Original Article

Theoretical value of the recommended expanded European Standard Set of STR loci for the identification of human remains

Q1 William Goodwin and Carole Peel

University of Central Lancashire – School of Forensic and Investigative Sciences, Corporation Street, Preston PR1 2HE, UK
Correspondence: William Goodwin. Email: whgoodwin@uclan.ac.uk

Abstract

We have undertaken a series of simulations to assess the effectiveness of commercially available sets of STR loci, including the loci recommended for inclusion in the expanded European Standard Set, for the purpose of human identification. A total of 9200 genotype simulations were performed using DNA · VIEW. The software was used to calculate likelihood ratios (LRs) for 23 groups of relatives, and to determine the probability of identification given scenarios that ranged between 10 and 250,000 victims. The additional loci included in the recommended expanded European Standard Set, when used in conjunction with the Identifiler® kit, significantly improved the typical LRs for tested scenarios and the likely success of providing correct identifications.

Med Sci Law 2011; 00: 1–7. DOI: 10.1258/msl.2011.011068

Introduction

With the availability of commercial STR kits that allow the typing of multiple loci, the use of DNA analysis has become a powerful tool for the identification of victims of natural disasters, man-made disasters and conflict.[1–4] Currently, kinship testing laboratories typically use two commercially available kits, Promega's PowerPlex®5 and Q2 Applied Biosystems' AmpFℓSTR® Identifiler®,[6] both of which amplify 15 STR loci.

Fifteen STR loci comprise a powerful battery both for parentage testing and for the identification of human remains.[7–11] However, there are circumstances when it is desirable to type more loci, such as standard parentage testing when there has been a mutation, kinship tests where direct relatives (i.e. parents/children) are not available for testing and in victim identification cases where in addition to the need to test complex relationships there is a large number of victim-relative comparisons to be made.[7,9,12,13]

The identification of human remains becomes more complex as the number of victims increases, in particular from open systems such as natural disasters and conflicts. In order to limit the possibility of false identifications it is necessary to apply thresholds to ensure that the strength of the evidence is sufficient. This can be done by applying appropriate prior probabilities, which are normally directly related to the number of victims, so if there are 100 victims the prior probability of a given body being a specified victim is 1/100 or 0.01. As the number of victims increases the posterior probability of obtaining a correct identification decreases unless higher likelihood ratios (LRs) can be obtained. Moreover, in large-scale identification programmes it is desirable to set a threshold for the identification programme as a whole,[9,11] for example, with the identification of victims of the World Trade Center a target LR for each identification was set at 10 billion, which provided a 99% chance of making no errors in 10,000 identifications.[12]

In order to meet the challenges presented by testing for complex relationships laboratories are now able to increase their battery of tests by, for example: using both the PowerPlex® 16 and Identifiler®, which provide a total of 17 loci; by using the FFFL kit (Promega) that types an additional four loci (LPL, F13A1, FESFPS, F13B); by using other commercial kits that incorporate additional loci, such as the AmpFℓSTR® SEfiler™ (Applied Biosystems) and PowerPlex ES system (Promega), which both incorporate the highly polymorphic SE33 locus; and by using their own primer sets for individual loci or multiplexes.[14] In 2008, the European Network of Forensic Science Institutes recommended expanding the current European Standard Set of STR (ESS) loci to include D10S1248, D22S1045, D2S441, D1S1656 and D12S391[15] (from here on referred as e-ESS) – these have now become available in several commercial kits. While these loci were selected to improve the effectiveness of data sharing between national DNA databases in Europe they also make a welcome addition to the STR loci that are readily available for kinship testing.

In this paper, using DNA · VIEW's simulation function,[16,17] we describe the effectiveness of the Identifiler® loci with and without the addition of the additional ESS loci and the FFFL loci, to provide sufficiently powerful matches for the identification of human remains. Identifiler® was also compared with PowerPlex® 16. In total, 23 different sets of relatives have been tested,

including those recommended by the International Society for Forensic Genetics.[11]

Materials and methods

Population databases

An allele reference database was constructed based on US Caucasian data. The database included frequency data for all loci in the Identifiler® loci[18]; Penta D and Penta E[19]; the e-ESS for D12S391, D1S1656, D10S1248, D22S1045, D2S441[20]; and FFFL (data for LPL, F13A1, FESFPS, F13B).[19] The database was imported to DNA · VIEW as described in the user's manual.[17]

LR calculations

LR calculations were performed in DNA · VIEW version 29.13 using the Automatic Kinship feature of DNA · VIEW, as described in the DNA · VIEW manual.[16,17] The relationship scenario to be tested was entered as a series of statements defining relationships. Individuals for whom genotypes were available were represented by a single-letter 'role code'. All un-typed individuals were represented by a name. The LR was calculated by entering a primary and alternative hypothesis; e.g. victim V is the child of M and F, versus another unidentified person is the child of M and F. This would be entered into DNA · VIEW as:

$$\mathrm{V} : \mathrm{M} + \mathrm{F}$$
$$/? : \mathrm{M} + \mathrm{F}$$

A total of 9200 simulations were carried out for 23 relationship scenarios, using four different STR panels: PowerPlex® 16, Identifiler®, Identifiler® + FFFL and Identifiler® + e-ESS loci. DNA · VIEW was set to perform 100 simulations per scenario. The programme simulated genotypes for any individuals who were defined by a one-letter role code. The LR was then automatically calculated. A 'Typical LR' was reported, based on the results of 100 simulations. Individual LR data for each simulation were accessed by returning to the basic menu, selecting 'Freeze the current simulated types and exit simulation' and 'Show (log) report'. The DNA · VIEW report displayed individual LRs within a string of text; a basic Excel tool was created to extract LRs from the reports. Analysis of these data provided a list of the 100 individual LRs for each simulation exercise.

Calculating posterior probabilities

The posterior probability of identification was calculated from raw LR data, using the following formula:

$$\text{Posterior probability} = \frac{(\text{Prior probability} \times \text{LR})}{(\text{Prior probability} \times \text{LR}) + (1 - \text{Prior probability})}$$

Calculating minimum LRs

The minimum LR required to achieve a target posterior probability for all the identifications being correct was calculated using the following formula (adapted from Brenner and Weir (2003)[9]):

$$\frac{V^2}{1 - (\text{Target posterior probability})} \approx \text{Minimum Likelihood Ratio}$$

where V = number of victims.

Table 1 Typical likelihood ratios (LRs) for 23 relationship sets using different STRs

Test	Relatives available for testing	Identifiler	PowerPlex 16	Identifiler + ESS	Identifiler + FFFL
A	1 Sibling	14,000	17,000	670,000	47,000
B	3 Grandparents	63,000	35,000	780,000	110,000
C	1 Parent + half-sibling (opposite side)	2,000,000	24,000,000	1,500,000	2,400,000,000
D	1 Child	97,000	74,000	4,900,000	720,000
E	1 Parent	110,000	690,000	5,700,000	1,200,000
F	2 Grandparents (different sides) + 1 child	3,300,000	12,000,000	960,000,000	96,000,000
G	1 Parent + aunt/uncle (opposite side)	3,100,000	8,500,000	2,400,000,000	140,000,000
H	2 Siblings	30,000,000	15,000,000	11,000,000,000	14,000,000
I	Spouse + 1 child	23,000,000	55,000,000	29,000,000,000	880,000,000
J	1 Parent + 1 sibling	31,000,000	70,000,000	32,000,000,000	1,000,000,000
K	2 Grandparents (same side) + 1 child	2,600,000	230,000,000	61,000,000,000	450,000
L	4 Grandparents	7,600,000	21,000,000	2E + 11	18,000,000,000
M	3 Siblings	1,400,000,000	150,000,000	2.5E + 11	2E + 11
N	1 Parent + 1 child	1,700,000,000	760,000,000	7.1E + 11	2,300,000,000
O	1 Grandparent + 2 children	87,000,000	9.5E + 11	1.7E + 12	7,500,000,000
P	2 Children	190,000,000	540,000,000	3.9E + 13	81,000,000,000
Q	1 Parent, spouse + child	2,100,000,000	53,000,000,000	7.5E + 13	39,000,000,000
R	2 Grandparents (same side) + 2 children	320,000,000	7.6E + 11	1.4E + 14	34,000,000,000
S	3 Children	4,600,000,000	5.5E + 12	4.2E + 14	1.7E + 13
T	1 Parent + 2 children	16,000,000,000	1.6E + 12	8.1E + 15	2.6E + 11
U	Spouse + 2 children	2.4E + 13	3E + 12	1.4E + 16	2.7E + 14
V	2 Parents	5.2E + 11	4.9E + 13	2E + 17	1.7E + 15
W	1 Parent, spouse + 2 children	2.1E + 11	1.9E + 15	1.8E + 18	6.1E + 14

Calculation of typical LRs for macro-generated data

Typical LRs were calculated for each scenario by taking the mean of the log values of individual LRs. The antilog of the mean was the 'typical LR'.

Statistical analysis in R

All statistical analyses were performed using the freely available package 'R' version 2.11.1.[21] All data frames were generated in Microsoft® Excel.

Results

The simulation feature of DNA · VIEW was used to assess the usefulness of testing different relatives, or sets of relatives, for victim identification.[16,17] This was achieved by comparing typical LRs for each scenario. Genotypes were simulated using four different panels of STRs to assess the effect of which and how many loci were tested.

DNA · VIEW simulations

A total of 9200 simulations were carried out in DNA · VIEW to test 23 relationship scenarios. Genotypes were simulated using four different panels of STRs: Identifiler®, Identifiler® plus the e-ESS loci, Identifiler® plus the FFFL loci and PowerPlex® 16. One hundred simulations were performed for each scenario to generate the typical LR. The results are shown in Table 1.

Usefulness of testing different relatives

The results, as expected, show that in general the more relatives that are tested, the higher the LR that can be obtained.

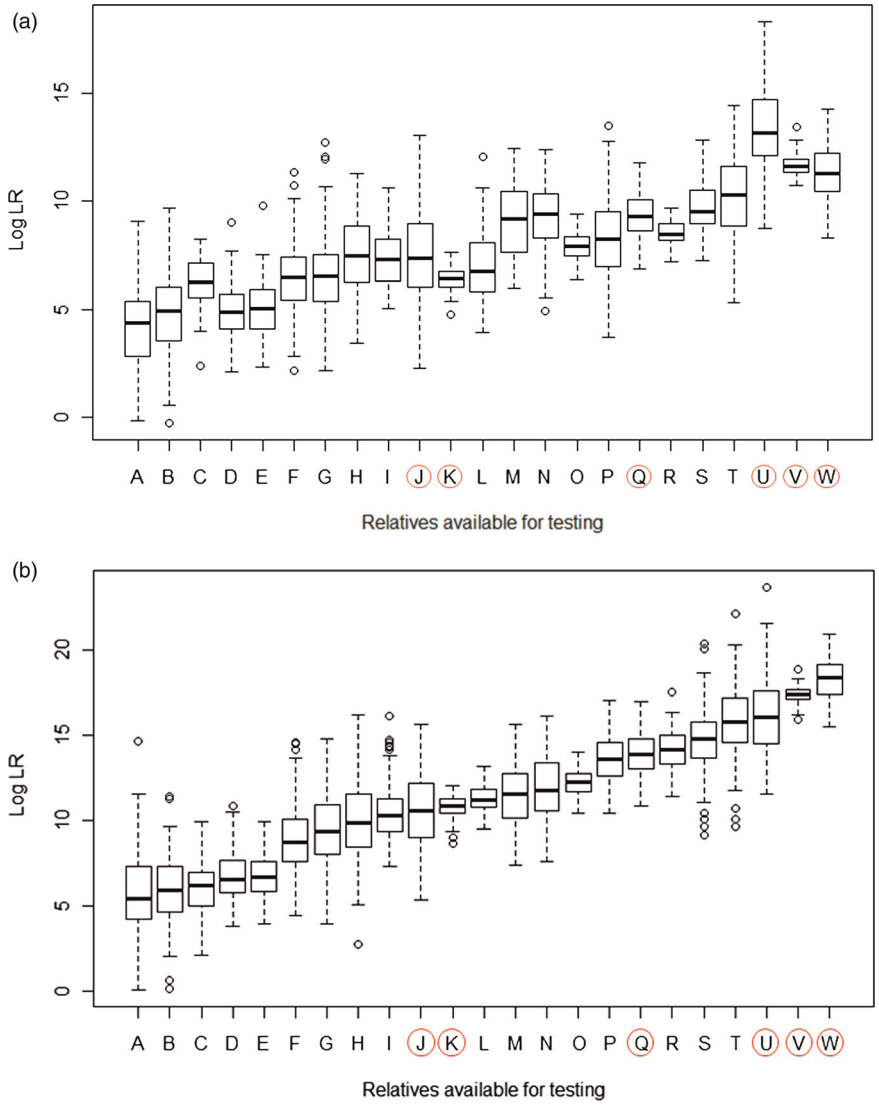

Figure 1 Boxplot comparing the typical likelihood ratios (LRs) of 23 relationship scenarios. The groups of relatives labelled A–W in the box plot are ordered as in Table 1. The plot is based on the data from (a) Identifiler® simulations; (b) Identifiler® plus e-ESS loci simulations. The groups of relatives recommended by International Society for Forensic Genetics are circled. These are: one parent and sibling (J); children and spouse (I = 1 child, U = 2 children); one parent, spouse and children (Q = 1 child, W = 2 children); and two parents (V) Q3

Close relatives such as a parent or child give a higher LR than more distant relatives such as a sibling or grandparent. It was noted that the order of 'usefulness' of each set of relatives was different for each STR panel tested. Some of the results are also anomalous, e.g. Identifiler® data gave a higher typical LR for a spouse and two children than for a parent, spouse and two children. These anomalies illustrate that 100 simulations are not sufficient to give an accurate prediction of the typical LR in all simulations. The individual LR data tend to be distributed over a large range, as shown in Figure 1, and results at the extreme end may skew the data. However, the data provide a very useful overview regarding which relatives are most beneficial to test and the range of values that would be expected.

Significance of using different STR systems

The effect of using different sets of STR loci on individual sets of relatives was assessed using raw LR data. Six relationship scenarios were selected that typically gave

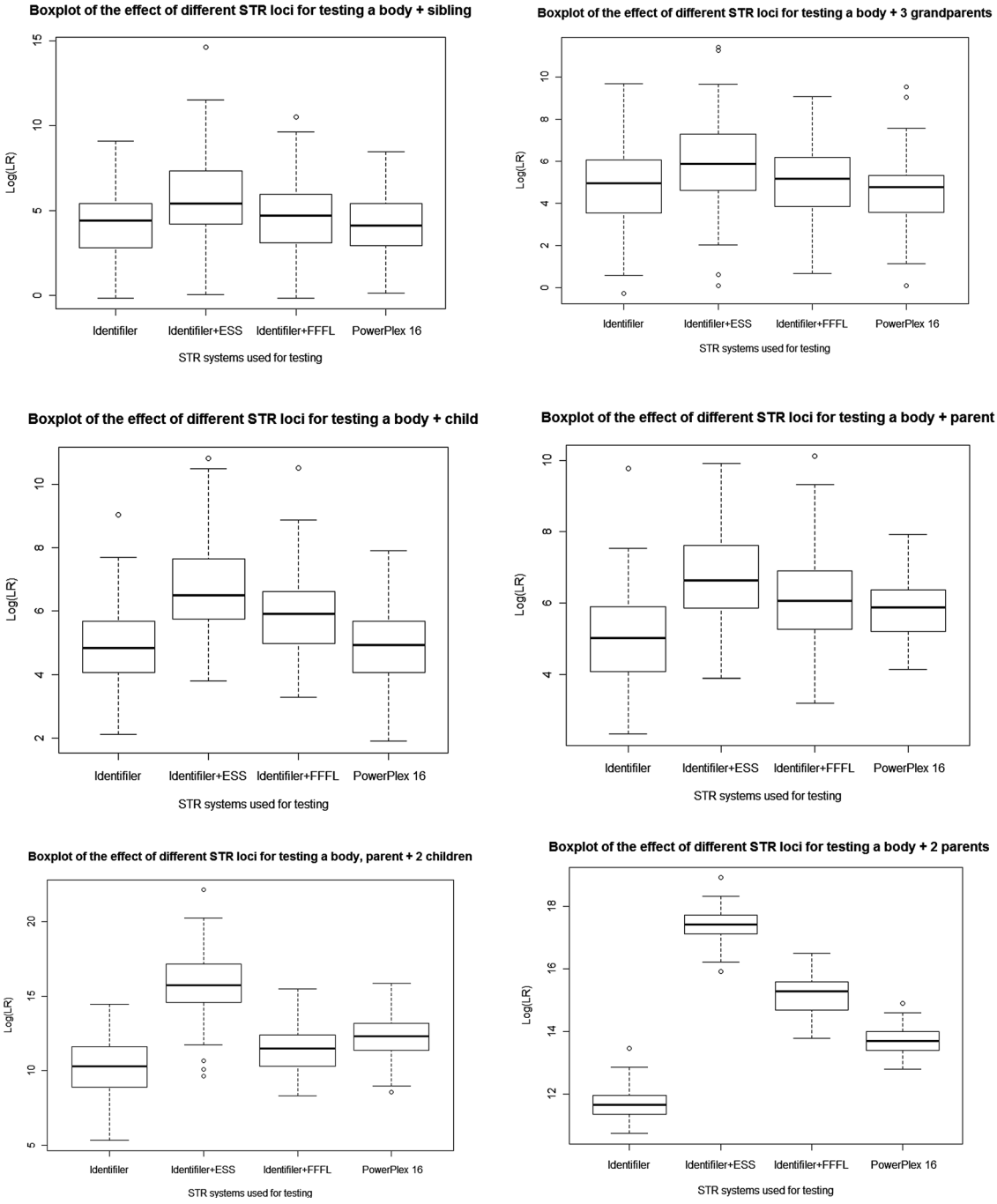

Figure 2 Effect of using different STR systems on LR for selected relationship scenarios. The plots are based on individual LR results from 100 DNA · VIEW simulations for each test

particularly low LRs (1 sibling, 3 grandparents, 1 child and 1 parent) or high LRs (1 parent and 2 children and 2 parents). The analysis was carried out to test whether the expected LRs may be improved by using one combination of loci in preference to another; the results for each scenario are shown in Figure 2. Analysis of variance followed by a Tukey's Honestly significant differences test for all six tested scenarios showed that there were significant differences between the STR loci tested ($P < 0.00138$) (Table 2).

Effectiveness of different STR systems to achieve target posterior probabilities

The LRs shown in Figure 1 and Table 1 provide an indication of the strength of evidence likely with different combinations of relatives. However, prior probabilities have to be considered when dealing with multiple victims. In order to assess the impact of multiple victims, in this study we have assessed five different scenarios with a range of between 10 and 250,000 victims. The prior probability has been calculated by assuming that each set of human remains was equally as likely to be any given one of the victims, therefore the prior probabilities are: 0.1 (10 victims); 0.004 (250 victims); 0.0004 (2500 victims); 0.00001 (100,000 victims); and 0.000004 (250,000 victims).

When identifying human remains following events that have resulted in mass fatalities it is advisable to establish a target posterior probability of obtaining correct identifications for the whole DNA-based identification process.[9,11,13] The target LR that is required to achieve defined posterior probabilities for all identifications is presented for posterior probabilities of 99%, 99.9% and 99.99% (Table 3). The percentage of cases where the target LR was reached when using a posterior probability of 99.9% for all identifications is shown in Table 4. The use of Identifiler® plus the e-ESS loci has a marked effect, with the percentage of cases that would meet the 99.9% threshold increasing from 33% to 61%.

Table 3 Minimum LRs required to achieve a set statistical threshold

	Minimum required LR when the statistical threshold is:		
Number of victims	**99.99%**	**99.9%**	**99%**
10	1,000,000	100,000	10,000
250	625,000,000	62,500,000	6,250,000
2500	62,500,000,000	6,250,000,000	625,000,000
100,000	1E + 14	1E + 13	1E + 12
250,000	6.25E + 14	6.25E + 13	6.25E + 12

The statistical threshold is the minimum posterior probability of correct identification of all the victims

Discussion/conclusions

Under ideal circumstances comparisons would be made between human remains and direct reference samples, which provide extremely high LRs. However, this approach is often hampered by the lack of antemortem DNA records (or cellular material that can be used to generate a DNA profile) and the only option is to use the DNA profiles of relatives for comparison. The simulation exercise has provided a typical LR for 23 combinations of relatives and, equally importantly, provided a range of values that would be expected to be encountered in casework. The data presented are based on Caucasian data, and the values presented would change with different allele reference databases. The only data available for comparison are based on Identifiler® loci with a Hispanic database; the values for comparable relationships are all very similar (values are within 0.1% of each other).

The different STR systems performed as expected in relation to each other. PowerPlex® 16, in most cases outperformed Identifiler®. This is also seen when comparing typical paternity indices, and can be attributed to the Penta D and Penta E loci being more polymorphic than D2S1338 and D19S433 that are found in the Identifiler®. The FFFL loci did improve the LRs, but the e-ESS loci had a greater impact, partially because it comprised an extra five loci, as compared with the four loci in the FFFL, but also because the loci are more polymorphic.[14,19] In real casework the e-ESS loci will be a significant improvement over the FFFL loci as they have been selected in part to work with small PCR amplicons, and therefore are predicted to have a higher success rate with degraded DNA; this has already been demonstrated in multilab trials with DNA recovered from crime scenes.[15]

The simulation function in DNA · VIEW provides a valuable tool to estimate the usefulness of different sets of relatives, and has already been used to provide guidelines to the forensic community.[8,11] In reality, when carrying out casework involving the identification of mass fatalities there may well be several confounding factors that are experienced that are not seen in computer-based simulations, for example: complications over which allele frequency database to use when there are victims of different geographical origins among the dead[7,13]; ascertaining the correct

Table 2 Pairwise comparison of STR multiplexes for six relationship scenarios

Pairwise comparison			1 Sibling	3 Grand-parents	1 Child	1 Parent	1 Parent + 2 children	2 Parents
Identifiler + e-ESS	v	Identifiler	*	*	*	*	*	*
Identifiler + FFFL	v	Identifiler			*	*	*	*
PowerPlex 16	v	Identifiler				*	*	*
Identifiler + FFFL	v	Identifiler + e-ESS			*	*	*	*
PowerPlex 16	v	Identifiler + e-ESS	*	*	*	*	*	*
PowerPlex 16	v	Identifiler + FFFL			*			*

The asterisks indicate pairs of STR multiplexes that produced significantly different LR ($P < 0.00138$). Log LR data from 100 DNA · VIEW simulations were used for each relationship scenario

biological relationship of individuals who are providing reference samples, the obvious example being samples from fathers who are not the actual biological father; the presence of multiple relatives among the victims, the likelihoods provided are based on the alternate hypothesis being that they are unrelated; and the profiles in the simulation are all complete, whereas in casework the degradation of DNA is a problem, with partial profiles often resulting, which will impact on the potential LR that can be obtained. Notwithstanding, the data presented provides a useful reference for organizations that are executing or planning identification programmes with a DNA component, illustrating what can potentially be achieved using different combinations of reference profiles in different scenarios.

Table 4 Percentage of simulated cases resulting in identification using (a) Identifiler® genotypes and (b) Identifiler® + recommended expanded-ESS

	Percentage of samples resulting in successful identification when the number of victims is:				
Relatives available for testing	**10**	**250**	**2500**	**100,000**	**250,000**
(a) Identifiler® genotypes					
1 Sibling	36	5	0	0	0
3 Grandparents	48	4	0	0	0
1 Parent + half-sibling (opposite side)	92	10	0	0	0
1 Child	47	1	0	0	0
1 Parent	52	1	0	0	0
2 Grandparents (different sides) + 1 child	81	23	3	0	0
1 Parent + aunt/uncle (opposite side)	79	19	4	0	0
2 Siblings	86	47	10	0	0
Spouse + 1 child	100	38	5	0	0
1 Parent + 1 sibling	88	42	15	1	0
2 Grandparents (same side) + 1 child	99	0	0	0	0
4 Grandparents	89	32	2	0	0
3 Siblings	100	73	42	0	0
1 Parent + 1 child	99	83	35	0	0
1 Grandparent + 2 children	100	60	0	0	0
2 Children	98	61	22	1	0
1 Parent, spouse + child	100	92	35	0	0
2 Grandparents (same side) + 2 children	100	89	0	0	0
3 Children	100	93	44	0	0
1 Parent + 2 children	100	88	60	9	5
Spouse + 2 children	100	100	99	55	35
2 Parents	100	100	100	1	0
1 Parent, spouse + 2 children	100	100	91	11	2
Required LR=	*100,000*	*6.25E + 07*	*6.25E + 09*	*1.00E + 13*	*6.25E + 13*
(b) Identifiler® + recommended expanded-ESS					
1 Sibling	60	20	8	1	1
3 Grandparents	68	15	2	0	0
1 Parent + half-sibling (opposite side)	75	15	1	0	0
1 Child	89	22	4	0	0
1 Parent	92	22	1	0	0
2 Grandparents (different sides) + 1 child	98	70	31	7	3
1 Parent + aunt/uncle (opposite side)	99	77	41	4	1
2 Siblings	99	82	52	12	8
Spouse + 1 child	100	95	60	8	7
1 Parent + 1 sibling	100	85	65	17	8
2 Grandparents (same side) + 1 child	100	100	93	0	0
4 Grandparents	100	100	97	4	0
3 Siblings	100	96	82	23	13
1 Parent + 1 child	100	99	88	27	17
1 Grandparent + 2 children	100	100	100	14	4
2 Children	100	100	100	66	44
1 Parent, spouse + child	100	100	100	75	52
2 Grandparents (same side) + 2 children	100	100	100	81	65
3 Children	100	100	98	82	71
1 Parent + 2 children	100	100	99	90	86
Spouse + 2 children	100	100	100	95	86
2 Parents	100	100	100	100	100
1 Parent, spouse + 2 children	100	100	100	100	100
Required LR=	*100,000*	*6.25E + 07*	*6.25E + 09*	*1.00E + 13*	*6.25E + 13*

The success rates are based on the percentage of DNA · VIEW simulations that, with different prior probabilities, met the required LR to obtain a posterior probability of 99.9%

The simulations shown here illustrate the value of using the recommended expanded ESS loci, in addition to STR markers that are already in routine use.

REFERENCES

1 Budimlija ZM, Prinz MK, Zelson-Mundorff A, *et al.* World trade center human identification project: experiences with individual body identification cases. *Croat Med J* 2003;**44**:259–63
2 Davoren J, Vanek D, Konjhodzic R, Crews J, Huffine E, Parsons TJ. Highly effective DNA extraction method for nuclear short tandem repeat testing of skeletal remains from mass graves. *Croat Med J* 2007;**48**:478–85
3 Huffine E, Crews J, Kennedy B, Bomberger K, Zinbo A. Mass identification of persons missing from the break-up of the former Yugoslavia: structure, function, and role of the International Commission on Missing Persons. *Croat Med J* 2001;**42**:271–5
4 Primorac D, Andelinovic S, Definis-Gojanovic M, *et al.* Identification of war victims from mass graves in Croatia, Bosnia, and Herzegovina by use of standard forensic methods and DNA typing. *J Forensic Sci* 1996;**41**:891–4
5 Krenke BE, Tereba A, Anderson SJ, *et al.* Validation of a 16-locus fluorescent multiplex system. *J Forensic Sci* 2002;**47**:773–85
6 Collins PJ, Hennessy LK, Leibelt CS, Roby RK, Reeder DJ, Foxall PA. Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: the AmpFlSTR® Identifiler® PCR amplification kit. *J Forensic Sci* 2004;**49**:1265–77
7 Brenner CH. Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities. *Forensic Sci Int* 2006;**157**:172–80
8 Brenner CH. Reuniting El Salvador Families. See http://dna-view.com/ProBusqueda.htm (last checked 14 April 2011)
9 Brenner CH, Weir BS. Issues and strategies in the DNA identification of World Trade Center victims. *Theor Popul Biol* 2003;**63**:173–8
10 Buckleton J, Triggs CM, Walsh SJ. *Forensic DNA Evidence Interpretation*. Boca Raton: CRC Press, 2005
11 Prinz M, Carracedo A, Mayr WR, *et al.* DNA Commission of the International Society for Forensic Genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI). *Forensic Sci Int Genet* 2007;**1**:3–12
12 Budowle B, Bieber FR, Eisenberg AJ. Forensic aspects of mass disasters: strategic considerations for DNA-based human identification. *Legal Med* 2005;**7**:230–43
13 National Institute of Justice. *Lessons Learned from 9/11: DNA Identification in Mass Fatality Incidents*. Washington, DC: US Department of Justice, 2005. See http://www.massfatality.dna.gov (last checked 14 April 2011)
14 Hill CR, Butler JM, Vallone PM. A 26plex autosomal STR assay to aid human identity testing*(dagger). *J Forensic Sci* 2009;**54**:1008–15
15 Schneider PM. Expansion of the European standard set of DNA database loci-the current situation. *Profiles DNA* 2009;**12**:6–7
16 Brenner CH. Symbolic kinship program. *Genetics* 1997;**145**:535–42
17 Brenner CH. *DNA • VIEW User's Manual*. Oakland: CH Brenner, 2009
18 Butler JM, Schoske R, Vallone PM, Redman JW, Kline MC. Allele frequencies for 15 autosomal STIR loci on US caucasian, African American, and Hispanic populations. *J Forensic Sci* 2003;**48**:908–11
19 Promega. Genetic identity reference information: population statistics. See http://www.promega.com/applications/hmnid/referenceinformation/popstat/custstat_Allelefreq.htm (last checked 24 June 2010)
20 Hill CR, Duewer DL, Kline MC, *et al.* Concordance and population studies along with stutter and peak height ratio analysis for the PowerPlex® ESX 17 and ESI 17 Systems. *Forensic Sci Int Genet* [doi: DOI: 10.1016/j.fsigen.2010.03.014]. In press,corrected proof **Q4**
21 CRAN. The comprehensive R archive network. See http://cran.r-project.org (last checked 14 April 2011)

QUERY FORM

Royal Society of Medicine

Journal Title: **MSL**
Article No: **11-068**

AUTHOR: The following queries have arisen during the editing of your manuscript. Please answer the queries by making the requisite corrections at the appropriate positions in the text.

Query No.	Nature of Query	Author's Response
Q1	Please provide up to two qualifications for both the authors.	
Q2	Please provide location of supplier Promega and Applied Biosystems.	
Q3	Please check the expansion of ISFG.	
Q4	Please provide the publication year, volume number and page range for ref. [20].	