Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Population Biology 63 (2003) 173–178
Theoretical Population Biology
http://www.elsevier.com/locate/ytpbi

Issues and strategies in the DNA identification of World Trade Center victims

C.H. Brenner[a,*] and B.S. Weir[b]

[a] 6568 Sobrante Road, Oakland, CA 94611-1123, USA
[b] Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, NC 27695-7566, USA

Received 11 November 2002

Abstract

Identification of the nearly 3000 victims of the World Trade Center attack, represented by about 15,000 body parts, rests heavily on DNA. Reference DNA profiles are often from relatives rather than from the deceased themselves. With so large a set of victims, coincidental similarities between non-relatives abound. Therefore considerable care is necessary to succeed in correlating references with correct victims while avoiding spurious assignments. Typically multiple relatives are necessary to establish the identity of a victim. We describe a 3-stage paradigm—collapse, screen, test—to organize the work of sorting out the identities.

Inter alia we present a simple and general formula for the likelihood ratio governing practically any potential relationship between two DNA profiles.

1. Introduction

The list of people who died in the World Trade Center (WTC) disaster of September 11, 2001 numbers 2792 and five times that number of body parts have been recovered. Even where physical characteristics can be discerned there is great danger of ambiguity—an incorrect identification of a firefighter was announced based on a gold necklace and a rare, but not rare enough, congenital malformation of a neck bone (NY Times, November 28, 2001). Consequently, DNA has been by far the most reliable tool for establishing identity. DNA profiles have found increasing use as a means to identify remains after war or mass disasters. Successful identifications have been made in recent years following aircraft crashes (Ballantyne, 1997; Olaisen et al., 1997; Brenner, 1999; Leclair et al., 1999; Goodwin et al., 1999), and for misplaced crematory corpses (T. Parsons, pers. comm.). Projects to identify war victims in mass graves in Bosnia–Hertzegovena (Vastag, 2002), and WWII Japanese soldiers in Russia (K. Tamaki, pers. comm.) are under way. Probably the largest and most complex single disaster site to date, however, is that of the attack and collapse of the WTC Towers in New York. This paper focuses on the statistical, combinatorial, and population genetic issues faced in this very large task. Direction and responsibility for the DNA identification effort belongs to the Forensic Biology Department of the New York Office of the Chief Medical Examiner, headed by Dr. R. Shaler, with assistance from various outside companies for the bulk of the DNA typing, for custom software, and from consultants including an advisory Kinship and Data Analysis Panel (KADAP). Administrative matters, including sample tracking, coordination with vendors, families, and multiple computer databases, which we do not discuss in this paper, may well constitute an even more complicated task than the technical and theoretical issues that we do consider. DNA profiles have been generally those of the CODIS core set of 13 microsatellite, or short tandem repeat (STR), loci that are used routinely for forensic purposes in the United States (e.g. Budowle and Moretti, 1999).

Identification issues can be formulated in the standard language of forensic science (e.g. Evett and Weir, 1998). In each instance there are alternative propositions such as:

H_1: This WTC sample is from victim X.
H_0: This WTC sample is unrelated to victim X.

*Corresponding author.
E-mail address: chb@dna-view.com (C.H. Brenner).

0040-5809/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0040-5809(03)00008-X

Before the DNA evidence is examined, there is some prior probability that H_1 is correct. This may plausibly be taken as $1/(v+1)$ if there are $v+1$ victims, which is to say the prior odds on H_1 are $1/v$. The strength of the DNA evidence E is captured by the likelihood ratio

$$m = \frac{\Pr(E|H_1)}{\Pr(E|H_0)}$$

and Bayes' theorem provides posterior odds on H_1 of m/v.

Early on in the investigation, in October 2001, KADAP adopted the view that there are two ways that a DNA match can occur. A body part is identified if it bears a sufficiently persuasive similarity either to a personal or direct reference—a known biological relic of a victim himself (80% of the recovered victim DNA profiles are male) mostly (78%) obtained from toothbrushes but also including hair (11%) or razors (9%), etc.—or to kin (indirect reference).

In the direct matching case, if it is decided to declare identification when m exceeds some minimum value M, then the posterior probability of correctly identifying all $v+1$ victims is $[m/(m+v)]^{v+1} \approx 1 - v^2/m$. For this probability to be 99.9% for 1000 victims it would be necessary for M to be 10^9 and this is well within the limits of the CODIS system if all 26 alleles are assumed to be independent. For kin cases, practicality suggests a less ambitious standard. A goal instead of 99.9% confidence for each kinship case—attained with a likelihood ratio of 3×10^6, assuming 1/3000 as the prior probability—seems reasonable. Admittedly, these rules suggested by KADAP, beg the question of what threshold applies for a combined identification. It turns out that for some victims 3×10^6 can be achieved only by a combination of kinship and direct reference.

Practical difficulties in making the identifications abound, including incorrectly labeled references (toothbrushes are shared, relationships are confounded). There is no need to belabor those issues here.

2. Assigning identities

The process of assigning identities has three stages: Collapsing, Screening, and Testing.

2.1. Collapsing

The Collapsing stage consists in associating like profiles in order to condense the amount of data. One interesting question that arises is the likely number of victims represented. The number of recovered body parts that produced any DNA typing result to date is 14,249; initial typing was unsuccessful on a further 6000 samples. An upper limit can be computed by assuming that s samples are collected at random, with replacement, from $v+1$ victims. The probability to sample any given victim would be $\{1-[1-1/(v+1)]^s\} \approx 1 - e^{-s/(v+1)}$. Using $v+1 = 2792$ and $s = 14,249$ the expected number of victims represented would be all but 17 of the total. This model, which ignores the DNA data, has the virtue of taking into account the thousands of null or highly deficient DNA profiles. Of course this is an unrealistic approach; the sampling is surely not random. Indeed, one-twelfth of the identified victims are represented by 10 or more pieces, accounting for nearly half of the pieces that have been identified. Under a Poisson distribution, only 1% of the victims would be represented by 10 or more pieces, and that would account for only 1% of the pieces. Somewhat more plausible would be to look at the DNA results and assume that any profile except one that is a proper subset of another represents an additional victim. Vacuous and highly deficient profiles thus contribute nothing to the tally, but still this rule results in 3205 profile bins which is obviously an overestimate.

Many of the profiles are "full"—13 STR loci—or nearly so, with typical probability under 1/13 per locus against matching between unrelated people. Such high-likelihood samples can easily be categorized into separate victim identities. The number of unambiguously distinct profiles at present is 1487; this is a lower bound for the number of distinct victims represented.

The uncertainty is because many of the profiles are partial. To refine the estimate further we construct a probabilistic approach to binning. Several complexities arise. Suppose the alternative propositions for two WTC samples are

H_1: These two WTC samples are from the same victim.
H_0: These two WTC samples are from different victims.

At loci for which both samples are typed and the genotypes are equal, the likelihood ratio m is the reciprocal of the genotype probability. At loci for which only one sample is typed, $m = 1$. At loci for which both samples are typed but the genotypes do not match, setting $m = 0$ is not tenable because of allelic dropout with degraded samples. Table 1 shows examples of this phenomenon: a genotype 7,8 may be observed as 7—or, less frequently, as 8—in a degraded sample. Hence a correctly evaluated m is non-zero between any pair of these three possibilities. As a very rough shortcut, we assign $m = 0.5$ in each such case. A more accurate computation of m would include a profile-specific estimate of the probability of one allele dropping out at a locus—a complicated question that depends on the number of loci altogether absent from the profile, and less obviously—but judging by the frequent occurrence of partial profiles that have only one allele at every locus—by homozygous appearance at other loci.

Table 1
Three disaster profiles, probabilistically representing 1.48 victims

Profile	D3	VWA	FGA	D8	D21	D18	D5	D13	D7	D16	TH01	TPOX	CSF
1	14	16	22	13	28	12	11	8	11	11	**7**	8	11
	18	17	24	13	29	13	12	14	12	13	**8**	8	12
2	14	16	22	13	28	12	11	8	11		**8**	8	11
	18	17	24	13	29	12	12	14	12		**8**	8	12
3	14	16		13			11				**7**		
	18	17		13			12				**7**		

The first is identical to that from a toothbrush of victim V. The second must be from victim V as well. The third is quite ambiguous ("D3," etc. abbreviate STR locus names).

Assuming that the likelihood ratio m for a common origin of two samples is computed correctly, two samples randomly selected from among $v+1$ victims represent the same person with probability $m/(m+v)$ and different people with probability $v/(m+v)$. Hence the expected number of people that they represent is $1+v/(m+v)$. Calculating based on this idea suggests that at least a partial DNA profile has been obtained for about 2100 victims.

2.2. Screening

Screening refers to the stage of comparing every victim profile (in the collapsed list) with every reference profile using a heuristic that must be very rapid—even at some expense in accuracy—and which produces a list, mostly correct, of tentative victim identities within a reasonable amount of time. Millions of comparisons are involved: 10,290 reference profiles (representing altogether 2652 victims—2366 by on average 2.9 relatives, 2234 by on average 1.6 direct references) times 3205 distinguishable victim profiles. Picking matches to direct references is not usually a problem, but picking relatives can be difficult.

An obvious approach would be to compute a likelihood ratio between every reference profile and every victim profile based on the putative relationship that the reference holds to its corresponding victim. The reference–victim pairs are then listed by descending likelihood ratio, and at least the pairs near the top of the list should represent true identities.

For two reasons this approach is not very effective. First, there is a very large variety of claimed relationships between surviving family members and victims, and experience soon showed that the records of claimed relationships can be unreliable. Therefore only three relationships are considered for screening: parent–child, sibling, and identity. These relationships are assumed to be an adequate surrogate for the others, and all three are computed for every reference–victim pair.

To that end, for any pair of relatives write P_0, P_1, P_2 for the probabilities that they share 0, 1, or 2 pairs of alleles identical by descent. These three probabilities sum to one, and for siblings $P_0 = 0.25, P_1 = 0.50, P_2 = 0.25$. If two genotypes with this suspected relationship are ab and cd, define variables u_i for the four mating combinations $1:ac; 2:ad; 3:bc; 4:bd$. If the two alleles in the ith pair are the same type, then u_i is the reciprocal of the frequency of that allele. Otherwise, $u_i = 0$. Let U be the average $(u_1+u_2+u_3+u_4)/4$ and W be the average between-individual product $(u_1u_4+u_2u_3)/2$. Then the likelihood ratio for the propositions that the individuals have the stated relationship versus they are unrelated is

$$m = P_0 + UP_1 + WP_2. \quad (1)$$

This expression has an obvious similarity to the two-allele I, T, O method of Li and Sacks (1954), and it can be derived from the treatment for non-inbred relatives given by Evett and Weir (1998). It has the computational advantages of a single compact equation for any pair of genotypes and any relationship, and of efficiently computing several relationships at once.

For the parent–child case, $P_1 = 1$, but instead of $m = U$ from Eq. (1) we put $m = \max(U, \mu)$ where μ is some expression that takes into account mutation. For screening purposes an adequate approximation is to let μ be a constant for each locus. See http://dna-view.com/mutext.htm

Second is the problem of false positives. With so large a number of victims, many unrelated pairs have higher likelihood ratios for relationship than do many related pairs. Table 2 illustrates the magnitude of the problem in the case of the sibling relationship. The first three columns were estimated by simulation. Siblings were created by random mating from parents generated by selecting alleles at random according to the frequencies of the OCME Caucasian population study for CODIS loci (data available at http://dna-view.com/ocme/). In order to include even 3/4 of the true sibling pairs it is necessary to consider all pairs with a likelihood ratio at least 100. However, since one of every thousand false sibling pairs also has such a likelihood ratio and the number of victims is nearly 3000, false siblings with

Table 2
Attribution of siblingship by various likelihood ratio values

Threshold for consideration m at least	True positive rate	False positive rate	Number of false positives per true positive[a]	
			WTC ($v = 3,000$)	Air crash ($v = 200$)
20000	0.3	1/200 000	0	0
2000	0.5	1/30 000	0	0
500	0.6	1/8000	1	0
100	3/4	1/1000	4	0
14	7/8	1/300	11	1
3	15/16	1/80	38	3
1	31/32	1/40	73	5
1/8	99/100	1/10	286	20
1/30	299/300	1/6	473	33
1/100	999/1000	1/3	944	Not relevant
1/1000	4999/5000	1/2	1415	Not relevant

[a] $v\times$ (false positive rate)/(true positive rate).

Table 3
Mutation-tolerant attribution of paternity

No. of matching loci at least	Incidence among		No. of false positives per true positive	
	Parent/child (%)	Unrelated	$v = 3000$	$v = 200$
13	97	1/1000	3	0
12	99.95	1/100	30	2
11	100	1/20	150	Not relevant

$m = 100$ out-number true siblings by 4:1. True siblings are lost in the noise. And even a true positive rate of 3/4 would not be sufficient; to identify every victim we need somehow to visit even the last line of Table 2.

Parents and children are typically characterized by sharing an allele at each autosomal locus, but so do 1/1000 of random pairs. Table 3 compares calculations (assuming a Bernoulli process for the occurrence of mutations) for parent–child pairs. "Incidence among unrelated" is from simulations. "Incidence among Parent/Child" is calculated assuming 1/400 mutations per locus. With this allowance for the possibility of mutation between parent and child, then false positives outnumber true positives by 30:1 when $v = 3000$.

In order to surmount the problem of drowning in false positives, a new strategy is necessary. The approach that has proven useful is to "triangulate"—to look for potential victim-family associations that are indicated by at least two members of the same family. Finding a formula like Eq. (1) for three-person relationships is too difficult, but an ad hoc alternative statistic can be used instead. For each family i and victim profile j, a score $S_{i,j}$ is computed as the largest product of any two relationship indices between distinct family members and j, or the largest single index. Then for family i the score $S_i = \max_j S_{i,j}$, if it is large, points to a good candidate victim. In order to bring the easiest cases to earliest attention, the right strategy is to sort the families by S_i.

A final point about screening is that in order to utilize both direct and kin references simultaneously, the direct references are considered simply as a particular kind of relative, with $P_2 = 1$. As in the Collapse stage, the computation of matching is slightly modified from the nominal $m = W$ in order to allow for allelic dropout.

2.3. Testing

When a victim sample is tentatively associated with family, it is usually necessary to perform a confirming computation. The example of Table 4 shows a candidate victim that was suggested by screening by the combination of toothbrush–victim matching odds of 30,000 and mother–victim parentage index of 20,000, i.e. by $S = 6 \times 10^8$. In those profiles, the Amelogenin locus indicates gender: x for female and xy for male.

Confirmation is of necessity case-by-case. In this case, the evidence for those loci for which the toothbrush yielded typing is the toothbrush profile matching odds. For the remaining loci the kinship likelihood ratio (Brenner, 1997), evaluating the sister–mother–victim trio, comes to 62,000. Multiplying the two odds ratios

Table 4
Genotypes (indicated symbolically) for a tentative identification of victim V suggested by screening, and confirmed by explicit computation

Profile	D3	VWA	FGA	Amel	D8	D21	D18	D5	D13	D7	D16	TH01	TPOX	CSF
Sample, possibly V	qr	ps	pr	xy	qr	p	pq	pq	pr	pr	r	pr	qr	qr
Toothbrush of V	qr	ps		xy				pq						
Sister of V	p	pq	pq	x	qr	pq	pr	rq	pq	q	pr	pq	qr	pr
Mother of V	qp	pr	pr	x	qp	pq	qr	rq	pr	qr	qr	pr	pr	r

and assuming a prior probability of 1/2792 gives posterior odds of over 400,000 that the possible victim is correctly identified, assuming that no other relative of the family also perished.

2.4. "Closed system"

The prior odds, initially $1/v$, continually increase as new victim identifications are made. Similarly, but less obviously, if a victim sample resembling family i can be excluded from a number of other families, even unsolved ones, the prior odds can also be adjusted upwards. If all but one family can thus be dismissed from consideration, then the prior odds become infinite for i and declaring identification is as easy as fitting the last piece into a jigsaw puzzle. This "closed system" situation is thus no different in principle from a non-closed system. The apparent difference arises only because prior to achieving a closed system status sometimes the need to consider not only the evidence from the similarity of a victim sample to a particular family, but also from its dissimilarity to other families, is overlooked.

3. Discussion

Probably every mass disaster identification effort reveals new special problems and complications. Airplane crashes are characterized by related people perishing together, and this confounds identification by DNA (Brenner, 1999). The mass graves in Bosnia–Hertzegovena also include related victims, but to some extent can be regarded as a collection of sub-disasters (but not altogether—graves were often moved and co-mingled (E. Huffine, pers. comm.)). Disasters vary greatly in the extent to which physical clues aid identification.

The WTC disaster includes some pairs of relatives among the victims, but related victims was not a salient feature. Nonetheless, it is an issue to be borne constantly in mind in assessing the reliability of any identification through kin. Many of the victims disappeared without a trace. The remainder were typically massively fragmented by the collapse of the towers, then buried for weeks in hostile conditions. In most cases little but DNA can possibly be used to identify them. Without doubt, the most conspicuous unique feature of the WTC situation is the sheer number of victims. As the comparisons with a smaller disaster in Tables 2 and 3 show, the larger number aggravates the problem of distinguishing true from false relatives in several ways. The number of false relatives, at any given likelihood ratio threshold, is proportional both to the size of the reference list and to the size of the victim list—in other words, to the square of the number of victims. Moreover, increasing the number of victims increases the number of victims who coincidentally bear only a modest genetic similarity to their kin references, thus depressing the likelihood ratio threshold that one must consider.

Currently there are efforts underway to use mitochondrial and nuclear single nucleotide polymorphisms (SNPs) on the unresolved WTC samples. Our estimate, 2100, of the number of victims for whom there exists at least a partial DNA profile, is a plausible upper bound for the eventual number of identifications that might be made if these additional technologies are successful. In any case, there is no prospect of attaining a closed-system. However, as we have noted and contrary to previous conceptions, this is not a difference of kind.

We note that Eq. (1) can be interpreted as showing how a large class of two-person likelihood ratios can be viewed as linear combinations of those for just two primitive relationships: $m = W$ for two profiles having the same source, and $m = U$ for two profiles coming from parent and child—the one-parent or "motherless" parentage index (Brenner, 1993).

Acknowledgments

CHB is a member of KADAP and a consultant to the New York OCME for WTC identification. BSW is supported in part by NIH Grant GM 45344 to North Carolina State University. Helpful comments were made by Drs. J. Ballantyne, G. Carmody and R. Shaler.

References

Ballantyne, J., 1997. Mass disaster genetics. Nat. Genet. 15, 329–331.
Brenner, C.H., 1993. A note on motherless paternity case computation. Transfusion 33, 51–54.

Brenner, C.H., 1997. Symbolic kinship program. Genetics 145, 535–542.

Brenner, C.H., 1999. Kinship analysis by DNA when there are many possibilities. In: Sensabaugh, G. (Ed.), Progress in Forensic Genetics, Vol. 8, pp. 94–96. Elsevier Science B.V.

Budowle, B., Moretti, T., 1999. Genotype profiles for six population groups at the 13 CODIS short tandem repeat core loci and other PCR-based loci. Forensic Science Communications 1. (http://www.fbi.gov/hq/lab/fsc/backissu/july1999/budowle.htm)

Evett, I.W., Weir, B.S., 1998. Interpreting DNA Evidence. Sinauer, Sunderland, MA.

Goodwin, W., Linacre, A., Vanezis, P., 1999. The use of mitochondrial DNA and short tandem repeat typing in the identification of air crash victims. Electrophoresis 20, 1707–1711.

Leclair, B., Frégeau, C.J., Bowen, K.L., Borys, S.B., Elliott, J., Fourney, R.M., 1999. Enhanced kinship analysis and STR-based DNA typing for human identification in mass disasters. In: Sensabaugh G., et al. (Eds.), Progress in Forensic Genetics, Vol. 8, pp. 91–93. Elsevier Science B.V.

Li, C.C., Sacks, L., 1954. The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. Biometrics 10, 347–360.

Olaisen, B., Stenersen, M., Mevåg, B., 1997. Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. Nat. Genet. 15, 402–405.

Vastag, B., 2002. Out of tragedy, identification innovation. J. Am. Med. Assoc. 288, 1221–1223.