# Bounding the Number of Contributors to Mixed DNA Stains*


Steffen L. Lauritzen† — Aalborg University

Julia Mortera — Università Roma Tre


February 18, 2002


**Abstract**

We derive a simple inequality for the probability of observing a given DNA profile when assuming a fixed number of unknown persons have contributed to the mixed stain. We then show how this inequality can be used to obtain an upper bound for the number of unknown contributors needed to be considered.


## 1 Introduction

Suppose that evidence in a given crime case is available in the form of a DNA profile for a mixed stain, which suggests that two or more persons may have contributed to the mixture. The contributors to the mixture could possibly be persons with known DNA profiles, such as the victim and suspect, but possibly also unknown individuals. Furthermore, suppose there is uncertainty or dispute about the total number of contributors involved.

As it is impossible to evaluate the strength of evidence for all possible numbers of contributors, it is of interest to identify an upper bound $b$ on the unknown number of contributors worth considering.

This issue has, for example, been discussed by Brenner, Fimmers and Baur (1996), Weir, Triggs, Starling, Stowell, Walsh and Buckleton (1997), and Buckleton, Evett and Weir (1998). It can be argued that, usually, this upper bound can be set to be equal to the minimal number of contributors necessary to explain the number of different alleles observed in the profile. However, the argument leading to this is inexact and the bound is not universally true for small numbers of contributors, as also pointed out by Weir et al. (1997).


*This is Research Report R-02-2003, Department of Mathematical Sciences, Aalborg University.



†Address for correspondence: Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, DK-9220 Aalborg, Denmark.

The purpose of this note is to derive exact bounds and illustrate their potential use.

## 2 The inequality

A typical hypothesis $H$ to be evaluated would specify both a set of known individuals as contributors to the mixed stain and a number $x$ of unknown contributors. We assume $M$ markers are used and denote the observed evidence profile by $E = (E_1, \ldots, E_M)$, with $E_m$ being the observed set of alleles at marker $m$. Similarly $K = (K_1, \ldots, K_M)$ are the alleles carried by the known individuals, and $U = (U_1, \ldots, U_M)$ the alleles supplied by the unknown individuals.

For a hypothesis $H$ involving $x$ unknown individuals, the relevant likelihood is

$$P_x(E \mid H) = P(U_m \cup K_m = E_m, m = 1, \ldots, M \mid H).$$

Weir et al. (1997) denoted this probability by $P_x(U \mid E)$.

Our bound relies on the simple fact that this probability is necessarily smaller than the probability that none of the alleles of the unknown contributors are outside those in $E$, i.e.

$$P_x(E \mid H) \leq P(U_m \subseteq E_m, m = 1, \ldots, M \mid H). \tag{1}$$

Suppose the $M$ markers are independent and all the unknown individuals are from the same population. Then the latter probability is

$$P(U_m \subseteq E_m, m = 1, \ldots, M \mid H) = \prod_{m=1}^{M} \left( \sum_{a \in E_m} p_a^m \right)^{2x},$$

where $p_a^m$ denotes the frequency of allele $a$ at marker $m$. From (1) we thus have that for any $x$

$$P_x(E \mid H) \leq \prod_{m=1}^{M} \left( \sum_{a \in E_m} p_a^m \right)^{2x}. \tag{2}$$

If all possible alleles for all markers are represented in the evidence profile, the right-hand side of this inequality is equal to one and therefore useless. And, indeed in this case, the probability of observing the given evidence tends to one as the number of unknown contributors tends to infinity.

However, if just some alleles are not represented, the right-hand side of (2) will tend to zero at an exponential rate and yield a useful bound for the probability on the left-hand side. In fact, the right-hand side of (2) is the leading term in the expansion of $P_x(E \mid H)$ given by Weir et al. (1997).

To ensure that $P_x(E \mid H)$ is smaller than a specified number, say $y$, $x$ should thus satisfy

$$\prod_{m=1}^{M} \left( \sum_{a \in E_m} p_a^m \right)^{2x} < y.$$

Taking logarithms and isolating $x$ yields the bound

$$x > b(y) = \frac{\ln y}{2 \sum_{m=1}^{M} \ln \left( \sum_{a \in E_m} p_a^m \right)},$$

where the inequality sign has been reversed because the denominator is negative. Thus

$$x > b(y) \Rightarrow P_x(E \mid H) \leq y. \tag{3}$$

The likelihood ratio $L$ needed to evaluate the evidence in favour of a hypothesis $H_0$ against an alternative hypothesis $H_1$ has the form

$$L = \frac{P_{x_0}(E \mid H_0)}{P_{x_1}(E \mid H_1)}$$

where $x_i$ denotes the number of unknown individuals involved in the hypothesis $H_i$. As argued by Brenner et al. (1996) it is sufficient to give a lower bound and thus consider a 'worst case' scenario in the denominator. So assume that the probability of the evidence for a given alternative $H_1$ has been considered.

For $y = P_{x_1}(E \mid H_1)$ it now follows from (3) that if the number of contributors for a given hypothesis $x$ is greater than $b(y)$, this hypothesis is less likely than $H_1$ and therefore need not be considered.

# 3 An example

We illustrate the use of the bound in an example discussed by Weir et al. (1997) involving five DNA markers. The data and appropriate gene frequencies are given in Table 1.

The hypothesis $H_0$ of the prosecutor is that the evidence profile consists of DNA from the victim and the suspect, and under this hypothesis the probability of the evidence is equal to one. A possible alternative hypothesis $H_1$ is that the evidence profile consists of DNA from the victim and a single unknown contributor. The relevant probabilities for each marker under this hypothesis were computed by Weir et al. (1997), as displayed in the last line of Table 1. Combining over all markers we find

$$P_1(E \mid H_1) = 0.321 \times 0.990 \times 0.352 = 0.11186.$$

The denominator of the bound $b(y)$ is evaluated to be

$$2 \sum_{m=1}^{M} \ln \left( \sum_{a \in E_m} p_a^m \right) = 2(\ln 0.567 + \ln(0.566 + 0.429)) = -1.1448$$

Table 1: Weir's example

| Profile | LDLR | GYPA | HBGG | D7S8 | GC |
|---|---|---|---|---|---|
| Evidence | B | AB | AB | AB | ABC |
| Victim | B | AB | AB | AB | AC |
| Suspect | B | A | A | A | B |
| $p_A$ | | 0.538 | 0.566 | 0.543 | 0.253 |
| $p_B$ | 0.567 | 0.462 | 0.429 | 0.457 | 0.195 |
| $p_C$ | | | | | 0.552 |
| $P_1(E \mid H_1)$ | 0.321 | 1.000 | 0.990 | 1.000 | 0.352 |

yielding the bound

$$b(0.11186) = \frac{\ln(0.11186)}{-1.1448} = \frac{-2.1905}{-1.1448} = 1.91$$

and therefore alternative hypotheses $H^*$ with more than a single unknown contributor need not be considered, at least their likelihood will be smaller than that of $H_1$.

## Acknowledgement

This research was partially supported by a Leverhulme Trust Research Interchange Grant. The manuscript has benefited from comments of Phil Dawid on an earlier version.

## References

Brenner, C. H., Fimmers, R. and Baur, M. P.: 1996, Likelihood ratios for mixed stains when the number of donors cannot be agreed, *International Journal of Legal Medicine* **109**, 218–219.

Buckleton, J. S., Evett, I. W. and Weir, B. S.: 1998, Setting bounds for the likelihood ratio when multiple hypotheses are postulated, *Science and Justice* **38**(1), 23–26.

Weir, B. S., Triggs, C. M., Starling, L., Stowell, L. I., Walsh, K. A. J. and Buckleton, J.: 1997, Interpreting DNA mixtures, *Journal of Forensic Science* **42**(2), 213–222.