Int J Legal Med (2001) 115 : 48–53

TECHNICAL NOTE

Wing K. Fung · Yue-Qing Hu

The evaluation of mixed stains from different ethnic origins: general result and common cases

Received: 24 July 2000 / Accepted: 7 January 2001

Abstract In some situations, it can be inferred from the crime circumstances that the mixed stain donors are of different ethnic groups. The evaluation of DNA mixtures with contributors coming from more than one ethnic group is considered under the assumption of independence of alleles within and between ethnic groups. A general formula is derived for the assessment of the weight of evidence in mixed stain problems. This formula is equivalent to that of Fukshansky and Bär, but we give a different derivation. For the convenience of practitioners, the explicit expressions of the likelihood ratios for 14 common cases are presented. The effect of different ethnic groups to the assessment of the evidence is shown in the well-known Simpson case.

Keywords DNA mixture · Ethnic groups · Forensic science · Likelihood ratio · Hardy-Weinberg law

Introduction

The evaluation of mixed DNA stains has been discussed by several authors and the likelihood ratio (*LR*) is a convenient tool for assigning the weight of evidence. Weir et al. (1997) and Fukshansky and Bär (1998) developed a general formula for the calculation of *LR* under the assumption of independence of alleles within and between loci. Taking the population structure into account, Curran et al. (1999) and Fung and Hu (2000a) derived general formulae for the calculation of *LR*. As a special case, the calculating formulae in Weir et al. (1997) and Fukshansky and Bär (1998) can be obtained when the co-ancestry coefficient is taken as zero.

In the interpretation of DNA mixtures, one fact we have to face is that of the contributors coming from different ethnic or racial groups. For example, the mixed stain collected from the crime scene contains material from the victim and the offender and they belong to the Caucasian and Negro ethnic groups, respectively. Extensive studies from a wide variety of databases show that there are indeed substantial frequency differences among the major racial and linguistic groups (Second National Research Council Report, hereafter called NRC-II, 1996). The presence of this phenomenon may not be ignored and may be misleading in the presentation of evidence in the court. Harbison and Buckleton (1998), following the sampling formula developed by Balding and Nichols (1994), discussed the interpretation of DNA mixtures when the two contributors come from two different races. Recently, Fukshansky and Bär (1999) obtained a formula for the calculation of *LR* for the situation that the contributors belonged to different ethnic origins. The independence of alleles within and between ethnic groups was employed for the calculations. Other recent discussions on mixtures can be referred to Buckleton et al. (1997), Clayton et al. (1998), Fung and Hu (2000b), and Triggs et al. (2000).

The purpose of this paper is three-fold. First, we give an alternative proof to the general formula obtained by Fukshansky and Bär (1999). We also notice that the match probability depends only on the allele frequencies of the ethnic groups to which the unknown contributors of the mixed stain belong. Second, the *LR*s of some common cases are provided explicitly, which are useful to practitioners who can get a better understanding of the problem and can get more feelings about the formulas. Third, a computer programme is developed for evaluating the *LR* under the general situation and the effect of ethnic groups on the weight of evidence is shown in the well-known Simpson case. Interested readers can contact the second author on the availability of the programme.

W. K. Fung (✉)
Department of Statistics and Actuarial Science,
The University of Hong Kong, Pokfulam Road, Hong Kong,
People's Republic of China
e-mail: hrntfwk@hku.hk,
Tel.: +852-28591988, Fax: +852-28589041

Y.-Q. Hu (✉)
Department of Applied Mathematics, Southeast University,
Nanjing 210018, People's Republic of China

Notations and formula

Suppose a crime was committed and a mixed stain was collected from the scene and some people, for example the victim and the suspects, were typed. Let M denote the distinct alleles of the mixed stain at a particular locus and K denotes the corresponding statistical profile (not necessarily distinct) of the known persons, where we assume the independence of alleles within and between loci. Thus the evidence could be written as (M,K). There will be alternative propositions about who are the contributors to the mixed stain. We let H_p be the prosecution proposition and H_d be the defence proposition. The likelihood ratio defined by:

$$LR = \frac{P(\text{Evidence} \mid H_p)}{P(\text{Evidence} \mid H_d)} = \frac{P(M,K \mid H_p)}{P(M,K \mid H_d)} \tag{1}$$

is usually used for assigning the weight of evidence for one locus. The overall LR can be obtained by multiplying over all loci.

Under either proposition H_p or H_d, the known contributors to the mixed stain and the number of unknown contributors were specified. Under proposition H_p for example, let X denote the genetic profile (Curran et al. 1999) of the x unknown contributors. Also let U denote the alleles present in M with the alleles of the specified known contributors removed, then we have $U \subset X \subset M$. So the probability $P(M,K|H_p)$ could be transformed as $P(K|H_p)P(U \subset X \subset M)$ using the independence between and within racial groups. After applying a similar procedure to the denominator of LR, it is easy to see that the LR is the ratio of two such probabilities expressed as $P(U \subset X \subset M)$ because $P(K|H_p) = P(K|H_d)$. In the following, we adopt the notation $P_x(U,M)$ instead of $P(U \subset X \subset M)$ to represent the probability that x unknown contributors have all the alleles in set U, but have no alleles not in M. Hence, the calculation of LR is converted to the calculation of the probability $P_x(U,M)$.

In the following, we consider the situation that the persons associated with the trial come from the same or different ethnic groups. Let {1, 2, 3,...} stand for the alleles $\{A_1, A_2, A_3,...\}$ for simplicity and $G = \{a, b,...\}$ denotes the group indices. The allele frequencies of types {1, 2, 3,...} in racial group g ($g = a, b,...$) are denoted as $\{p_{g1}, p_{g2},...\}$. The number of unknown contributors is: $x = \sum_g x_g$ where x_g is the number of unknown contributors belonging to ethnic group g, $g = a, b,....$ Then the formula for calculating the probability could be expressed as:

$$\begin{aligned} P_x(U,M) = & \prod_{g\in G}\left(\sum_{i\in M} p_{gi}\right)^{2x_g} \\ & - \sum_{j\in U}\prod_{g\in G}\left(\sum_{i\in M\setminus\{j\}} p_{gi}\right)^{2x_g} \\ & + \sum_{j,k\in U}\prod_{g\in G}\left(\sum_{i\in M\setminus\{j,k\}} p_{gi}\right)^{2x_g} \\ & - \sum_{j,k,l\in U}\prod_{g\in G}\left(\sum_{i\in M\setminus\{j,k,l\}} p_{gi}\right)^{2x_g} \\ & + \cdots + (-1)^{|U|}\prod_{g\in G}\left(\sum_{i\in M\setminus U} p_{gi}\right)^{2x_g}, \end{aligned} \tag{2}$$

where $|U|$ denotes the cardinality of set U. The proof of this formula can be referred to in the Appendix.

This formula can be transformed easily for numerical evaluation based on computers. From the right side of the above equation, it is interesting to note that the probability $P_x(U,M)$ depends merely on the ethnic group in which the unknown contributor number, x_g is positive. In other words, for any proposition H, we only need to specify to which groups the unknown contributors belong and need not care about the groups that the other known contributors belong to. This provides a great convenience in programming the calculation of $P_x(U,M)$. One special case is when $x = 0$. In this case, U must be empty and the probability $P_x(\phi,M)$ is always one.

*LR*s for some common cases

Since the general formula in Eq. (2) is in a non-trivial algebraic form, it will be helpful to practitioners if explicit expressions of the *LR*s are available to them. This section presents the expressions for some common cases with the number of unknown contributors being one or two. For simplicity, the suspect is abbreviated as *S*, the victim as *V*, the numerator of the *LR* is abbreviated as *NUM*, and the denominator as *DEN*. In addition, some general expressions of $P_x(U,M)$ for cases $x_a = x_b = 1$ are given for the purpose of practical applications.

One victim, one suspect and one unknown

Suppose a mixed stain M was recovered, and the victim and a suspect were typed. The two alternative propositions are:

H_p: The contributors of M were the victim and the suspect,
H_d: The contributors of M were the victim and one unknown person. (3)

Based on the discussion on the calculation formula in Eq. (2), we should specify to which ethnic group the unknown contributor belongs. Without loss of generality, it is assumed that he comes from ethnic group a.

Consider the case that $M = A_1A_2A_3$, $V = A_1A_2$ and $S = A_3A_3$. Since the unknown contributor under H_p is zero, so $NUM = 1$. Under H_d, the number of unknown contributors is 1, i.e. $x = 1$ and this unknown contributor must contain the allele $U = A_3$. Therefore, $DEN = P_1(A_3,A_1A_2A_3) \equiv P_1(3, 123) = (p_{a1} + p_{a2} + p_{a3})^2 - (p_{a1} + p_{a2})^2 = (2p_{a1} + 2p_{a2} + p_{a3})p_{a3}$. So $LR = 1/(2p_{a1} + 2p_{a2} + p_{a3})p_{a3}$.

Based on a similar approach, we can derive the *LR*s for the cases:

Table 1 Likelihood ratio for one victim, one suspect and one unknown case with H_p: the contributors were the victim and the suspect and H_d: the contributors were the victim and one unknown person of group *a*

M	V	S	Likelihood ratio
$A_1A_2A_3$	A_1A_2	A_3A_3	$1/[(2p_{a1} + 2p_{a2} + p_{a3})p_{a3}]$
$A_1A_2A_3$	A_1A_2	A_1A_3	$1/[(2p_{a1} + 2p_{a2} + p_{a3})p_{a3}]$
$A_1A_2A_3$	A_1A_1	A_2A_3	$1/(2p_{a2}p_{a3})$
$A_1A_2A_3A_4$	A_1A_2	A_3A_4	$1/(2p_{a3}p_{a4})$

(i) $M = A_1A_2A_3$, $V = A_1A_2$, $S = A_1A_3$
(ii) $M = A_1A_2A_3$, $V = A_1A_1$, $S = A_2A_3$
(iii) $M = A_1A_2A_3A_4$, $V = A_1A_2$, $S = A_3A_4$

All these *LR*s are listed in Table 1.

One suspect, two unknowns

In cases that the mixed stain did not originate from the victim and one suspect was identified, the two alternative propositions could be:

H_p: The contributors were the suspect and one unknown X_1 (ethnic group *a*);
H_d: The contributors were two unknown persons X_1 and X_2 (4)

Consider the situation that $M = A_1A_2A_3$, $S = A_1A_2$ and the two unknowns come from ethnic group *a*. Under H_p, there is an unknown contributor, $x = 1$, and the known contributor is the suspect, so $U = A_3$. Hence $NUM = P_1(3, 123) = (2p_{a1} + 2p_{a2} + p_{a3})p_{a3}$. Under H_d, $x = 2$, and there is no known contributor, so $U = A_1A_2A_3$. Hence $DEN = P_2(123, 123) = 12p_{a1}p_{a2}p_{a3}(p_{a1} + p_{a2} + p_{a3})$ after simplification. Thus, $LR = (2p_{a1} + 2p_{a2} + p_{a3})/[12p_{a1}p_{a2}(p_{a1} + p_{a2} + p_{a3})]$.

Based on a similar derivation, the *LR*s can be obtained for the following cases:

(i) $M = A_1A_2A_3$ and $S = A_1A_2$ with two unknowns from ethnic groups *a* and *b*
(ii) $M = A_1A_2A_3$ and $S = A_1A_1$ with two unknowns from the same ethnic group or different groups
(iii) $M = A_1A_2A_3A_4$ and $S = A_1A_2$

All the derived results are shown in Table 2.

Two suspects, two unknowns

When two suspects were identified, we may consider the two alternative propositions as:

H_p: The contributors were the two suspects, (5)
H_d: The contributors were two unknown persons X_1 and X_2.

Suppose the mixed stain is $M = A_1A_2A_3$, suspect 1 has $S_1 = A_1A_2$, and suspect 2 has $S_2 = A_1A_3$ (or $S_1 = A_1A_1$, $S_2 = A_2A_3$). Under H_p, there is no unknown contributor and so $NUM = 1$. Under H_d, there is no known contributor, so $U = A_1A_2A_3$. The denominator is $DEN = P_2(123, 123)$.

When the two unknowns come from ethnic group *a*, we have $DEN = 12p_{a1}p_{a2}p_{a3}(p_{a1} + p_{a2} + p_{a3})$. Thus, $LR = 1/[12p_{a1}p_{a2}p_{a3}(p_{a1} + p_{a2} + p_{a3})]$.

Three other scenarios relating to this 2-suspects/2-unknowns case were also considered. All these *LR*s are reported in Table 3.

Miscellaneous

For completion, Table 4 lists some expressions of $P_x(U,M)$ for $x_a = x_b = 1$.

Example

The Simpson case is considered here to show the effect of different ethnic groups. A three-banded RFLP mixed profile $A_1A_2A_3$ at the locus D2S44 was recovered from the centre console of an automobile owned by the defendant (Weir et al. 1997). The profiles of the defendant OJ and a victim RG were found to be A_1A_2 and A_1A_3, respectively.

Table 2 Likelihood ratio for one suspect and two unknowns case with H_p: The contributors were the suspect and one unknown X_1, and H_d: The contributors were two unknowns X_1 and X_2

M	S	Ethnicity		Likelihood ratio
		X_1	X_2	
$A_1A_2A_3$	A_1A_2	a	a	$(2p_{a1} + 2p_{a2} + p_{a3})/[12p_{a1}p_{a2}(p_{a1} + p_{a2} + p_{a3})]$
$A_1A_2A_3$	A_1A_2	a	b	$(2p_{a1} + 2p_{a2} + p_{a3})p_{a3}/(2p^2_{a1}p_{b2}p_{b3} + 2p^2_{a2}p_{b1}p_{b3} + 2p^2_{a3}p_{b1}p_{b2} + 2p^2_{b1}p_{a2}p_{a3} + 2p^2_{b2}p_{a1}p_{a3} + 2p^2_{b3}p_{a1}p_{a2} + 4p_{a1}p_{a2}p_{b1}p_{b3} + 4p_{a1}p_{a3}p_{b1}p_{b2} + 4p_{a2}p_{a1}p_{b2}p_{b3} + 4p_{a2}p_{a3}p_{b2}p_{b1} + 4p_{a3}p_{a1}p_{b3}p_{b2} + 4p_{a3}p_{a2}p_{b3}p_{b1})$
$A_1A_2A_3$	A_1A_1	a	a	$1/[6p_{a1}(p_{a1} + p_{a2} + p_{a3})]$
$A_1A_2A_3$	A_1A_1	a	b	$2p_{a2}p_{a3}/(2p^2_{a1}p_{b2}p_{b3} + 2p^2_{a2}p_{b1}p_{b3} + 2p^2_{a3}p_{b1}p_{b2} + 2p^2_{b1}p_{a2}p_{a3} + 2p^2_{b2}p_{a1}p_{a3} + 2p^2_{b3}p_{a1}p_{a2} + 4p_{a1}p_{a2}p_{b1}p_{b3} + 4p_{a1}p_{a3}p_{b1}p_{b2} + 4p_{a2}p_{a1}p_{b2}p_{b3} + 4p_{a2}p_{a3}p_{b2}p_{b1} + 4p_{a3}p_{a1}p_{b3}p_{b2} + 4p_{a3}p_{a2}p_{b3}p_{b1})$
$A_1A_2A_3A_4$	A_1A_2	a	a	$1/(12p_{a1}p_{a2})$
$A_1A_2A_3A_4$	A_1A_2	a	b	$2p_{a3}p_{a4}/(4p_{a1}p_{a2}p_{b3}p_{b4} + 4p_{a1}p_{a3}p_{b2}p_{b4} + 4p_{a1}p_{a4}p_{b2}p_{b3} + 4p_{a2}p_{a3}p_{b1}p_{b4} + 4p_{a2}p_{a4}p_{b1}p_{b3} + 4p_{a3}p_{a4}p_{b1}p_{b2})$

Table 3 Likelihood ratio for two suspects and two unknowns case with H_p: The contributors were two suspects and H_d: The contributors were two unknowns X_1 and X_2

M	S_1	S_2	Ethnicity		Likelihood ratio
			X_1	X_2	
$A_1A_2A_3$	A_1A_2	A_1A_3[a]	a	a	$1/[12p_{a1}p_{a2}p_{a3}(p_{a1} + p_{a2} + p_{a3})]$
$A_1A_2A_3$	A_1A_2	A_1A_3[a]	a	b	$1/(2p^2_{a1}p_{b2}p_{b3} + 2p^2_{a2}p_{b1}p_{b3} + 2p^2_{a3}p_{b1}p_{b2} + 2p^2_{b1}p_{a2}p_{a3} + 2p^2_{b2}p_{a1}p_{a3} + 2p^2_{b3}p_{a1}p_{a2} + 4p_{a1}p_{a2}p_{b1}p_{b3} + 4p_{a1}p_{a3}p_{b1}p_{b2} + 4p_{a2}p_{a1}p_{b2}p_{b3} + 4p_{a2}p_{a3}p_{b2}p_{b1} + 4p_{a3}p_{a1}p_{b3}p_{b2} + 4p_{a3}p_{a2}p_{b3}p_{b1})$
$A_1A_2A_3A_4$	A_1A_2	A_3A_4	a	a	$1/(24p_{a1}p_{a2}p_{a3}p_{a4})$
$A_1A_2A_3A_4$	A_1A_2	A_3A_4	a	b	$1/(4p_{a1}p_{a2}p_{b3}p_{b4} + 4p_{a1}p_{a3}p_{b2}p_{b4} + 4p_{a1}p_{a4}p_{b2}p_{b3} + 4p_{a2}p_{a3}p_{b1}p_{b4} + 4p_{a2}p_{a4}p_{b1}p_{b3} + 4p_{a3}p_{a4}p_{b1}p_{b2})$

[a] also holds for S1 = A1A1, S2 = A2A3

Table 4 Cases of $P_2(U,M)$, when the two unknown contributors come from ethnic groups *a* and *b*, respectively

One allele	
$P_2(\phi, 1)$	$(p_{a1})^2(p_{b1})^2$
$P_2(1, 1)$	$(p_{a1})^2(p_{b1})^2$
Two alleles	
$P_2(\phi, 12)$	$(p_{a1} + p_{a2})^2(p_{b1} + p_{b2})^2$
$P_2(1, 12)$	$(p_{a1} + p_{a2})^2(p_{b1} + p_{b2})^2 - (p_{a2})^2(p_{b2})^2$
$P_2(12, 12)$	$(p_{a1} + p_{a2})^2(p_{b1} + p_{b2})^2 - (p_{a1})^2(p_{b1})^2 - (p_{a2})^2(p_{b2})^2$
Three alleles	
$P_2(\phi, 123)$	$(p_{a1} + p_{a2} + p_{a3})^2(p_{b1} + p_{b2} + p_{b3})^2$
$P_2(1, 123)$	$(p_{a1} + p_{a2} + p_{a3})^2(p_{b1} + p_{b2} + p_{b3})^2 - (p_{a2} + p_{a3})^2(p_{b2} + p_{b3})^2$
$P_2(12, 123)$	$(p_{a1} + p_{a2} + p_{a3})^2(p_{b1} + p_{b2} + p_{b3})^2 - (p_{a2} + p_{a3})^2(p_{b2} + p_{b3})^2 - (p_{a1} + p_{a3})^2(p_{b1} + p_{b3})^2 + (p_{a3})^2(p_{b3})^2$
$P_2(123, 123)$	$2p^2_{a1}p_{b2}p_{b3} + 2p^2_{a2}p_{b1}p_{b3} + 2p^2_{a3}p_{b1}p_{b2} + 2p^2_{b1}p_{a2}p_{a3} + 2p^2_{b2}p_{a1}p_{a3} + 2p^2_{b3}p_{a1}p_{a2} + 4p_{a1}p_{a2}p_{b1}p_{b3} + 4p_{a1}p_{a3}p_{b1}p_{b2} + 4p_{a2}p_{a1}p_{b2}p_{b3} + 4p_{a2}p_{a3}p_{b2}p_{b1} + 4p_{a3}p_{a1}p_{b3}p_{b2} + 4p_{a3}p_{a2}p_{b3}p_{b1}$
Four alleles	
$P_2(\phi, 1234)$	$(p_{a1} + p_{a2} + p_{a3} + p_{a4})^2(p_{b1} + p_{b2} + p_{b3} + p_{b4})^2$
$P_2(1, 1234)$	$(p_{a1} + p_{a2} + p_{a3} + p_{a4})^2(p_{b1} + p_{b2} + p_{b3} + p_{b4})^2 - (p_{a2} + p_{a3} + p_{a4})^2 (p_{b2} + p_{b3} + p_{b4})^2$
$P_2(12, 1234)$	$(p_{a1} + p_{a2} + p_{a3} + p_{a4})^2(p_{b1} + p_{b2} + p_{b3} + p_{b4})^2 - (p_{a2} + p_{a3} + p_{a4})^2 (p_{b2} + p_{b3} + p_{b4})^2 - (p_{a1} + p_{a3} + p_{a4})^2(p_{b1} + p_{b3} + p_{b4})^2 + (p_{a3} + p_{a4})^2 (p_{b3} + p_{b4})^2$
$P_2(123, 1234)$	$2p^2_{a1}p_{b2}p_{b3} + 2p^2_{b1}p_{a2}p_{a3} + 4p_{a1}p_{a2}p_{b1}p_{b3} + 4p_{a1}p_{a3}p_{b1}p_{b2} + 2p^2_{a2}p_{b1}p_{b3} + 2p^2_{b2}p_{a1}p_{a3} + 4p_{a2}p_{a1}p_{b2}p_{b3} + 4p_{a2}p_{a3}p_{b2}p_{b1} + 2p^2_{a3}p_{b1}p_{b2} + 2p^2_{b3}p_{a1}p_{a2} + 4p_{a3}p_{a1}p_{b3}p_{b2} + 4p_{a3}p_{a2}p_{b3}p_{b1} + 4p_{a1}p_{a2}p_{b3}p_{b4} + 4p_{a1}p_{a3}p_{b2}p_{b4} + 4p_{a1}p_{a4}p_{b2}p_{b3} + 4p_{a2}p_{a3}p_{b1}p_{b4} + 4p_{a2}p_{a4}p_{b1}p_{b3} + 4p_{a3}p_{a4}p_{b1}p_{b2}$
$P_2(1234, 1234)$	$4p_{a1}p_{a2}p_{b3}p_{b4} + 4p_{a1}p_{a3}p_{b2}p_{b4} + 4p_{a1}p_{a4}p_{b2}p_{b3} + 4p_{a2}p_{a3}p_{b1}p_{b4} + 4p_{a2}p_{a4}p_{b1}p_{b3} + 4p_{a3}p_{a4}p_{b1}p_{b2}$

The court ordered that for the statistical calculation, the number of contributors to the mixed samples should be taken as at least two, three and four but here we only take it equal to two for illustration. The following propositions are of interest.

H_p: Contributors were thevictim and the suspect;
H_d: Contributors were 2 unknown persons. (6)

The defendant was an African-American and the victim was a Caucasian. We regard the 2 unknown persons as being African-American or Caucasians. The following allele frequencies are taken, African-American: $p_{a1} = 0.0316$, $p_{a2} = 0.0842$, $p_{a3} = 0.0926$, and Caucasian: $p_{b1} = 0.0859$, $p_{b2} = 0.0827$, $p_{b3} = 0.1073$; also see Budowle et al. (1991) and Fung (1996). Taking the single-banded alleles as true homozygotes, the effect of different ethnic groups to the likelihood ratio is investigated and the results are shown in Table 5. Weir et al. (1997) obtained the likelihood ratio 1,623, using the allele frequencies of African-Americans for the 2 unknown persons. If the two unknown persons were one African-American and one Caucasian, the *LR* drops to 727 (less than one half) and drops further to 396 (less than one-quarter) if the two unknowns were Cau-

Table 5 Likelihood ratios with two unknowns belonging to ethnic groups of African-American(*AA*) and Caucasian (*C*)

2*AA*s	(1*AA*, 1*C*)	2*C*s
1623	727	396

casians. Thus, the effect of different ethnic groups could be large. It is to be noticed that the ethnic group of the defendant does not matter for the *LR*, only the ethnic groups of the unknowns matter.

In conclusion, when the contributors to the mixed stain come from more than one ethnic group, the evaluation of the evidence is considered using the likelihood ratio, under the assumption of Hardy-Weinberg law and linkage equilibrium. In some cases, the independence between alleles does not hold, for example, due to population substructure (Curran et al. 1999; Fung and Hu 2000a). It is expected that the method considered here could be modified to meet this need.

In this paper, we assume the number of unknown contributors to the mixed stain to be known, but this is not always the case. We can use a range of values for the number of unknowns if it is not certain and let the courtroom decide on which is the appropriate one.

The consideration of peak area and peak height can enhance the interpretation of the evidence (Clayton et al. 1998), but we have ignored this information and treat each combination of genotypes with equal possibility. The interpretation of mixtures taking information on peak height into account, is currently under investigation by the authors.

Appendix

Proof. First of all, we assume that the x unknown contributors come from two different ethnic groups, say groups a and b, i.e. $x = x_a + x_b$, $x_a > 0$, $x_b > 0$. That means that x_a unknown contributors come from ethnic group a, x_b unknown contributors come from group b and the total number of unknown contributors is x. Without loss of generality, we can assume $M = \{1,2,...,m\}$ and $U = \{1,2,...,n\}$, where $0 \le n \le m$. Note that $n = 0$ corresponds to the case $U = \phi$. Since the alleles of the x unknown contributors must have the alleles contained in U and cannot have alleles not found in M, the x_1 unknown contributors have $2x_1$ alleles and the x_2 unknown contributors have $2x_2$ alleles, so the probability $P_x(U,M)$ is just the sum

$$\sum_{\substack{i_1+i_2+\cdots+i_m=2x_1\\ j_1+j_2+\cdots+j_m=2x_2\\ i_1+j_1\ge 1, i_2+j_2\ge 1,\ldots,i_n+j_n\ge 1}} (p_{a1})^{i_1}(p_{a2})^{i_2}\cdots (p_{am})^{i_m}(p_{b1})^{j_1}(p_{b2})^{j_2}\cdots(p_{bm})^{j_m} \tag{7}$$

by using the Hardy-Weinberg law and independence of alleles between ethnic groups.

Let $I = (i_1,i_2,...,i_m)$, $J = (j_1,j_2,...,j_m)$, $B_1 = \{i_1 + j_1 = 0\}$, $B_2 = \{i_2 + j_2 = 0\},\ldots, B_n = \{i_n + j_n = 0\}$, $B = \{i_1 + i_2 + \ldots + i_m = 2x_1, j_1 + j_2 + \ldots + j_m = 2x_2\}$, $f(I,J) = (p_{a1})^{i_1}(p_{a2})^{i_2}\ldots(p_{am})^{i_m}(p_{b1})^{j_1}(p_{b2})^{j_2}\ldots(p_{bm})^{j_m}$ and $A = B \cap (\bar{B}_1 \cap \bar{B}_2 \cap \ldots \cap \bar{B}_n)$
then $A = B \cap \overline{B_1 \cup B_2 \cup \ldots B_n} = B\backslash(\cup_{i=1}^{n} BB_i)$ and

$$P_x(U,M) = \sum_{I\in A,\, J\in A} f(I,J) = \sum_{I\in B,\, J\in B} f(I,J) - \sum_{I\in \cup_{i=1}^{n} BB_i,\, J\in \cup_{i=1}^{n} BB_i} f(I,J) = \sum_{I\in B,\, J\in B} f(I,J)$$

$$-\sum_{j}\left(\sum_{I\in BB_j,\, J\in BB_j} f(I,J)\right) + \sum_{j,k}\left(\sum_{I\in BB_jB_k,\, J\in BB_jB_k} f(I,J)\right) - \sum_{j,k,l}\left(\sum_{I\in BB_jB_kB_l,\, J\in BB_jB_kB_l} f(I,J)\right)+\cdots \tag{8}$$

It is obvious that

$$\sum_{I\in B,\, J\in B} f(I,J) = \sum_{i_1+i_2+\cdots+i_m=2x_1} (p_{a1})^{i_1}(p_{a2})^{i_2}\cdots(p_{am})^{i_m} \sum_{j_1+j_2+\cdots+j_m=2x_2} (p_{b1})^{j_1}(p_{b2})^{j_2}\cdots(p_{bm})^{j_m} = \left(\sum_{i\in M} p_{ai}\right)^{2x_1}\left(\sum_{i\in M} p_{bi}\right)^{2x_2} \tag{9}$$

$$\sum_{I\in BB_j,\, J\in BB_j} f(I,J) = \left(\sum_{i\in M\backslash\{j\}} p_{ai}\right)^{2x_1}\left(\sum_{i\in M\backslash\{j\}} p_{bi}\right)^{2x_2}$$

$$\sum_{\substack{I\in BB_jB_k,\\ J\in BB_jB_k}} f(I,J) = \left(\sum_{i\in M\backslash\{j,k\}} p_{ai}\right)^{2x_1}\left(\sum_{i\in M\backslash\{j,k\}} p_{bi}\right)^{2x_2} \tag{10}$$

$$\sum_{\substack{I\in BB_jB_kB_l,\\ J\in BB_jB_kB_l}} f(I,J) = \left(\sum_{i\in M\backslash\{j,k,l\}} p_{ai}\right)^{2x_1}\left(\sum_{i\in M\backslash\{j,k,l\}} p_{bi}\right)^{2x_2}$$

Thus we have

$$\begin{aligned} P_x(U,M) = {} & \left(\sum_{i\in M} p_{ai}\right)^{2x_1}\left(\sum_{i\in M} p_{bi}\right)^{2x_2} \\ & - \sum_{j\in U}\left(\sum_{i\in M\backslash\{j\}} p_{ai}\right)^{2x_1}\left(\sum_{i\in M\backslash\{j\}} p_{bi}\right)^{2x_2} \\ & + \sum_{j,k\in U}\left(\sum_{i\in M\backslash\{j,k\}} p_{ai}\right)^{2x_1}\left(\sum_{i\in M\backslash\{j,k\}} p_{bi}\right)^{2x_2} \\ & - \sum_{j,k,l\in U}\left(\sum_{i\in M\backslash\{j,k,l\}} p_{ai}\right)^{2x_1}\left(\sum_{i\in M\backslash\{j,k,l\}} p_{bi}\right)^{2x_2} \\ & +\cdots+(-1)^{|U|}\left(\sum_{i\in M\backslash U} p_{ai}\right)^{2x_1}\left(\sum_{i\in M\backslash U} p_{bi}\right)^{2x_2}, \end{aligned} \tag{11}$$

and hence Eq. (2) holds.

From the process of the proof for $G = \{a,b\}$, it is not difficult to understand the calculation formula Eq. (2) holds for general G by the principle of induction.

Acknowledgements The authors thank the two referees and the Coordinating Editor for helpful comments. The first author is partially supported by the Hong Kong RGC Competitive Earmarked Research Grant HKU 7136/97H.

References

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

Buckleton JS, Evett IW, Weir BS (1997) Setting bounds for the likelihood ratio when multiple hypotheses are postulated. Sci Justice 37:23–26

Budowle B, Monson KL, Anoe K, Baechtel FS, Bergman D, et al, (1991) A preliminary report on binned general population data on six VNTR loci in Caucasians, Blacks, and Hispanics from the United States. Crime Lab Digest 18:9–26

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

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

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

Fukshansky N, Bär W (1999) Biostatistical evaluation of mixed stains with contributors of different ethnic origin. Int J Legal Med 112:383–387

Fung WK (1996) 10% or 5% match window in DNA profiling. Forensic Sci Int 78:111–118

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

Fung WK, Hu YQ (2000b) Interpreting DNA mixtures based on the NRC-II Recommendation 4.1. Forensic Sci Commun 2(4) (available at http: //www/fbi/gov/programs/lab/fsc/backissu/oct2000/fung.htm)

Harbison SA, Buckleton JS (1998) Applications and extensions of subpopulation theory: a caseworkers guide. Sci Justice 38:249–254

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

Triggs CM, Harbison SA, Buckleton J (2000) The calculation of DNA match probabilities in mixed race populations. Sci Justice 40:33–38

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