Int J Legal Med (2000) 114 : 61–65 

ORIGINAL ARTICLE

**Francesc Calafell**

# The probability distribution of the number of loci indicating exclusion in a core set of STR markers

Received: 15 September 1999 / Accepted: 31 January 2000

**Abstract** The distribution of the number of loci out of the 13 in the CODIS STR set that would show an exclusion (i.e., a genotype set incompatible either with the prosecution hypothesis or with Mendelian transmission) was estimated in different scenarios. The knowledge of this distribution would provide a framework against which casework evidence can be compared. I used allele frequencies in Iberian and in Italian populations to generate individual genotypes at random and to test in 1 million simulation replicates, how many of the 13 loci would give an exclusion in an individual identification case, a paternity case, and a double parenthood case. All three scenarios were tested under an expected overall exclusion, both for unrelated individuals and for cases in which the suspect or the alleged father was the brother of the real culprit or real father. Paternity and double parenthood cases were also tested in the true scenario, with exclusionary loci due to mutation. In individual identification cases, the average number of exclusionary loci was 11.95 with a minimum of 7. This STR set also showed sufficient power to resolve identification cases in which the evidence sample came from a suspect's sib. False paternity cases yielded an average of 7.65 exclusionary loci and exclusions with only one (0.0108%) or two (0.14%) exclusionary loci were obtained only rarely. The cases of exclusion with one locus could lead to likelihood ratios in favour of paternity, while both true and false paternity cases with two exclusionary loci would often lead to non-conclusive likelihood ratios. The average number of exclusionary loci in a paternity case where the alleged father was the real father's brother was 3.82, with a significant number of cases where no exclusions were obtained.

F. Calafell (✉)
Unitat de Biologia Evolutiva,
Facultat de Cincies de la Salut i de la Vida,
Universitat Pompeu Fabra, Doctor Aiguader 80,
08003 Barcelona, Catalonia, Spain
e-mail: francesc.calafell@cexs.upf.es;
Tel.: +34-93-5422841; Fax: +34-93-5422802

**Key words** STR · Identification · Paternity · Parenthood · Exclusion

## Introduction

The forensic genetics community seems to be on the way to adopting a common set of short tandem repeat (STR) loci as core markers in routine practice. The availability of commercial kits for typing sets of STRs as well as the adoption of standard sets for data banking (such as CODIS in the USA) have fuelled this trend towards the de facto adoption of a common set of STRs. Although most of the technical aspects in the production of genotypes for these STRs are well known and standardised, the statistical properties of the genotype distributions for the same loci are less well known. I have explored one such aspect: the distribution of the number of exclusionary loci in the CODIS STR set. The term exclusionary locus is used here as shorthand for those loci where the genotype of the suspect in a particular case seems to be incompatible with the hypothesis of the prosecution i.e., an exclusionary locus shows different genotypes in the suspect and in the evidence. Or, in a paternity case with a known and typed mother, an exclusionary locus is where no combination of the genotypes of the alleged father and mother would produce the genotype of the child. It should be noted that in a paternity case, a genotype set can be exclusionary either because the paternity hypothesis is not correct, or because the Mendelian model of transmission is violated e.g., because of mutation. Therefore, both possibilities have to be taken into account when estimating the number of exclusionary loci.

If the suspect is not the real culprit, or if the alleged father is not actually the father, it is to be expected that a number of loci in a STR set would show an exclusion. The actual number of exclusionary loci would follow a probability distribution that depends on the number of markers used, on the genotype frequencies in the population, on the nature of the case (i.e., individual identification or paternity) and on the degree of relatedness (if any) between the actual donor of the sample (or father) and the suspect

or alleged father. The knowledge of this distribution would provide a framework against which the case evidence can be compared. Certain circumstances may result in a number of exclusionary loci that is different from the expected. For instance, relatedness between the actual provider of the evidence and the suspect may reduce the number of exclusionary loci. Or, in a case of true paternity, one or a few loci may have mutated and appear as exclusionary loci. The relative expected probability distributions of exclusionary loci under different hypotheses can be used as a first step in assessing the different hypotheses.

The distribution of the number of exclusionary loci has been estimated by means of a Monte Carlo simulation in the CODIS STR set from allele frequencies in Iberian and in Italian populations with different scenarios (individual identification, paternity, joint parenthood), in the case of an innocent, unrelated suspect (or non-paternity), or in the case of a suspect or alleged father who is a sib of the culprit or father.

## Materials and methods

The distribution of the number of exclusionary loci has been estimated in the set of 13 STRs in CODIS, which are also those in the AmpF*l*STR Profiler Plus and AmpF*l*STR Cofiler commercial kits (Perkin Elmer, Foster City, Calif.), as well as in GenePrint PowerPlex 1 and 2 (Promega, Madison Wis.). These loci are CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, TH01, TPOX and vWA. I have used a Monte Carlo simulation approach because, although the probability distribution of the number of exclusionary loci can be computed algebraically, the resulting equations would become too cumbersome to be resolved in practice. The following scenarios have been simulated:

a) Individual identification (e.g., as in crime scene evidence vs suspect's sample)
   1. Samples come from two different,
   2. Samples come from two sibs, children of unrelated parents

b) Paternity, mother known and typed
   1. Alleged father is the actual father
   2. Alleged father unrelated to real father
   3. Alleged father is the real father's brother (again, they are children of unrelated parents)

c) Exclusion of a couple as a child's parents (as in a baby exchange case)
   1. The couple are the actual parents
   2. The couple are two unrelated individuals and are not the child's parents
   3. One of the members of the couple is a sib of one of the actual parents.

A Monte Carlo computer simulation method was used to generate the probability distribution of the number of exclusionary loci under the scenarios listed under points a–c. The starting point was the allele frequencies for the Iberian populations which were determined from the data of Pérez-Lezaun et al. (2000) and contained allele frequencies in four different Iberian populations (Basques, Catalans, northern Portuguese and Andalusians). A set of Iberian allele frequencies was created by pooling the maximum number of populations that would be significantly homogeneous according to an exact test (Raymond and Rousset 1995) as implemented in the Arlequin software package (Schneider et al. 1997). Thus, the Basque allele frequencies were excluded for D21S11, D3S1358 and D13S317; the Andalusians were excluded for D7S820, and the northern Portuguese for D3S1358, D7S820 and D18S51 because for all these loci, *p*-values for the hypothesis of allele frequency homogeneity were lower than 0.006. The simulation was also run independently with the allele frequencies given for 223 samples from central and southern Italy (Garofano et al. 1998).

The 13-locus genotypes for the individuals needed for scenarios a–c were generated at random from the set of Iberian (or Italian) allele frequencies, under Hardy-Weinberg equilibrium and independence of loci. At each parent-child transmission, mutations were introduced with a specific probability for each locus (Table 1). Each mutation rate was either derived from actual paternity data (Ambach et al. 1997; Brinkmann et al. 1998; Mornhinweg et al. 1998) if those data were available and at least one mutation event had been observed, or it was estimated from the geometric mean of the repeat length, by using the exponential function in Brinkmann et al. (1998). Under each scenario, the loci with exclusionary genotypes were noted and a total number of exclusionary loci was obtained for the whole 13-locus set. This procedure was iterated 1 million times and the probability distribution of the number of exclusionary loci was estimated as the proportion of iterations in which 0,1,2,…,13 exclusionary loci were obtained. Moreover, for each iteration in scenarios b 1 and b 2, the likelihood ratio for paternity was computed from the equations in Table 6.2 in Evett and Weir (1998), modified by considering the locus-specific mutation rates.

**Table 1** Mutation rates (%) used in the simulation procedure for each locus (*1* estimated from the geometric mean of allele length by using the exponential function in Brinkmann et al. 1998)

| Locus | Mutation rate estimate (%) | Source |
|---|---|---|
| CSF1PO | 0.168 | (1) |
| D3S1358 | 0.769 | Mornhinweg et al. 1998 |
| D5S818 | 0.192 | (1) |
| D7S820 | 0.110 | (1) |
| D8S1179 | 0.311 | (1) |
| D13S317 | 0.164 | (1) |
| D16S539 | 0.166 | (1) |
| D18S51 | 0.565 | (1) |
| D21S11 | 0.180 | Brinkmann et al. 1998 |
| FGA | 0.401 | Brinkmann et al. 1998 |
| TH01 | 0.062 | (1) |
| TPOX | 0.086 | (1) |
| vWA | 0.251 | Pooled from Ambach et al. 1997 and Brinkmann et al. 1998 |

## Results and discussion

The distribution of the number of exclusionary loci under several scenarios is given in Table 2 for the Iberian data set and in Table 3 for the Italian data set. Since all distributions were practically identical for Iberians and Italians, I will further discuss the Iberian results only. The distribution of the number of expected loci in a case of individual identification (scenario a) is given in Fig. 1 both for unrelated individuals and for a sib pair. For unrelated individuals, the number of exclusionary loci obtained was 7–13, with an expectation of 11.95, 9 or more exclusionary loci were obtained with a 99.8% probability, and conversely, the probability of obtaining 6 or fewer exclusionary loci was less than 1 in a million. This is additional demonstration of the ample power of discrimination of the CODIS STR set (Garofano et al. 1998; Pérez-Lezaun et al.

**Table 2** Probability (%) of obtaining *k* exclusionary loci among the 13 CODIS STR loci in different scenarios, in the Iberian population (*Ind. ID* individual identification case with two unrelated individuals, *Ind. ID (sib)* individual identification case with two full sibs, *Paternity* non-paternity case, alleged father unrelated to real father, *Paternity (brother)* non-paternity, alleged father is real father's brother, *Couple parenthood* test whether a couple are the parents of a child, *Couple parenthood (sib)* same as previous, one member of the couple is a sib of one the real parents)

| *k* | Ind. ID (%) | Ind ID (sib %) | Paternity (real %) | Paternity (unrelated %) | Paternity (brother %) | Couple parenthood (real %) | Couple parenthood (unrelated %) | Couple parenthood (sib %) |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 95.0 | 0 | 0.982 | 93.6 | 0 | 0 |
| 1 | 0 | 0.0061 | 4.85 | 0.0108 | 5.40 | 6.20 | 0 | 0 |
| 2 | 0 | 0.0650 | 0.107 | 0.142 | 13.9 | 0.183 | 0 | 0.0078 |
| 3 | 0 | 0.375 | 0.003 | 0.744 | 22.0 | 0.0079 | 0.0061 | 0.0934 |
| 4 | 0 | 1.56 | 0 | 2.69 | 23.6 | 0 | 0.0353 | 0.445 |
| 5 | 0 | 4.62 | 0 | 7.48 | 18.2 | 0 | 0.261 | 1.92 |
| 6 | 0 | 10.8 | 0 | 14.1 | 10.1 | 0 | 1.18 | 5.62 |
| 7 | 0.0156 | 17.7 | 0 | 20.8 | 4.29 | 0 | 4.11 | 12.5 |
| 8 | 0.211 | 22.3 | 0 | 22.0 | 1.32 | 0 | 10.6 | 20.3 |
| 9 | 1.38 | 20.7 | 0 | 17.5 | 0.298 | 0 | 19.9 | 23.9 |
| 10 | 6.49 | 13.5 | 0 | 9.79 | 0.0436 | 0 | 25.9 | 19.7 |
| 11 | 20.3 | 6.30 | 0 | 3.72 | 0.0031 | 0 | 23.0 | 11.3 |
| 12 | 38.7 | 1.85 | 0 | 0.85 | 0 | 0 | 12.1 | 3.70 |
| 13 | 32.9 | 0.225 | 0 | 0.0983 | 0 | 0 | 2.85 | 0.585 |

**Table 3** Probability (%) of obtaining *k* exclusionary loci among the 13 CODIS STR loci in different scenarios in the Italian population (column headings as in Table 2)

| *k* | Ind. ID (%) | Ind ID (sib %) | Paternity (real %) | Paternity (unrelated %) | Paternity (brother %) | Couple parenthood (real %) | Couple parenthood (unrelated %) | Couple parenthood (sib %) |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 94.9 | 0.0015 | 0.942 | 93.8 | 0 | 0 |
| 1 | 0 | 0.0031 | 4.98 | 0.0080 | 5.49 | 6.08 | 0 | 0 |
| 2 | 0 | 0.0681 | 0.128 | 0.156 | 13.9 | 0.163 | 0 | 0.0092 |
| 3 | 0 | 0.354 | 0.0015 | 0.696 | 21.9 | 0.0015 | 0.0031 | 0.0653 |
| 4 | 0 | 1.58 | 0 | 2.81 | 23.6 | 0 | 0.0297 | 0.490 |
| 5 | 0.0016 | 4.73 | 0 | 7.35 | 18.1 | 0 | 0.274 | 2.04 |
| 6 | 0.0016 | 10.5 | 0 | 14.4 | 10.1 | 0 | 1.32 | 5.76 |
| 7 | 0.0330 | 18.0 | 0 | 20.9 | 4.23 | 0 | 4.29 | 12.9 |
| 8 | 0.217 | 22.2 | 0 | 22.1 | 1.35 | 0 | 10.7 | 20.2 |
| 9 | 1.37 | 20.8 | 0 | 17.4 | 0.328 | 0 | 19.7 | 23.8 |
| 10 | 6.72 | 13.7 | 0 | 9.61 | 0.0554 | 0 | 26.3 | 19.6 |
| 11 | 20.82 | 6.13 | 0 | 3.68 | 0.0049 | 0 | 22.7 | 11.0 |
| 12 | 38.44 | 1.72 | 0 | 0.800 | 0.0015 | 0 | 12.0 | 3.61 |
| 13 | 32.41 | 0.222 | 0 | 0.0839 | 0 | 0 | 2.78 | 0.547 |

2000). It should be noted that if the evidence sample does originate from the subject, exclusionary genotypes can result from a number of circumstances (e.g., sample contamination, sample mix-up, laboratory errors) that are difficult to quantify and to incorporate into a model. In the case of a sib pair, the average number of exclusionary loci was 8.13, and the probability of obtaining five or more exclusionary loci was 97.9%. In all of the 1 million iterations, at least 1 exclusionary locus was obtained. Thus, the chance that a brother of the suspect would share an identical 13-locus profile is roughly 1 in a million or less, which is a strong argument in those cases where the defence alleges that it was a brother of the defendant who committed the crime (Evett 1992). If the hypotheses that the evidence sample belongs to an unrelated individual or to a sib of the suspect are given equal a priori probabilities, the likelihood of obtaining only 7, 8 or 9 exclusionary loci is 1135, 105.4 or 15.02 times greater for sib pairs than for unrelated individuals, respectively. If a real case yields such a number of (or fewer) exclusionary loci, the considerations and the decision rule suggested by Sjerps and Kloosterman (1999) may be applied to decide whether the biological evidence was left by a suspect's sib.

In a simple paternity case (Fig. 2), an average of 7.65 exclusionary loci are expected, with three or more exclusionary loci in 99.85% of the cases. The probability of excluding with only one locus was 0.0108% (or 1 in 9259), and with two it was 0.14% (1 in 704). If the man tested was indeed the child's father, 95.04% of the cases did not show any exclusionary loci, 4.85% of the cases yielded 1 exclusionary locus, 0.11% (1 in 936) yielded 2 exclusionary loci, and 30 in 1 million yielded 3 exclusionary

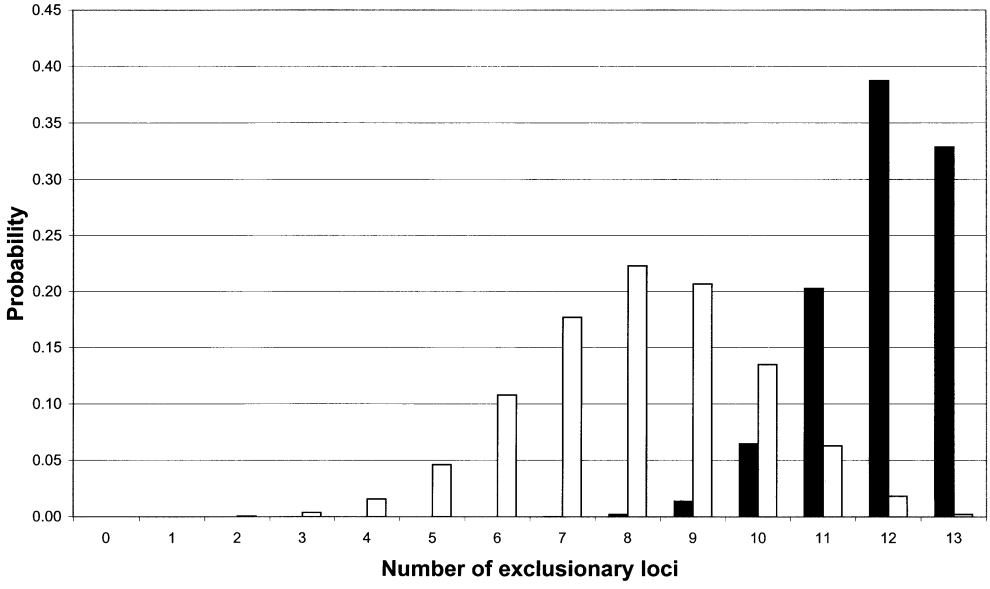


**Fig. 1** Number of exclusionary loci in an individual identification case using Iberian data (*black bars* two unrelated individuals, *open bars* two sibs)

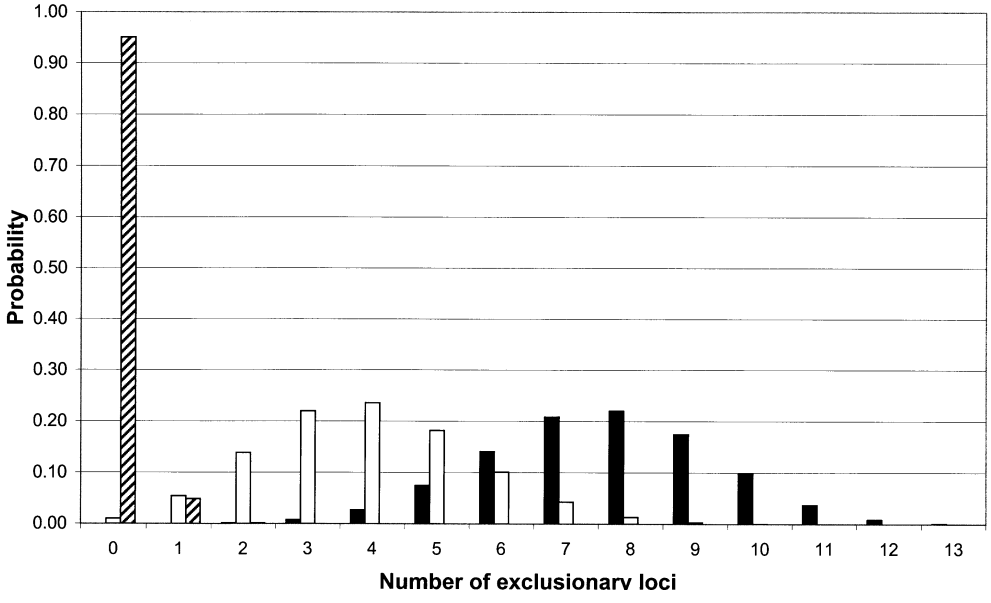


**Fig. 2** Number of exclusionary loci in a paternity case using Iberian data (*black bars* alleged father is unrelated to real father, *open bars* alleged father is real father's brother, *hatched bars* real paternity case)

loci. Thus, the distribution of the number of exclusionary loci in the cases of false and real paternity overlap slightly. The usual way of reaching a decision on paternity is using a likelihood ratio. In the 950,441 real paternity cases that did not show any exclusionary locus, the overall likelihood ratio (LR) ranged from 111.79 to over 18,000,000 with an average of 57,907 and at a 95% confidence interval 1,745–332,584. In the 48,461 cases with 1 exclusionary locus, the average overall LR (which included the exclusionary locus as well as the 12 loci that did not show an exclusion) was 611, and a LR > 20 was obtained in 94.2% of the cases. However, if the alleged father was unrelated to the real father, in the 108 cases in which only one exclusionary locus was obtained, the average LR was 55, and a LR > 20 was obtained in 71.3% of the cases. Therefore, although false paternities with only one exclusionary locus are rare, in most cases they would be mistaken for real paternities. Further refinement of the model in the likelihood ratio that would take into account additional circumstances such as the difference in the repeat size between the non-maternal allele of the child and the alleged father's alleles, allele length (Brinkmann et al. 1998) or father's age (Henke and Henke 1999) may help in correctly assessing those cases. If 2 exclusionary loci are present, in real paternities ($n$ = 1,068) the average LR was 8.07, it was larger than 20 only in 13% of the cases, and it was even below 1 in 41.8% of the cases. In 1,420 false paternity cases with two exclusionary loci, the average LR was 0.7, it was never above 20, but in 19.6% of the cases it was below 1 in 20. In this situation it is most likely that a non-conclusive LR will be obtained, and that typing further loci would be necessary to reach a decision.

If the alleged father is the actual father's brother, the number of exclusionary loci drops to an average of 3.89,

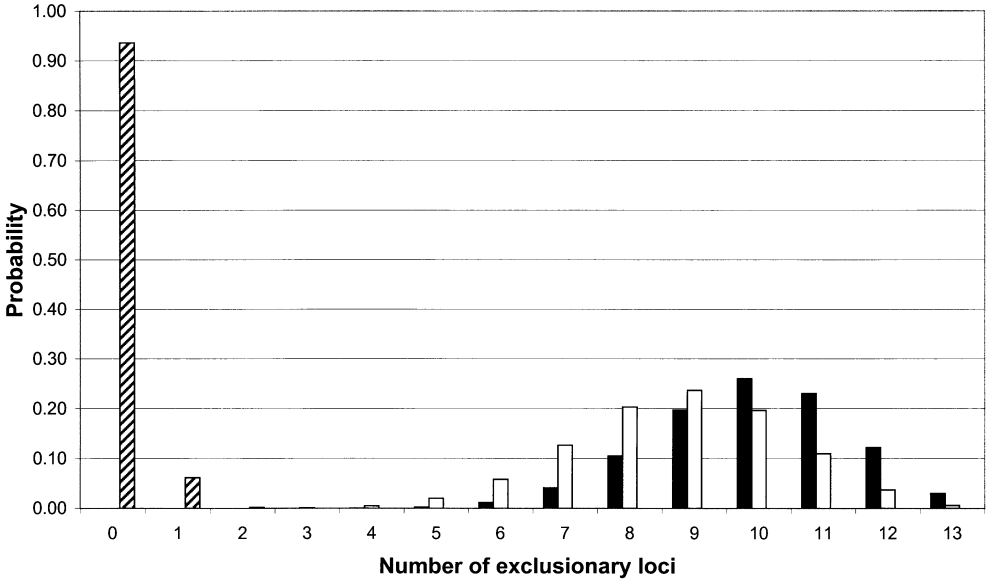


**Fig. 3** Number of exclusionary loci in a case in which a couple is tested for parenthood of a child (*black bars* couple unrelated to actual parents, *open bars* one of the members of the couple is a sib of one of the real parents, *hatched bars* real parenthood case)

with 0.98% of cases in which no exclusion could be obtained. If the hypothesis that the alleged father is a brother of the actual father and that he is actually the father are given equal a priori probabilities, obtaining only 0, 1, 2 or 3 exclusionary loci is 0.0103, 1.11, 130 or 7,317 times more likely, respectively if the alleged father is the real father's brother than if he is actually the father. If we now compare paternity by a brother with paternity by a man unrelated to the alleged father, it is > 9,815, 500, 97.6, and 29.5 times more likely to obtain 0, 1, 2 or 3 exclusionary loci , respectively, if the alleged father is a brother of the real father than if the alleged father is unrelated to the real father. If there is a suspicion that a brother of the alleged father might be the actual father, a low number of exclusionary loci may be an indication that this is the case. The genotype evidence should then be weighed with the appropriate likelihood ratio for the two hypotheses being contemplated.

A different scenario consists of testing whether a couple could be the parents of a child, as in a baby exchange case where exclusion of either of the two alleged parents excludes both individuals simultaneously. The average number of exclusionary loci that the CODIS set would provide in this scenario is 9.96, with a minimum of 3 (in 61 out of 1 million cases), and 6 or more in 99.70% of the cases. Again, involvement of a sib decreases the number of expected exclusionary loci and the average is then 8.83 for the cases in which a member of the couple is a sib of one of the actual parents. In this case, the decrease is not so large because the unrelated member of the couple can still be excluded with undiminished power.

**Acknowledgements** This research was stimulated by the baby exchange problem posed as a quality control exercise by the Spanish and Portuguese Group of the ISFH, in which related individuals and mutated genotypes had been included. Anna Pérez-Lezaun made useful comments on a previous version of this manuscript and two anonymous reviewers also contributed to improve this manuscript. During this work, the author was supported by a return contract awarded within DGICYT (Spain) project PB95-0267-C02-01 to Jaume Bertranpetit. Additional support was provided by Direcci General de Recerca, Generalitat de Catalunya (Catalonia; project 1998SGR00009).

## References

Ambach E, Parson W, Niederstatter H, Budowle B (1997) Austrian Caucasian population data for the quadruplex plus amelogenin: refined mutation rate for HumvWFA31/A. J Forensic Sci 42: 1136–1139

Brinkmann B, Klintschar M, Neuhuber F, Hhne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62: 1408–1415

Evett IW (1992) Evaluating DNA profiles in a case where the defence is "It was my brother". J Forensic Sci Soc 32: 5–14

Evett IW, Weir BS (1998) Interpreting DNA evidence: statistical genetics for forensic scientists. Sinauer Associates, Sunderland Mass.

Garofano L, Pizzamiglio M, Vecchio C, Lago G, Floris T, D'Errico G, Brembilla G, Romano A, Budowle B (1998) Italian population data on thirteen short tandem repeat loci: HUMTH01, D21S11, D18S51, HUMVWFA31, HUMFIBRA, D8S1179, HUMTPOX, HUMCSF1PO, D16S539, D7S820, D13S317, D5S818, D3S1358. Forensic Sci Int 97: 53–60

Henke J, Henke L (1999) Mutation rate in human microsatellites. Am J Hum Genet 64: 1473

Pérez-Lezaun A, Calafell F, Clarimón J, Bosch E, Mateu E, Gusmão L, Amorim A, Benchemsi N, Bertranpetit J (2000) Allele frequencies of 13 short tandem repeat polymorphisms in population samples of the Iberian Peninsula and Northern Africa. Int J Legal Med 113: 208–214

Mornhinweg E, Luckenbach C, Fimmers R, Ritter H (1998) D3S1358: sequence analysis and gene frequency in a German population. Forensic Sci Int 95: 173–178

Raymond M, Rousset F (1995) An exact test for population differentiation. Evolution 49: 1280–1283

Schneider S, Kueffer J-M, Excoffier L (1997) Arlequin ver 1.1: a software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland

Sjerps M, Kloosterman AD (1999) On the consequences of DNA profile mismatches for close relatives of an excluded suspect. Int J Legal Med 112: 176–180