

# Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities

Charles H. Brenner<sup>1</sup>

<sup>1</sup> School of Public Health, Forensic Science Group, U.C. Berkeley  
6801 Thornhill Drive, Oakland, California 94611-1336  
☎++1 510 339 1911, fax -1181  
cbrenner@berkeley.edu

## Abstract

DNA is a major and essential identification tool for mass fatality incidents including the hundreds of thousands of victims of the 2004 Indian Ocean tsunami. Mathematical complications characteristic of this sort of mass fatality include prevalence of related victims, the many races represented among the victims, and various identification modalities in tandem with DNA. Four mathematical problems of interest are discussed in this paper. (1) Other quantifiable factors (i.e. geography) can be formally accounted for by including a likelihood ratio that can be thought of as reducing the “effective number of victims.” (2) When a victim is found and tentatively identified as V, but then it comes to light that the victim has a relative W who is also missing, confidence in the identity is depressed. To account for the existence of W, increment the effective number of victims by the likelihood ratio supporting W as the identity of the victim. (3) When several apparently related victims are found, their mutual identities should be calculated simultaneously. Compared to one-at-a-time, serial identifications, this is both logical and may lead to much more confidence in the identities. (4) Although there may be many different population groups represented among the missing, it is generally sufficient to consider population statistics for only a few of them in deciding whether to declare an identification.

Keywords: DNA identification, Mass fatality, kinship, tsunami identification

## 1. Introduction

DNA is unique among identification modalities in lending itself well to mathematical treatment.

The purpose of this paper is to give some mathematical guidance and suggestions for determining identities of victims in the case of mass disasters similar to the December 2004 tsunami deaths in the Indian Ocean. The general outline of strategy for DNA-based mass fatality identification is reasonably clear from experience of this and earlier incidents [1, 2, 3, 4, 5]. DNA profiles of victim DNA (“PM” for post-mortem) are compared with reference profiles that come from relatives (indirect references) and/or from relics such as a toothbrush of the victim (direct reference). A likelihood ratio comparing appropriate hypotheses may then reasonably be interpreted using a Bayesian model [6] – in other words, assuming some prior odds – to give posterior odds of correct identification (1). A typical policy is that identification will be declared provided the posterior odds are high enough that the chance of a misidentification is very small (2).

On the other hand, the discipline of mass identification is not yet mature. Every disaster, particularly a gigantic one, has unique aspects, and it is already apparent what some of the novel mathematical complications of tsunami identifications are. The emphasis in this paper is on DNA calculations and a Bayesian approach. Beyond that the only unifying theme is relevance to the practical problem.

Assume that either through formal computer DNA screening as described previously [5], or via physical clues, tentative identifications exist.

The first issue is the context within which DNA information is to be interpreted – that is, the prior odds inferred from other, moderately quantifiable factors, like geography or the victim’s size. Physical location where remains were recovered was considered only slightly in making World Trade Center identifications, is generally useless in airplane crashes, and has unfortunately been of little help in the major effort to identify Bosnian war dead because of the chaotic and deadly history near Srebrenica. The Indian Ocean situation is different.

One then needs to evaluate: How strongly does the DNA evidence indicate that a given body is that of a particular missing person X? If relatives of X are also missing the analysis is more complex. Compared to the situation with no relatives, the identification is weaker if only X's body is found, but if several related bodies turn up their identities can reinforce one another, especially if the identification is based on comparison with living relatives rather than direct references.

Finally, what is the appropriate reference population for calculations? Obviously many races are represented among the dead, so it is helpful to recognize conditions under which some of those races can reasonably be ignored in making calculations.

## 2. Methods

### 2.1 Population allele frequencies

Population allele frequencies are from 15 Identifiler loci in Thai [7], Chinese [8], Caucasian [9], African-American [9], Japanese [10], Indian [8], Vietnamese [11], Korean [11], African [12, 13].

### 2.2 LR computation

Kinship LR computations are performed using the Symbolic Kinship [14] module of the author's DNA·VIEW software. The module also includes a facility to create simulated identification examples (<http://dna-view.com/simulate.htm>), used for several of the examples herein. Simulations used Thai allele frequencies [7].

### 2.3 Racial distances

For present purposes "race" simply means a population that has been isolated – i.e. endogamous – for a period of time and whose allele frequencies have therefore drifted compared to other populations. Allele frequency differences between races provide a possible way to assign individuals to race [15,16,17]. They also represent a possible pitfall when deciding on the confidence of a putative identification, because calculating with the wrong reference population creates a bias toward false confidence in the certainty of an identification. I therefore define "racial distance" as the typical ratio by which a DNA profile randomly selected from race  $r$  is expected to be more probable in race  $r$  than in race  $s$ . This is a useful statistic for evaluating the suitability of using a few reference races as proxies.

Racial distance estimates are calculated [15,16,17] from available population samples [7,8,9,10,11,12,13] as follows. Let the indices  $r$ ,  $l$ , and  $a$  range over races, loci, and alleles within a locus. Consider the probability  $q(r,l,a)$  of finding allele  $a$  at locus  $l$  in race  $r$ . Assuming the existence of allele population data for  $r$ , a collection of numbers  $c(r,l,a)$ , the numbers of times the allele was counted in a population sample, a plausible estimate (especially for length polymorphisms) of  $q$  is the sample frequency;

$$q(r,l,a)=c(r,l,a)/\sum_a c(r,l,a).$$

Also, define "extended-sample" frequencies  $q'$  based on extending the sample by any one allele:

$$q'(r,l,a)=(1+c(r,l,a))/(1+\sum_a c(r,l,a)).$$

If a profile  $E$  is simulated from population  $r$  by Monte Carlo selection from  $c(r,l,a)$  then  $L_E(r,s)=\prod_{a,l \in E} q(r,l,a)/q'(s,l,a)$  is the likelihood ratio favoring  $r$  over population  $s$  as the origin of  $E$ . (Extended-sample frequencies in the denominator compensate for ascertainment bias from using the numerator population data to choose a profile [15].) The "typical" (that is, expected in the sense of geometric mean) racial discrimination likelihood ratio from a full profile with two alleles per locus is then [15]

$$D(r,s) = \prod_{a,l} [q(r,l,a)/q'(s,l,a)]^{2q(r,l,a)}.$$

$D(r,s)$  can be calculated for any pair of races for which allele population data exists, and is a directed distance measure between the two races that is natural in a forensic context because it measures the expected error in using the wrong reference population for computing a random match probability.

By the Central Limit Theorem the distribution of  $\log L_E$  is nearly normal if the number of loci is large. Comparison of moments with Monte Carlo samples of 100,000 profiles confirms that the deviations from normality, with 15 locus profiles, are small. Kurtosis and skew are 3.3 and  $-1/6$ , versus 3 and 0 for a normal distribution.

### 3. Discussion

#### 3.1 Framework

A simplified example illustrates the Bayesian framework that I assume herein. Suppose  $v+1$  people are lost including a person Victor. A body  $V$  turns up. Without saying anything more the probability is  $1/(v+1)$  – i.e. the prior odds are  $1:v$  meaning  $1/v$  – that the body is Victor. Various evidence, such as DNA and the size and location of the body, combine to give a likelihood ratio of  $L$  supporting the proposition that the body is Victor rather than anyone else. Bayes' Theorem, formulated in terms of odds, says that

$$\text{posterior odds} = (\text{prior odds}) \cdot (\text{likelihood ratio}), \quad (1)$$

so the posterior odds are  $L/v$  that  $V$  is Victor. Since probability = odds/(odds+1), the posterior probability is very nearly  $1 - v/L$  so long as (as will be true in situations of interest)  $L$  is much larger than  $v$ . Each declared identification has some (small) probability  $v/L$  to be a mistake;  $v/L$  can also be interpreted as an *expected* (in the statistical sense) *number of misidentifications* in declaring that  $V=$ Victor.

Perhaps we have in mind a probabilistic budget – no doubt less than 1 – for the total number of expected errors to be accumulated over all declared identifications. Apportioning that budget among individual identifications suggests, as a baseline policy for declaring each identification, to insist on a stringent standard that

$$\text{expected number of misidentifications per identification} = v/L < \epsilon \quad (2)$$

for some small  $\epsilon$  such as  $\epsilon=1/10000$ . The simplest policy is to keep the same  $\epsilon$  for all victims, but it may be judicious to make exceptions occasionally while keeping an eye on the bottom line.

If  $L$  comes from several independent pieces of evidence, such as location found (geography), physical attributes, and DNA, then it can be expressed as a product of the likelihood ratios for each different piece of evidence:

$$L = (\text{geographical LR}) \cdot (\text{physical LR}) \cdot (\text{DNA LR}). \quad (3)$$

Substituting (3) into (1) and multiplying left to right (instead of multiplying all the likelihood ratio factors together first as in (3)) corresponds to the idea of climbing a ladder of evidence wherein each rung represents the posterior odds relative to the evidence below it, and at the same time is prior odds relative to the next evidence. That is,

$$\text{posterior-to-geography odds} = (\text{prior odds}) \cdot (\text{geographical LR}) \text{ and} \quad (4)$$

$$\text{prior-to-physical odds} = \text{posterior-to-geography odds};$$

$$\text{posterior-to-geography-and-physical odds} = (\text{prior-to-physical odds}) \cdot (\text{physical LR})$$

etc.

#### *Effective number of victims*

Suppose  $v+1=10000$  are missing in Thailand, 500 of them are missing from some particular seaside village  $SV$ , and assume that all 500 of these plus 10% of the 400 missing from a nearby village to the south will wash up near  $SV$ . If Kanya is missing from  $SV$ , what are the odds that a random body  $V$  is Kanya?

Geographical LR =  $X/Y$  where

$$X = \Pr(V \text{ appears near } SV \mid V=Kanya) = 1$$

$Y = \Pr(V \text{ appears near } SV \mid V \neq Kanya) = (499/v) + (400 \cdot 10\%/v) = 539/v$  since there are  $v$  victims in Thailand who are not Kanya, and 499 victims from  $SV$  who are not Kanya. So  $LR = v/539$  and

$$\text{posterior-to-geography odds} = (1/v) \cdot v/539 = 1/539,$$

a rather obvious and expected result. It is convenient to refer to 540 – or to 539, why split hairs? – as the *effective number of victims* after consideration of geography. This number is relative to the facts that the victim of interest comes from  $SV$  and that the body was found at  $SV$ ; otherwise it would be a different number.

#### **Related victims**

Since the tsunami hit people in their homes and on vacations, it is certain that many related victims are involved. Family

members perishing together is a typical complication of airplane crashes [3,18] – it was a conspicuous feature in sorting out some of the identities from the 1998 Swissair 111 crash [4] near Halifax and from the post-9/11 crash in Queens of AA587. Among the World Trade center victims there were not many who were related, and we did not systematically consider the problems of related victims – in one instance early on only by luck was a mistaken identification between brothers avoided.

Without related victims, the alternative identity hypotheses to consider are merely

$H_p$ : This PM profile represents missing person X,

$H_0$ : This PM profile is unrelated to X.

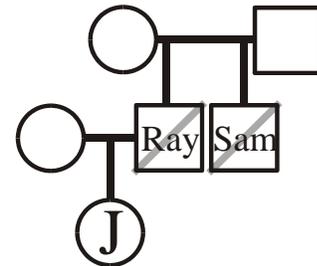
When relatives  $X_1, X_2, \dots$  are also possible victims, additional hypotheses must be calculated separately:

$H_i$ : This PM profile is  $X$ 's relative  $X_i$ .

The tsunami disaster scale, nature, and uncertainties require approaching related victim problems systematically. There are several different complications to consider.

### 3.2 One of related victims found

The tsunami differs from airplane crashes in that many victims will never be recovered at all, which brings an extra complication when there is a tentative identification for a missing victim who has relatives that are also, and still, missing. Suppose that brothers Ray and Sam are missing, and only Ray's daughter J is available as a reference (**Figure 1**). A body, V, is found, and comparing the DNA of V and of J suggests that they might be father and daughter ( $H_p$ ) – that V might be Ray. But of course the possibility  $H_1$  that V is Sam must also be considered; otherwise the probability  $\Pr(H_p)$  that V is Ray would be vastly overstated as a result of comparing it only with the “strawman”  $H_0$ .



**Figure 1** Identifying one of related victims. Is the body Ray or is it Sam?

In a typical situation there are prior odds  $\text{Odds}(\text{Ray})$  that the body is Ray and let's assume  $\text{Odds}(\text{Ray}) = \text{Odds}(\text{Sam}) = 1/v$ ,  $v+1$  being the effective number of victims. Let  $L(\text{Ray}, 0)$  be the likelihood ratio by which the DNA evidence favors Ray as the identity of V rather than an unrelated identity. Then if there were no Sam to consider, the posterior odds that V is Ray would be

$$\text{Odds}(V=\text{Ray} \mid \text{DNA, no Sam}) = L(\text{Ray}, 0)/v. \quad (5)$$

In view of the existence of Sam however, the posterior odds that  $V=\text{Ray}$  are different, and, from equation (14) (Appendix) – are expressed in a simple way as a combination of  $L(\text{Ray}, 0)$  and  $L(\text{Sam}, 0)$ :

$$\text{Odds}(V=\text{Ray} \mid \text{DNA}) \approx L(\text{Ray}, 0)/[L(\text{Sam}, 0) + v]. \quad (6)$$

Comparing (6) and (5) reveals the elegant rule that the possibility that the body is Sam rather than Ray is accounted for by imagining  $L(\text{Sam}, 0)$  additional effective victims.

For example  $L(\text{Ray}, 0) = 600,000$ ,  $L(\text{Sam}, 0) = 5400$ , and  $v = 600$ . Then if Sam were neglected the odds, from (5), that  $V=\text{Ray}$  would be 1000:1, but taking Sam into account per (6) this tenfold smaller, 100:1. The expected number of misidentifications, per (2), is correspondingly ten times larger.

### 3.3 Multiple related victims found

When several bodies are found that are suspected of being members of the same family (**Table 1.vw**) and are to be identified with reference to other, living, family members, a seemingly attractive plan is to assign the identities one at a time and let each victim identity once established participate in the identification of the subsequent bodies [3]. A serious objection to this serial approach is that it may squander a lot of the potential power of the evidence by failing to utilize the DNA profile of each victim as evidence to identify the other. As an extreme example, imagine a father and daughter as the only two related victims of a small airplane crash. The two of them can probably be picked out and therefore identified from their DNA similarity even if *no* reference relatives are available, so simultaneous consideration of their types is almost infinitely better than serial identification in this case. In general, the serial method misstates and tends to understate the true value of the evidence. It is preferable to consider all the identifications simultaneously and thereby use the fact that each dead body's identity is supported by its DNA similarity to the other dead bodies.

Example

Hypothesis $H_i, i=$	$\nu w$	$\nu$	$w$	$s$	$0$
	V=Sue, W=Joe	V=Sue, W unrelated	V unrelated, W=Joe	V&W sibs but not of F	V, W, F all unrelated
					
$L(i, 0)$ , LR favoring $H_i$ over unrelated, $H_0$	3,867,000	96.1	125.4	7406	1

correct analysis – simultaneous consideration of all data to evaluate both identities

$p_i$ = prior likelihood*	1	19	19	2**	359
$w_i$ = posterior likelihood†	3,867,000	1826	2382	14812	359
$W_i$ = posterior probability‡	<b>99.5%</b>	0.05%	0.06%	0.4%	0.01%

incorrect analysis – identify victims serially using only presumably known identities at each stage

$p_i$ = prior likelihood <sup>⌘</sup>		19	Evaluation of V=Sue (using reference F, ignoring W)	361 <sup>§</sup>
$w_i$ = posterior likelihood†		1826		361
$W_i$ = posterior probability‡		<b>83.5%</b>		16.5%
$p_i$ = prior likelihood <sup>⌘</sup>	1	19	Evaluation of W=Joe assuming (?!) V=Sue	
$w_i$ = posterior likelihood†	3,867,000	1826		
$W_i$ = posterior probability‡	<b>99.95%</b>	0.05%		

\* Prior probabilities – (1/20)(1/20), (19/20)(1/20) etc. assuming 20 missing children of each sex – are each multiplied by 400 for convenient comparison.

\*\* assuming 2 other missing brother-sister pairs

†  $w_i = L(i,0) \times p_i$ .

‡  $W_i = w_i / \sum_i w_i$ .

⌘ prior probabilities of 1/20, 19/20 scaled by factor of 19·20 for ease of comparison.

§ The difference between 361 here and 359 above comes from ignoring  $H_s$ .

⌘ prior probabilities of 1/20, 19/20 scaled by factor of 20 for ease of comparison.

**Table 1** Hypotheses to consider in identifying bodies of putative siblings. One or both of the two bodies, V and W, may represent missing children Sue and Joe of the typed reference father F.

**Table 1** shows a realistic situation where simultaneous consideration of the identities of two victims is important. Two bodies V and W may be the missing daughter and son of the living reference F (hypothesis  $H_{\nu w}$ ; **Table 1.vw**), and in addition there are at least four other possibilities (**Table 1**) to consider.

The general method of attack for these problem is to compute and compare the likelihoods  $L_i$  of the DNA conditional upon each of the hypothetical pedigrees  $H_i, i \in \{vw, \nu, w, s, 0\}$ .

$$L_i = \text{Prob}(\text{DNA as observed} \mid H_i).$$

The ratio  $L(i,k) = L_i / L_k$  is then the likelihood ratio by which the DNA supports  $H_i$  over  $H_k$ .

Note that  $L(i,k)$  can equally be expressed as  $L(i, 0) / L(k, 0)$ . Therefore, and since likelihoods are never used except in ratios, there is no harm in identifying the likelihood ratios  $\{L(i, 0)\}$  as the likelihoods  $\{L_i\}$ , and it is convenient to do so.

*Theory – simultaneous and serial identifications*

The likelihood ratio  $L$  supporting **Table 1.vw** – identification of both V and W – is at least the ratio  $L_{min} = \min_i \{L_{vw}/L_i\}$  by which this explanation exceeds any other, i.e. exceeds whatever is in second place. However, if the second-best explanation is inherently implausible then  $L_{min}$  is needlessly conservative. The accurate expression for  $L$  would involve some weighted average of the  $L_i$  in the denominator, the weights depending on the respective prior probabilities of the various alternatives  $H_i$ . Since therefore consideration of prior probabilities is unavoidable, we may as well use them to simply calculate the posterior probability of each hypothesis using Bayes’ Theorem. The tableau arrangement of **Table 1** illustrates the calculation: Assuming that relative prior probabilities (“likelihoods”) are given, multiply by the corresponding genetic likelihoods  $L_i$  to obtain the posterior likelihoods. These in turn are converted to posterior probabilities by scaling them to sum to 100%. Under the given assumptions it is thus seen to be 99.5% that V is Sue and W is Joe. (The identity of either body alone is slightly higher.)

Consider by contrast how the serial approach would work in this case. If we choose to identify V first, then we compare the hypothesis  $H_v$  (that V=Sue) with  $H_0$  (V≠Sue), as shown in **Table 1**. The conclusion that V=Sue is a modest 83.5%, much weaker than the actual value of the evidence. Moreover, to proceed to the next step, using the “established” identity of V to identify W, requires an illogical leap from 83.5% to 100%. Assume that leap is made. Then identifying W amounts to comparing the hypotheses  $H_{vw}$  (V=Sue and W=Joe) with  $H_v$  (V=Sue, W unrelated). The resulting confidence of 99.95% is of course an exaggeration, mainly because it rests on an illogical leap, partly because the best alternative explanation  $H_s$  (V & W siblings unrelated to F) is overlooked by the serial method.

*3.4 Appropriate reference population*

Thai	38%
US and European	57%
Chinese	2%
Japanese	1%
Arab	0.7%
Malay	0.3%
African	0.3%
Indian	0.2%
Korean	0.1%
Vietnamese	0.1%

**Table 2** Approximate victim proportions

Those who were killed in Thailand by the tsunami were about equally Caucasians and Thai, with many other racial groups represented in small proportions (**Table 2**, derived from <http://missingpersons.or.th/index.en.html>). An obvious question of practical importance: In making probability calculations is it reasonable to ignore completely those populations are only scantily represented? Practically speaking some sort of compromise is inevitable; just as is true in the forensic setting it is always impossible to characterize, let alone to obtain, the precisely appropriate reference population sample.

This section discusses the method and the mathematics to justify considering only the two major constituent victim races.

*Analysis for direct id*

Suppose a direct match between a body V and some personal effect belonging to a victim Victor. The likelihood ratio (known as “matching odds” in this context) between the hypotheses:

$H_p$ : V is Victor

$H_0$ : V is unrelated to Victor

is then  $L=1/Pr(M)$ , where M is defined as the event that the DNA profile of a randomly selected unrelated body would match that of Victor. The criterion (2) for declaring an identification takes the form

$$\Pr(M) < \vartheta, \text{ where the threshold probability } \vartheta = \vartheta(v) = \epsilon/v \quad (7)$$

and the corresponding contribution to the expected number of misidentifications is  $\Pr(M)v$ .

The probability to see any particular DNA profile at random varies from population to population. Therefore  $\Pr(M)$  is a weighted average

$$\Pr(M) = \sum_r w_r P_r, \text{ where } \sum_r w_r = 1 \quad (8)$$

of the probabilities  $P_r = \Pr(M|\text{race}=r)$  from each constituent victim population. The weights  $w_r$  may be the proportions with which each constituent race is represented among the victims – e.g. if subscripts T and C correspond to Thai and Caucasian,  $w_T \approx w_C \approx 1/2$  – but in principle  $w_r$  may also vary depending on the circumstances and body appearance in a particular case, i.e.  $w_r$  is the prior likelihood, based on any information other than DNA, that the body is of race  $r$ . Therefore it would not be appropriate simply to use average allele frequencies such as might be compiled from the victim profiles.

For the World Trade Center identifications [5], instead of (7) we adopted the more stringent condition

$$P_r < \vartheta \text{ for each } r. \quad (9)$$

Particularly in dealing with tsunami identifications (9) offers two advantages. One, it avoids the need to decide the appropriate weights  $\{w_r\}$ . Two, it is likely to be sufficiently conservative to compensate for the sin of not making calculations for some races that are in fact represented. That is, if (9) is satisfied for the main populations, probably (7) is satisfied. If the vast preponderance of the victims are either Thai or Caucasian, it is not far wrong to make calculations just for those two.

An example will illustrate the reason. Suppose  $w_J = 1/100$  is the prior probability that a given victim is Japanese, and suppose that  $P_T = 10^{-14}$ ,  $P_C = 10^{-17}$ ,  $P_J = 10^{-13}$  – perhaps because V is Japanese. Then by (8),  $\Pr(M) \approx 10^{-14}/2 + 10^{-17}/2 + 10^{-13}/100 \approx 0.51 \cdot 10^{-14}$ , so  $\Pr(M) < P_T$ . Therefore if (9) holds, so does (7) in this case.

In order to investigate the error in general in ignoring minor races, suppose that (9) holds for races T and C, and suppose for example that  $P_T = \max(P_T, P_C)$ . Then the relative error in ignoring minor races is the relative difference,  $\Pr(M)/P_T - 1$ , between  $P_T$  and  $\Pr(M)$ . It can be written as a weighted average of the relative errors that would be incurred by substituting  $P_T$  for each of the  $P_r$ ,  $r=T, C, 3, 4, \dots$ :

$$\Pr(M)/P_T - 1 = w_C(P_C/P_T - 1) + w_3(P_3/P_T - 1) + w_4(P_4/P_T - 1) + \dots \quad (10)$$

If this expression is non-positive, it is harmless to ignore the minor races for calculation. If it is positive, then it is the relative margin by which  $P_T$  must be less than  $\vartheta$  in order to achieve the target expected misidentification contribution despite ignoring minor race calculations. The first term on the right is non-positive, and although the proper choice of weights  $w_i$  may be case-specific, at least from the census data  $w_C$  is by far the largest of the coefficients. The next term is positive only if  $P_3 > P_T$  (an obvious prerequisite for ignoring race 3 to be a mistake). To analyze the formula further it is helpful to understand the behavior of  $P_r/P_s$ . First, note that it is the likelihood ratio supporting race  $r$  over race  $s$  as the origin of the profile. Therefore usually  $P_r/P_s > 1$  when the profile actually comes from population  $r$ . Second, I have calculated "typical" (expected in the geometric sense) values of this ratio comparing various populations among various populations of interest (**Table 3**).

	Chinese	Thai	Japanese	Indian	Caucasian	African-Am	Mozambican
Chinese		1.4 (×3.5)	3.5 (×5.9)	21 (×11)	130 (×27)	750 (×34)	4900 (×67)
Thai	1.8 (×4.5)		8.5 (×11)	16 (×12)	140 (×32)	920 (×33)	8000 (×72)
Japanese	2.5 (×5.3)	3.8 (×6.1)		14 (×9)	100 (×22)	790 (×33)	1900 (×45)
Indian	18 (×16)	8.5 (×11)	23 (×19)		16 (×15)	180 (×27)	480 (×45)
Caucasian	220 (×32)	98 (×20)	380 (×42)	29 (×13)		48 (×13)	400 (×32)
African-Am	1300 (×63)	1900 (×82)	6900 (×150)	570 (×56)	84 (×39)		1.9 (×6.1)
Mozambican	9400 (×84)	18000 (×120)	23000 (×170)	2000 (×71)	980 (×67)	3.2 (×5.4)	

**Table 3** Typical likelihood ratio between races.

$D(r, s)$  supporting the true origin ( $r$ =row) of a DNA profile, versus an alternative origin ( $s$ =column). Factors in parentheses are the calculated standard deviation. Example: the Japanese/Thai notation  $3.8 (\times 6.1)$  means that 68% of Japanese profiles would be between  $3.8 \times 6.1$  and  $3.8 \div 6.1$  times rarer among Thai than among Japanese.

Inter alia these data tend to refute the canard that “Forensic STR markers are very poorly suited to the task of describing ancestry [19].” DNA forensics has advanced considerably in the eight years since the research [15] that the quoted statement can be traced back to [20].

Note from **Table 3** that the East Asian allele frequencies are very similar between ethnicities (including Korean and Vietnamese, not shown). Therefore let’s group all the East Asian populations other than Thai into a single category with  $r=3$ , and for the moment think of  $r=4$  as a mixture of the remaining populations. (10) is thus truncated at three terms.

If the victim race is 3, East Asian – think Japanese for illustration – then typically  $P_3$  is only a few (e.g.  $D(3,T)=3.8$ ) times larger than  $P_T$ , but many times larger than  $P_C$  ( $D(3,C)=100$ ). Therefore we have  $P_C/P_T = D(3, T)/D(3, C) \approx 3.8/100$  – hence  $(P_C/P_T - 1)$  is typically not much different from  $-1$  and similar considerations show even more strongly that the last term,  $w_4(P_4/P_T - 1)$ , is very unlikely to be positive. Therefore the right hand side of (10) is about  $-w_C + w_3P_3/P_T$  or less, which is safely negative provided that

$$P_3/P_T < w_C/w_3. \tag{11}$$

Now, the right side of (11) is about 20 to judge from **Table 2**, and the typical value of  $P_3/P_T$  is  $D(3,T)$  – quite a lot smaller according to **Table 3** given the assumption that the victim is East Asian. Therefore (11) probably holds, so not calculating  $P_3$  is unlikely to be a mistake even when it corresponds to the victim’s true race. From simulations, this back-of-envelope estimate is correct about 90% of the time. In 10% of those cases where 3 is the race of origin, (7) is positive either because  $P_3/P_T$  is abnormally large or because  $P_C/P_T$  is nearly 1, but both of these circumstances are very unlikely unless  $P_T$  is particularly small [analysis of simulation data, not shown].

The situation, then, is this: When  $P_T$  is so large that (9) is barely satisfied, then (10) is negative, which implies (7), i.e. there is no harm in ignoring  $P_3$ . When (10) is positive (about 10% among the times that a simulated victim is actually East Asian), it is because all the probabilities  $P_i$  are so small that  $\Pr(M) < \emptyset$  anyway.

On the infrequent occasions that  $P_3 > P_C > P_T$ , again the frequencies are almost certainly all very small.

The same argument easily applies to populations that are not very different from European Caucasian, such as Finns, Arabs, U.S. Hispanic, and Indians. If the prior probability at which the race is represented as a possible identity for the victim is small compared to the “racial distance” (in the sense of **Table 3**) from a reference race (viz formula 11), then there is little need to bother with a calculation for that population.

The argument may be difficult to apply for those populations that are quite distant from both of the main populations – e.g. African and African-American. Fortunately these seem to be (**Table 2**) particularly scantily represented among the victims, which suggests that  $w_C/w_A$  – subscript A for African – is safely larger than  $P_A/P_C$  (vide (11)). On the other hand, the body might be in good enough condition to permit judging African skin color or features. Then  $w_C/w_A$  is not large. But in that case  $v$ , the effective number of victims taking into account the body appearance, will be much decreased, so (7) will be an easy standard to meet.

Nonetheless, it would seem stubborn to announce the identification of an African victim based on calculating only  $P_T$  and  $P_C$  when African-American [9] and African [12, 13] data are readily available. A sensible suggestion is to re-check (9) for the race of identification (or near alternative) before making an announcement.

#### *Analysis for id by relatives*

If only haplotypes are involved, or partial profiles, then the racial ratios and variations are diminished compared to **Table 3**, which gives more comfort to the policy of computing only for the major races. Some kinship calculations are in this category. For example, if a body is identified through comparison with a single parent then usually only one allele per locus participates in the likelihood ratio. On the other hand, if both parents are available as references then the ratios and variations are as large as given by **Table 3**, while unfortunately the likelihood ratio itself is depressed because of large combinatorial factors (e.g.  $1/8pq$  instead of  $1/2pq$  for mother-father vs direct id identification).

## 4. Summary

1. To declare an identification, normally the posterior odds supporting the identification should exceed some threshold value decided by policy, probably a policy with the expectation of no misidentifications in the entire mass identification effort. The policy can allow exceptions but with cognizance that exceptions gradually accumulate as accrued chances of error.
2. The “effective number of victims”  $v+1$ , or equivalently the prior odds  $1/v$  that a given body is a particular missing person, varies from case to case. The factors to consider include location found compared to expected, and physical attributes observable in the body (including non-DNA expertises such as dental evidence).
3. The confounding effect of related victims is a constant concern. The emphasis in this paper and the greatest concern is when the identification uses living relatives as references rather than direct references but the principle is the same either way. The unknown extra possibilities diminish the confidence of a possible identification.

When a direct reference is available for a missing person but there is no reference for a missing relatives of that person, the situation is logically similar to the forensic situation when a suspect tries to explain away his DNA at a crime scene by suggesting that his (untested) brother or other relative is the culprit. Therefore recommendations 4.2 and 4.4 of the NRC report [21] on forensic DNA evidence would be appropriate.

4. When several possibly related bodies are to be identified using kin references, all reasonable combinations of identities should be enumerated [4] and a calculation made for each combination, so that the identities are evaluated simultaneously.
5. The threshold value for identification should be met by calculations in the principally represented races of Thai and Caucasian, with a third calculation in the race of assignment if it is different. There is no end to possible complications – for example the reference relatives for a certain missing person might be of various races themselves, implying that the perfect calculation is rather complicated. All models are wrong, but some are useful [22]. There has to be a practical end to the complications we worry about.

## **Acknowledgements**

This paper arose out of discussions on identification strategy during my visit to Phuket, Thailand in March 2005. There were very few DNA profiles available at that time, hence the examples in the text are simulated data.

Thanks to P. Sribanditmongkol and David Gross for helpful information about victim demographics. The practical idea to make an extra calculation in the race of attribution is Jon Davoren's.

## References

- [1] B. Olaisen, M. Stenersen and B. Mevåg, Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nature Genetics* 15 (1997) 402–405.
- [2] J. Ballantyne, Mass disaster genetics. *Nature Genetics* 15 (1997) 329–331.
- [3] C.M. Hsu, N.E. Huang, L.C. Tsai, L.G. Kao, C.H. Chao, A. Linacre J.C.-I. Lee, Identification of victims of the 1998 Taoyuan Airbus crash accident using DNA analysis. *Int J Leg Med* 113(1) (1999) 43–46.
- [4] C.H. Brenner, Kinship analysis by DNA when there are many possibilities. *Progress in Forensic Genetics* 8 (1999) 94–96.
- [5] C.H. Brenner and B.S. Weir, Issues and strategies in the identification of World Trade Center victims. *Theor Pop Bio* 63 (2003) 173–178.
- [6] I.W. Evett and B.S. Weir, *Interpreting DNA Evidence*. Sinauer, Sunderland, MA, 1998.
- [7] B. Rerkamnuaychoke *et al* Rinhachai T, Shotivaranon J, Jomsawat U, Siriboonpiputtana T, Chaiatchanarat K, Pasomsab E, Chantratita W, Thai population data on fifteen tetrameric STR loci – D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, vWA, TPOX, D18S51, D5S818, and FGA. *Forensic Sci Int* (2005) *in press*
- [8] L.H. Seah *et al*, STR Data for the AmpFISTR Identifier loci in three ethnic groups (Malay, Chinese, Indian) of the Malaysian population. *Forensic Sci Int* 138 (2003) 134–7.
- [9] B. Budowle, T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh and K.M. Keys, Population data on the thirteen CODIS core short tandem repeat loci in African Americans, US Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. *J For Sci* 44(6) (1999) , 1277–1286.
- [10] M. Hashiyada, Y. Ikatura, T. Nagashima, M. Nata and M. Funayama, Polymorphism of 17 STRs by multiplex analysis in Japanese population. *Forensic Sci Int* 133 (2003) 250–3.
- [11] B. Budowle, B. Shea, S. Niezgoda and R. Chakraborty, CODIS STR loci data from 41 sample populations. *J For Sci* 46(3) (2001) 453–489.
- [12] C. Alves, L. Gusmão, A. Damasceno, B. Soares and A. Amorim, Contribution for an African autosomic STR database (AmpFISTR Identifier and Powerplex 16 System) and a report on genotypic variations. *Forensic Sci Int* 139 (2004) 201–205.
- [13] S. Beleza, *et al*, 17 STR data (AmpFISTR Identifier and Powerplex 16 System) from Cabinda (Angola). *Forensic Sci Int* 141 (2004) 193–196.
- [14] C.H. Brenner, Symbolic kinship program. *Genetics* 145 (1997) 535–542.
- [15] C.H. Brenner, Difficulties in estimating ethnic affiliation. *Am J Hum Gen* 62 (1998) 1558–1560.
- [16] C.H. Brenner, Probable Race of a Stain Donor, in: *Proceedings from the Seventh Human Identification Symposium*, Promega Corp, 1997, pp. 48–52.
- [17] B. Rannala and J.L. Mountain, Detecting immigration by using multi locus genotypes. *Proc Natl Acad Sci U S A* 94(17) (1997) 9197–9201.
- [18] B. Leclair Frégeau CJ, Bowen KL, Borys SB, Elliott J, Fournery RM, Enhanced kinship analysis and STR-based

DNA typing for human identification in mass disasters. Progress in Forensic Genetics 8 (1999) 91–93.

[19] M. Shriver, T. Frudakis and B. Budowle, Getting the science and the ethics right in forensic genetics. Nature Genetics 37 (2005) 449–450.

[20] T. Frudakis, *et al*, A classifier for the SNP-based inference of ancestry. J For Sci 48(4) (2003) 771–782.

[21] National Research Council, The Evaluation of Forensic DNA Evidence, J Crow editor, National Acad Press, Washington DC, 1996.

[22] G.E.P. Box, Robustness in scientific model building, in: R L Launer & G N Wilkinson (Eds), Robustness in statistics, Academic Press, New York, 1979.

## Appendix

Suppose one body is found, and from the DNA there are two possible identities to consider, R and S. This situation might arise if R and S are brothers and the identification is through comparison with relatives as discussed in the text.

Let

- E be the actual DNA type observed in the body
- M be the event that the body have DNA profile E
- $H_R$  mean “R is the name of the body”; prior probability  $\Pr(H_R) = p_R$
- $H_S$  mean “S is the name of the body”; prior probability  $\Pr(H_S) = p_S$
- $H_0$  mean neither  $H_R$  nor  $H_S$ ; prior probability  $\Pr(H_0) = 1 - p_R - p_S$
- $\Pr(M)$  mean the probability of M prior to DNA testing.

$$\begin{aligned} \text{Now, } \Pr(M) &= \Pr(H_R \& M) + \Pr(H_S \& M) + \Pr(H_0 \& M) \\ &= p_R \Pr(M|H_R) + p_S \Pr(M|H_S) + (1 - p_R - p_S) \Pr(M|H_0). \end{aligned} \quad (12)$$

Also,

$$\begin{aligned} \Pr(H_R \& M) &= \Pr(M) \cdot \Pr(H_R|M), \text{ hence} \\ \Pr(H_R|M) &= p_R \Pr(M|H_R) / \Pr(M) \end{aligned} \quad (13)$$

(Bayes’ Theorem) and from (12) and (13),

$$\Pr(\text{not } H_R|M) = [p_S \Pr(M|H_S) + (1 - p_R - p_S) \Pr(M|H_0)] / \Pr(M),$$

so the posterior odds  $\text{Odds}(H_R|M) = \frac{P(H_R|M)}{P(\text{not } H_R|M)}$  favoring  $H_R$  can be calculated

$$\begin{aligned} \text{Odds}(H_R|M) &= \frac{p_R P(M|H_R)}{p_S P(M|H_S) + (1 - p_R - p_S) P(M|H_0)} \\ &= \frac{L(R,0)}{(p_S / p_R)(L(S,0) - 1) + (1 - p_R) / p_R}, \text{ where } L(x,0) \text{ denotes the likelihood ratio} \\ &\quad \Pr(M|H_x) / \Pr(M|H_0) \text{ by which the genetic evidence M favors } H_x \text{ over } H_0; \\ &\approx \frac{L(R,0)}{(p_S / p_R)L(S,0) + 1 / O_R} \end{aligned} \quad (14)$$

where  $O_R = p_R / (1 - p_R)$  is the prior odds of  $H_R$ .