Int J Leg Med (2006) 120: 95–104
DOI 10.1007/s00414-005-0017-2

ORIGINAL ARTICLE

Amy D. Anderson · Bruce S. Weir

It was one of my brothers

Received: 25 January 2005 / Accepted: 10 June 2005 / Published online: 25 August 2005

Abstract When DNA evidence is used to implicate a suspect, it may be of interest to know whether it is likely that the suspect's near relatives also share the suspect's DNA profile. In this study we discuss methods for evaluating the probability that at least one of a set of the suspect's full or half-siblings shares the suspect's DNA profile. We present three such methods: exact calculation, estimation via Monte Carlo simulations, and estimation by means of sandwiching the probability between an upper and a lower bound. We show that, under many circumstances, this upper bound itself provides an extremely quick and accurate estimate of the probability that at least one of the relatives matches the suspect's profile.

Keywords Forensic science · Match probability · DNA profile · Relatives

Introduction

In this study, we consider the situation in which a suspect's DNA profile is found to match a profile found at a crime scene, and we wish to examine the hypothesis that one of a set of the suspect's near relatives might also have a profile that matches. This is of interest because DNA profiles of near relatives may be similar and, if a suspect has many such relatives, there is a chance that at least one of them shares the suspect's profile. In this situation, a possible defense would be for the suspect to claim that the DNA found at the scene belonged to an unspecified one of these relatives (i.e., "It was my one of my brothers"). If there is no other reason to suspect these relatives of committing the crime, they cannot be compelled to give their DNA for genotyping and it is assumed that they will not volunteer for such testing. The question arises: given that the suspect's DNA matches the sample found at the crime scene, what is the probability that at least one of a specified set of relatives also has a DNA profile that matches?

Forensic calculations involving related individuals have been considered in a number of contexts. Brookfield [5] and Donnelly [6] discussed the situation in which a suspect's DNA matched that found at a crime scene and examined the effect on the likelihood ratio of ignoring the possibility that the culprit might be a relative of the suspect. Within this context, Balding and Donnelly [2] suggested that relatives of the suspect should be explicitly included in the calculation of match probabilities. Balding and Nichols [3], Belin et al. [4], Evett [7], Weir [14] and Weir and Hill [16] all considered the calculation of match probabilities and/or likelihood ratios in the case in which a single relative of the suspect may have been the culprit. Similar calculations occur in paternity analysis. Lee et al. [12] and Fung et al. [11] derived formulae for use in paternity testing when the alleged father is a relative of the true father. More generally, Ayres [1] developed a methodology for testing whether two individuals had one of a number of genetic relationships (e.g., full siblings, half-siblings, etc.). The presence of relatives has also been considered in the evaluation of DNA mixtures [8, 10].

The situation of interest in this paper, in which we consider a set relatives of a suspect, has been examined by Evett [7], who considered the case in which the suspect had a number of full siblings. Evett's approach for multiple full siblings was first to calculate Q, the probability that a single full sibling matched the suspect's profile, then use nQ to approximate the probability that at least one of the suspect's n full siblings matched the profile. This approximation works well if the probability of having multiple siblings with the suspect's profile is negligible (as would be the case,

A. D. Anderson (✉)
Bioinformatics Research Center,
North Carolina State University,
Campus Box 7566 Raleigh, NC 27695-7566, USA
e-mail: amya@statgen.ncsu.edu
Tel.: +1-919-5133439
Fax: +1-919-5157315

B. S. Weir
Program in Statistical Genetics,
North Carolina State University,
Raleigh, NC 27695-7566, USA

for example, if sufficiently many loci were typed), but leaves two questions unanswered: how many loci are necessary in order for this estimate to be valid and how can the probability that at least one of the relatives matches the suspect's profile be calculated in cases in which this estimate is poor?

In this paper, we present a number of methods for evaluating the probability that at least one of a group of the suspect's relatives shares the suspect's DNA profile, including a method for exact calculation and techniques for estimating the probability when exact methods prove difficult. This work expands upon Evett's work by addressing the two questions listed above and by including both full siblings and half-siblings in our analyses (we also briefly include a simple situation with first cousins). In addition, we generalize the previous work by including a parameter, θ, which allows us to incorporate some amount of population structure into our calculations as described by Weir [15].

Materials and methods

In this treatment, we will assume that genotypes are measured without error and population allele frequencies are known. Those issues aside, we turn our attention to calculating the match probability, that is, the probability that at least one member of the suspect's set of relatives matches the suspect's multilocus genetic profile.

Throughout this work, we assume a model that allows populations to be composed of subpopulations. Such population structure leads to correlations between allele types of seemingly unrelated people within a family. To take this into account, we make use of a population parameter θ (also known as F_{ST}). In practice, θ represents the probability that two alleles in a subpopulation are of identical type because both are descended from a common (unknown) ancestral source. To facilitate ease of computation, we assume that allele frequencies in a subpopulation are distributed according to a Dirichlet distribution over subpopulations with expected values equal to the population allele frequencies. This model has been presented more fully by Weir [15] and has additional theoretical justification from population genetic theory [9].

Single locus calculations

In order to evaluate the match probability, we will need to calculate genotype frequencies for a single individual and joint genotype frequencies for groups of individuals. Note that, since we are ultimately interested only in whether or not an individual matches the genotype of the suspect, $\mathcal{P}$, we need consider at most three allelic types at each locus—the one or two types $\mathcal{P}$ has, plus another type that represents all alleles that $\mathcal{P}$ does not possess. With this in mind, let A_1, A_2, and A_3 be alleles at the locus of interest, with frequencies p_1, p_2, and p_3, respectively. With this notation, the single individual genotype frequencies are $\Pr(A_iA_i)=p_i[(1-\theta)p_i+\theta]$ and, if $i\neq j$, $\Pr(A_iA_j)=2(1-\theta)p_ip_j$ (see, for example, [15]).

Define $M_{i,j}=[(1-\theta)p_i+j\theta]$, for $i=1, 2, 3$ and $j=0, 1, \ldots$. When allele frequencies are assumed to follow a Dirichlet distribution, joint genotype frequencies among n individuals can be easily calculated. Let g be the joint genotype among the n people whose joint genotype frequency we wish to evaluate. Let h be the number of individuals that are heterozygous under g, and let t_i be the number of A_i alleles among the n individuals. Then

$$\Pr(g)=\frac{2^h\prod_{i=1}^{3}\prod_{j=0}^{t_i-1}M_{i,j}}{\prod_{j=0}^{2n-1}[1+(j-1)\theta]}. \tag{1}$$

We begin our calculations of the match probability with the case in which the suspect, $\mathcal{P}$, has n_f full siblings and no half-siblings. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ denote $\mathcal{P}$'s mother and father, respectively. Let P be the genotype of $\mathcal{P}$, P_1 and P_2 be the genotypes of $\mathcal{P}_1$ and $\mathcal{P}_2$, and $S_1,\ldots,S_{n_f}$ be the genotypes of $\mathcal{P}$'s full-siblings. In this problem, P is considered known and the other genotypes are all unknown. Noting that the genotypes of the offspring are independent once the genotypes of the parents are known, we can calculate the probability that at least one of the siblings matches $\mathcal{P}$'s profile by conditioning on P_1 and P_2 as follows:

$$\begin{aligned}\Pr(\text{at least one match}|P) &= 1-\Pr(\text{no match}\,|P) \qquad (2)\\ &= 1-\sum_{P_1,P_2}\Pr(S_1\neq P,\ldots,S_{n_f}\neq P|P_1,P_2,P)\\ &\quad\times\Pr(P_1,P_2|P)\\ &= 1-\sum_{P_1,P_2}\Pr(S_1\neq P|P_1,P_2,P)^{n_f}\Pr(P_1,P_2|P),\end{aligned}$$

where

$$\Pr(P_1,P_2|P)=\frac{\Pr(P|P_1,P_2)\Pr(P_1,P_2)}{\Pr(P)}. \tag{3}$$

When $\mathcal{P}$ is homozygous with genotype $P=A_1A_1$, we need consider only two possible alleles. In this case, the only possible parental genotypes are A_1A_1 and A_1A_2 and we may use Eqs. (1) and (3) to derive:

$$\begin{aligned}&\Pr(P_1=A_1A_1,P_2=A_1A_1|P=A_1A_1) \qquad (4)\\ &\quad=M_{1,3}M_{1,2}/[(1+2\theta)(1+\theta)]\end{aligned}$$

$$\begin{aligned}&\Pr(P_1=A_1A_1,P_2=A_1A_2|P=A_1A_1) \qquad (5)\\ &\quad=M_{1,2}M_{2,0}/[(1+2\theta)(1+\theta)]\end{aligned}$$

$$\Pr(P_1 = A_1A_2, P_2 = A_1A_2 | P = A_1A_1) \quad (6)$$
$$= M_{2,1}M_{2,0}/[(1+2\theta)(1+\theta)].$$

With these values, whenever $n_f \geq 1$, Eq. (2) yields:

$$\Pr(\text{at least one match} | P = A_1A_1)$$
$$= 1 - \left[0\frac{M_{1,3}M_{1,2}}{(1+2\theta)(1+\theta)} + \left(\frac{1}{2}\right)^{n_f}\frac{M_{1,2}M_{2,0}}{(1+2\theta)(1+\theta)} + \left(\frac{1}{2}\right)^{n_f}\frac{M_{1,2}M_{2,0}}{(1+2\theta)(1+\theta)} + \left(\frac{3}{4}\right)^{n_f}\frac{M_{2,1}M_{2,0}}{(1+2\theta)(1+\theta)}\right]$$
$$= 1 - \frac{M_{2,0}}{2^{n_f}(1+2\theta)(1+\theta)}\left[2M_{1,2} + \left(\frac{3}{2}\right)^{n_f}M_{2,1}\right],$$

provided $n_f \geq 1$. When $P = A_1A_2$, similar calculations (again, with $n_f \geq 1$) yield the following result:

$$\Pr(\text{at least one match} | P = A_1A_2)$$
$$= 1 - \frac{\left(M_{1,1} + \left(\frac{3}{2}\right)^{n_f}M_{3,0}\right)(1+2\theta) + M_{2,1}(M_{2,2} + M_{3,0})}{2^{n_f}(1+2\theta)(1+\theta)}.$$

When the family also includes half-siblings, calculations must proceed by conditioning upon the joint genotypes of all parents of $\mathcal{P}$ and the half-siblings. Let n_1 and n_2 be the number of mates of $\mathcal{P}_1$ and $\mathcal{P}_2$ (not including each other). For i=1, 2, let $\mathcal{P}_{i,j}$ denote $\mathcal{P}_i$'s jth mate and let the genotype of this individual be $P_{i,j}$, and let P^* denote the joint genotype of all the mates of $\mathcal{P}_1$ and $\mathcal{P}_2$. Moreover, let $n_{i,j}$ be the number of offspring of $\mathcal{P}_i$ and $\mathcal{P}_{i,j}$, and let the genotypes of these offspring be denoted $S_{i,j,1}, \ldots, S_{i,j,n_{i,j}}$. Let G be the set of all possible joint genotypes among the parents in the pedigree, where an observation, g, in G has the form $(P_1, P_2, P_{1,1}, P_{1,2}, \ldots, P_{2,n_2}) = (P_1, P_2, P^*)$.

Let H be the event that none of $\mathcal{P}$'s full siblings matches $\mathcal{P}$'s genotype, and let $H_{i,j}$ be the event that none of the offsprings of $\mathcal{P}_i$ and $\mathcal{P}_{i,j}$ matches $\mathcal{P}$'s genotype. With this notation, the probability of finding at least one match among $\mathcal{P}$'s full and half-siblings is as follows:

$$\Pr(\text{at least one match} | P)$$
$$= 1 - \sum_{g \in G}\Pr(\text{no match}|g)\Pr(g|P)$$
$$= 1 - \sum_{g \in G}\left[\left(\prod_{i=1}^{2}\prod_{j=1}^{n_i}\Pr(H_{i,j}|P_i, P_{i,j}, P)\right)\Pr(H|P_1, P_2, P)\Pr(g|P)\right] \quad (7)$$
$$= 1 - \sum_{g \in G}\left[\left(\prod_{i=1}^{2}\prod_{i=1}^{n_i}\Pr(S_{i,j,1} \neq P|P_i, P_{i,j}, P)^{n_{i,j}}\right) \times \Pr(S_1 \neq P|P_1, P_2, P)^{n_f}\Pr(g|P)\right]$$

where $(\Pr | gP)$ can be calculated according to Eq. (8):

$$\Pr(g|P) = \frac{\Pr(P|P_1, P_2)\Pr(P_1, P_2, P*)}{\Pr(P)} \quad (8)$$

In Eq. (8), $\Pr(P | P_1, P_2)$ is evaluated using the laws of Mendelian inheritance, whereas $\Pr(P_1, P_2, P^*)$ can be calculated using Eq. (1).

Multilocus calculations

Care must be taken in generalizing from single locus calculations to multilocus calculations when there are multiple relatives being considered. Whereas, for a single sibling, the multilocus match probability can be found by simply multiplying the single-locus match probabilities for each locus, the same is not true in the case of multiple relatives. At a single locus, the match probability is the probability that at least one of the relatives has a profile matching the

suspect. The product of these is the probability that some relative shares the suspect's genotype at each locus, not the probability that there is a single relative that shares at all loci.

The difference can be striking. For example, suppose a suspect with a single full-sibling has genotype *Aa* at each of ten loci, where both alleles *A* and *a* have a frequency of 0.20. In this case, the probability that the sibling matches the suspect's profile at one locus is 0.3817, so the probability of a match at all ten loci is $0.3817^{10}=6.56\times10^{-5}$. If the suspect had ten full siblings, the probability that at least one of the siblings matches at a single locus is 0.9678. The value $0.9678^{10}=0.72$ is the probability that a match can be found among the siblings at each locus, while the probability that at least one of the siblings matches at all ten loci is in fact 6.6×10^{-4} (or about 1 in 1,500).

The correct way to calculate multilocus calculations when there are multiple relatives of interest is similar to the approach used in calculating the single-locus match probabilities. Equations (7) and (8) still hold, except that now each genotype must be considered to be a multilocus genotype. More notation is necessary. If P is now the multilocus genotype of $\mathcal{P}$, let $P(k)$ be the genotype of $\mathcal{P}$ at the kth locus. Let this convention follow for all previously defined genotypes (for example, $g(k)$ is the joint genotype of all parents under consideration at the kth locus), and let m be the number of loci at which $\mathcal{P}$ was typed. Under the assumption that the loci are independent (i.e., unlinked and are not in linkage disequilibrium with each other), the individual components in Eq. (7) can be found as follows:

$$\Pr(S_1 \neq P|P_1,P_2,P) = 1-\prod_{k=1}^{m}\Pr\left(S_1(k)=P(k)|P_1(k),P_2(k),P(k)\right)$$

$$\Pr\left(S_{i,j,1}\neq P\middle|P_i,P_{i,j},P\right) = 1-\prod_{k=1}^{m}\Pr\left(S_{i,j,1}(k)=P(k)|P_i(k),P_{i,j}(k),P(k)\right)$$

$$\Pr\left(g|P\right)=\prod_{k=1}^{m}\Pr\left(g(k)|P(k)\right)$$

Note that the sum in Eq. (7) is now over the joint multilocus genotypes of all parents in the pedigree. The number of such genotypes may be quite large. If $\mathcal{P}$'s genotype at any locus is denoted A_1A_1 or A_1A_2, then, for each parent, at each locus there are up to five possible genotypes, namely A_1A_1, A_1A_2, A_2A_2, A_1A_3, and A_2A_3. Then, the number of possible joint multilocus genotypes among the parents is on the order of 5^{nm}, where n is the number of parents in the pedigree. Hence, unless there are relatively few parents to consider and only a small number of loci, the exact calculation may be computationally infeasible.

To evaluate the probability that at least one of $\mathcal{P}$'s siblings or half-siblings matches $\mathcal{P}$'s genetic profile in cases in which analytical calculations are infeasible, one may use Monte Carlo techniques. Using this approach, we simulate N sets of genotypes for the individuals in $\mathcal{P}$'s family and use the results to estimate the proportion of such configurations in which at least one of $\mathcal{P}$'s full or half-siblings shares $\mathcal{P}$'s genetic profile. For the ith simulated genotype configuration, we record $x_i=1$ if at least one of the siblings or half-siblings matches P's profile and $x_i=0$ otherwise. Our sampling strategy is as follows:

1. Simulate the genotypes of $\mathcal{P}_1$ and $\mathcal{P}_2$ conditioned on the genotype of $\mathcal{P}$, using Eq. (3).
2. Simulate the genotypes of $\mathcal{P}$'s siblings using the rules of Mendelian inheritance from P_1 and P_2. If any of these matches P, record $x_i=1$ and go on to the next simulation, starting with step 1. Otherwise, continue on to step 3.
3. For each of the other parents in the pedigree, simulate that parent's genotype conditional on all previous parental genotypes in the current simulation. Use the rules of Mendelian inheritance to simulate genotypes for each of the offspring of that parent. If, at any point, a half-sibling is generated that matches $\mathcal{P}$'s profile, record $x_i=1$ and move on to the next simulation. If the entire pedigree is simulated with no siblings or half-siblings whose genotype matches P, record $x_i=0$.

When all N simulated genotype configurations are completed, the estimate for the probability that at least one of $\mathcal{P}$'s siblings or half-siblings shares a genetic profile with $\mathcal{P}$ is $\hat{p}=\sum_{i=1}^{N}x_i/N$. The estimated standard error of this estimator is $\sqrt{\hat{p}(1-\hat{p})/N}$.

The Monte Carlo technique works well when the match probability is not too small (i.e., not too many loci, several relatives of interest). When the match probability is very low, a prohibitive number of samples must be generated in order to yield a standard error small enough to give us confidence in the accuracy of the estimated match probability. Fortunately, it is exactly in this situation that the following method, for estimating the match probability by sandwiching it between two bounds, works well.

Estimation by bounding

In this section, we demonstrate upper and lower bounds for the match probability, based on Bonferroni's inequality. When the match probability is quite small, these bounds are close together and so give a quite precise estimate of the match probability.

Let $\mathcal{S}_1,\ldots,\mathcal{S}_n$ be the relatives of interest, with profiles $S_1,\ldots S_n$. We are interested in the probability that at least

one of $S_1, \ldots S_n$ is equal to P, the profile of the suspect. It is always true that:

$$\Pr(\text{at least one match}) \leq \Pr(S_1 = P) + (S_2 = P) + \ldots + \Pr(S_n = P). \quad (9)$$

The right-hand side of Eq. (9) is often a good estimate for the match probability and is valid regardless of the nature of the relationships between $\mathcal{S}_1, \ldots, \mathcal{S}_n$ and $\mathcal{P}$. A list of formulae for single-locus match probabilities, Pr (S_i=P), for various relationships can be found, for example, in Weir [15]. Since in this case we are interested in the probability that a single individual matches P, the multilocus match probabilities can be found by taking the product of the single-locus match probabilities.

In the case in which $\mathcal{P}$ has n_f full siblings and n_h half-siblings, Eq. (9) becomes:

$$\Pr(\text{at least one match}) \leq n_\mathrm{f} \Pr(\text{a full} - \text{sib matches } P) + n_\mathrm{h} \Pr(\text{a half} - \text{sib matches } P),$$

which, if n_h=0 and the inequality sign is replaced by an approximation sign, reduces to Evett's formula for multiple full siblings [7].

A lower bound for the match probability can also be derived. For the lower bound, in addition to calculating the probability that each individual relative of interest matches $\mathcal{P}$'s profile (as in the calculation of the upper bound), one must also compute, for each possible pair of $\mathcal{P}$'s relatives of interest, the probability that both individuals in the pair match $\mathcal{P}$'s profile. Then,

$$\Pr(\text{at least one match}) \geq \sum_{i=1}^{n} \Pr(S_i = P) - \sum_{j=1}^{n} \sum_{k=j+1}^{n} \Pr(S_i = S_j = P).$$

In contrast to the upper bound, which relied only on single-individual match probabilities, the lower bound contains probabilities that both individuals in a pair will match $\mathcal{P}$'s profile. These probabilities will depend upon the relationships between the individuals in the pair. For example, when only full siblings and half-siblings of $\mathcal{P}$ are considered as people of interest, there are five possible types of relative pairs: (1) both individuals in the pair are full siblings to $\mathcal{P}$, (2) one individual is a full sibling and the other a half-sibling to $\mathcal{P}$, (3) both individuals are half-siblings to $\mathcal{P}$ but are full siblings to each other, (4) both individuals are half-siblings to $\mathcal{P}$ and are half-siblings to each other, and (5) both individuals are half-siblings to $\mathcal{P}$ but are unrelated to each other.

Formulae for the probabilities that both individuals match $\mathcal{P}$'s single-locus genotype in each of these situations are presented in Appendix A. Multilocus probabilities that both individuals in a pair match $\mathcal{P}$'s profile can be found by multiplying the single-locus probabilities.

One might also note that, in the case in which there is just one relative of interest, the lower and upper bounds are identical and equal to the exact value of the match probability. In the case with two relatives of interest, the lower bound gives the exact probability. With more than two relatives, the exact value will lie somewhere between the two bounds.

Results and discussion

Single-locus match probabilities

We begin by showing results for a single locus. To investigate how the probability of at least one match depends upon the family's configuration, we compared probabilities among the four pedigrees shown in Fig. 1. In pedigree A, the relatives of interest are full siblings to $\mathcal{P}$, while in the other three pedigrees they are half-siblings to $\mathcal{P}$, but vary in their relationships to each other. For pedigree B, we supposed that the half-siblings were evenly divided between $\mathcal{P}$'s two parents.

For each given pedigree, we calculated the probability (using θ=0.03) that at least one full sibling or half-sibling matches $\mathcal{P}$'s genotype and have plotted this as a function of allele frequencies in Fig. 2. For the case in which $\mathcal{P}$ is

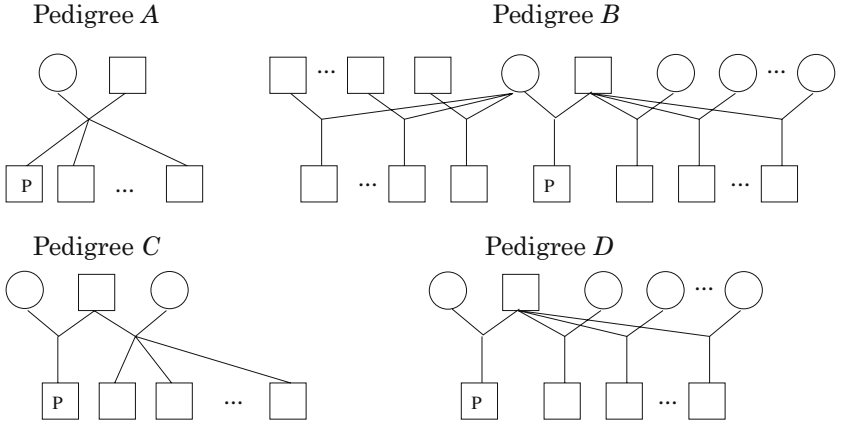

Fig. 1 Sample pedigrees. Pedigree A represents the situation in which the suspect, $\mathcal{P}$, has a number of full-siblings. Pedigrees B, C, and D show a few of the different family configurations possible when the relatives of interest are the suspect's half-siblings

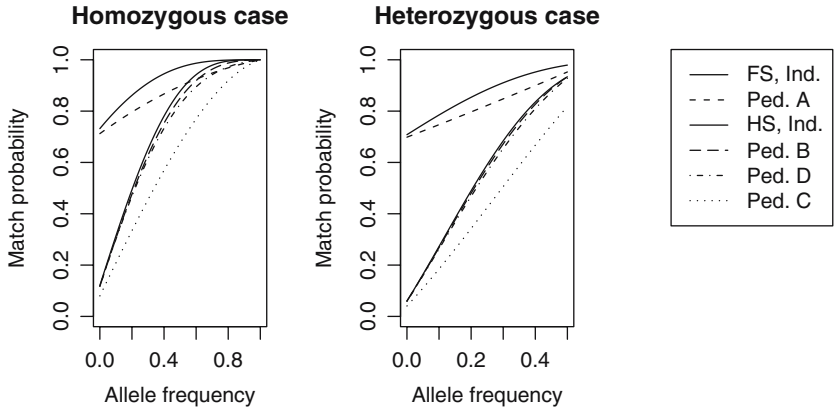

Fig. 2 Single-locus comparisons between various pedigrees. Here we have plotted the relationship between allele frequencies and match probabilities for the situation in which $\mathcal{P}$ has four siblings according to pedigrees *A*, *B*, *C*, and *D*. The two plots represent the situations in which $\mathcal{P}$ is homozygous and heterozygous (with both alleles having equal population frequencies). For comparison with the exact match probabilities (*dashed lines*), we have also plotted the match probabilities that would be calculated in each case by assuming that the genotype of each sibling was independent of the others (*solid lines*)

homozygous, the allele frequency used was that of $\mathcal{P}$'s only allele. In the heterozygous case, we assumed that both of the suspect's alleles had equal frequencies and plotted the match probability against that frequency. A second heterozygous case, in which one of $\mathcal{P}$'s alleles was twice as common as the other was also examined, but yielded a plot quite similar to the equifrequent case presented here.

In addition to the probability calculations based upon the pedigrees in Fig. 1, Fig. 2 also displays the match probabilities (solid lines) that would be obtained if only the number of full siblings or half-siblings were known and analysis proceeded as if the genotypes of these individuals were independent. For example, in the full-sibling case, we calculated p, the probability that a single full sibling of $\mathcal{P}$ shared $\mathcal{P}$'s genotype (using population parameter θ=0.03), then supposed that the number of matching siblings followed a binomial distribution with parameters n=4 and p. This approach is not correct because it ignores the relationships between $\mathcal{P}$'s siblings, but we have included these values as a reference.

In both the homozygous and heterozygous cases, pedigrees *A* and *C*, in which there are strong relationships between the full siblings or half-siblings gave results that differed strongly from the results produced under the independence assumption. The match probabilities for the other pedigrees, *B* and *D*, in which the half-siblings were not as closely related to each other, showed close agreement with each other and with the probabilities generated under independence. Since these two pedigrees showed such close agreement, and to save computational time, we have decided to exclude pedigree *B* from the remaining analyses.

In each of the situations exhibited in Fig. 2, the independence calculation proved to be conservative, that is, it gave a higher estimate of the match probability than the calculation that took into account the relationships among the full siblings or half-siblings. While it is true that probabilities calculated under the independence assumption will generally be conservative, situations can be devised in which the reverse is true.

In addition to comparing the results under the given pedigrees to the probabilities resulting under the indepen-

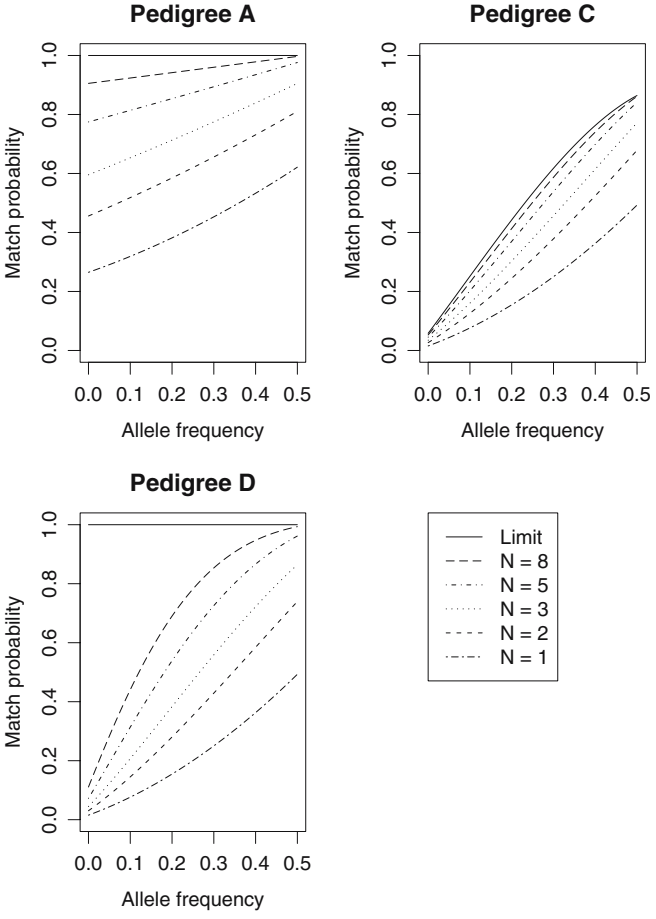

Fig. 3 The effect of the number of siblings. Here, we compare the match probability under pedigrees *A*, *C*, and *D* for various numbers (N=1, 2, 3, 5, 8, ∞) of siblings

dence assumption, we may also compare the pedigrees with each other. Following the principle that the presence of more parents in the pedigree offers more opportunities for alleles matching $\mathcal{P}$'s alleles to enter the family, we see that the pedigree in which the half-siblings are full siblings to each other (pedigree *C*) gives the lowest match probabilities.

One might also note that Fig. 2 was produced under θ=0.03. Plots produced under other reasonable values of $\theta(0 \leq \theta \leq 0.05)$ were produced, but were found to look qualitatively similar to those presented and so were not included here. Generally, the effect of raising or lowering θ was to raise or lower the match probabilities. This effect was substantial for low allele frequencies but rather slight when $\mathcal{P}$ had more common alleles. Throughout the remainder of this discussion, we use a value of θ=0.03, as suggested by the National Research Council [13].

While Fig. 2 shows results for the case in which $\mathcal{P}$ has four full siblings or half-siblings, it may be of interest to examine more carefully how the match probabilities vary with the number of siblings. In Fig. 3, we show the relationship between match probability, allele frequency, and number of full siblings or half-siblings for pedigrees *A*, *C*, and *D* in the case in which $\mathcal{P}$ is heterozygous and both alleles are equifrequent. For each pedigree, the solid line represents the limiting case in which there are infinitely many siblings.

When the relatives of interest are the suspect's full siblings (pedigree *A*), the number of siblings has a strong effect on the match probability for small or moderately sized families, but the effect of each additional sibling decreases as the number of siblings grows.

The results for pedigree *C* are substantially different than the other pedigrees. In this case, $\mathcal{P}$ has *N* paternal half-siblings, but these are all full siblings to each other. Since it is possible that the allele passed to $\mathcal{P}$ from $\mathcal{P}$'s mother will not be present in the mother of the half-siblings, there will always be a considerable chance that none of the half-siblings will share $\mathcal{P}$'s genotype. Hence, pedigree *C* yields lower limiting match probabilities than the other pedigrees examined here.

In pedigree *D*, $\mathcal{P}$ has *N* paternal half-siblings, but each of these has a different mother. Like the full sibling case, the limiting probability as the number of siblings becomes large is 1.0 for this pedigree. In contrast to the previous pedigrees, increasing the number of half-siblings beyond 8 will have appreciable effects on the match probability for pedigree *D* because each additional half-sibling brings with it an additional mother and, hence, a larger chance for alleles matching $\mathcal{P}$ to enter the pedigree.

Multilocus match probabilities

As seen in Figs. 2 and 3, when only one locus is typed, the probability that at least one of the suspect's relatives shares the suspect's profile can be quite large. When numerous loci are typed, however, the match probability decreases substantially. Figure 4 shows the relationship between the match probability and number of loci for pedigrees *A* and *C* with various numbers of siblings. We chose to present results for only one of the half-sibling cases because, for eight or more markers, all half-sibling pedigrees gave similar results. For this figure, the suspect is considered to be heterozygous at each locus with alleles that each have a population frequency of 0.20. This scenario was chosen as a representation of a "typical" profile.

The values used to generate Fig. 4 are a mixture of exact values, values found by simulation, and values found by taking the median between upper and lower bound probabilities. We used exact values where these could be calculated quickly. We used the median between the upper and lower bound probabilities whenever the difference between these was less than 2% of the upper bound. When we were unable to obtain suitably accurate estimates using the bounding method, we estimated the match probability by simulation. For the estimates by simulation, we chose sample sizes large enough to ensure that the standard error be less than 1% of the estimated value.

As can be seen in Fig. 4, the match probability decreases rapidly with an increase in number of loci typed. As a result, when even a moderate number of loci are typed, the probability of finding a match for the suspect's profile

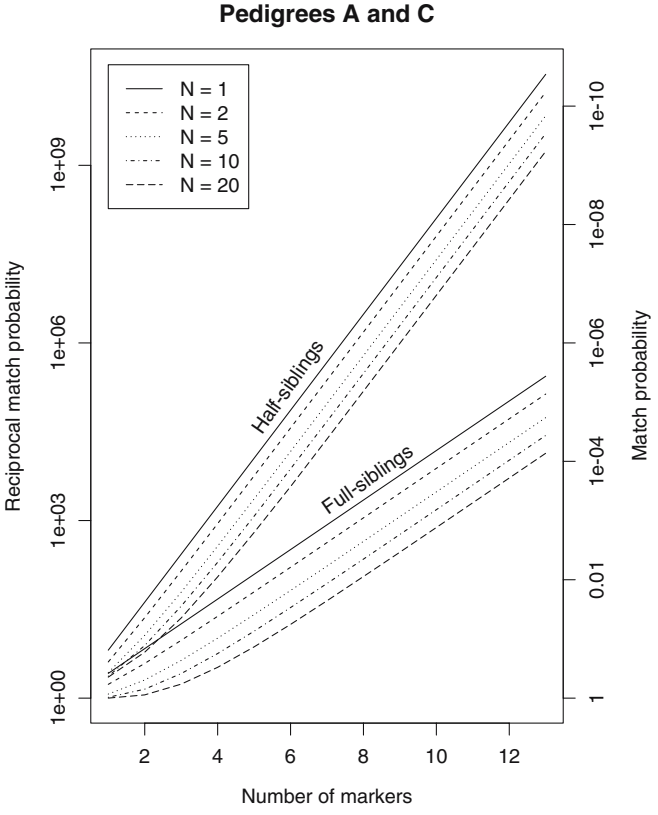

Fig. 4 Multilocus match probabilities viewed on the log scale. This plot depicts the decrease in match probability as the number of loci increases for a "typical" situation in which the suspect's profile is heterozygous at each locus and each allele has a frequency of 0.20. Note that, even for large families, the match probability becomes small when a moderate amount of loci are typed. Pedigree *C* was used for the half-sibling case

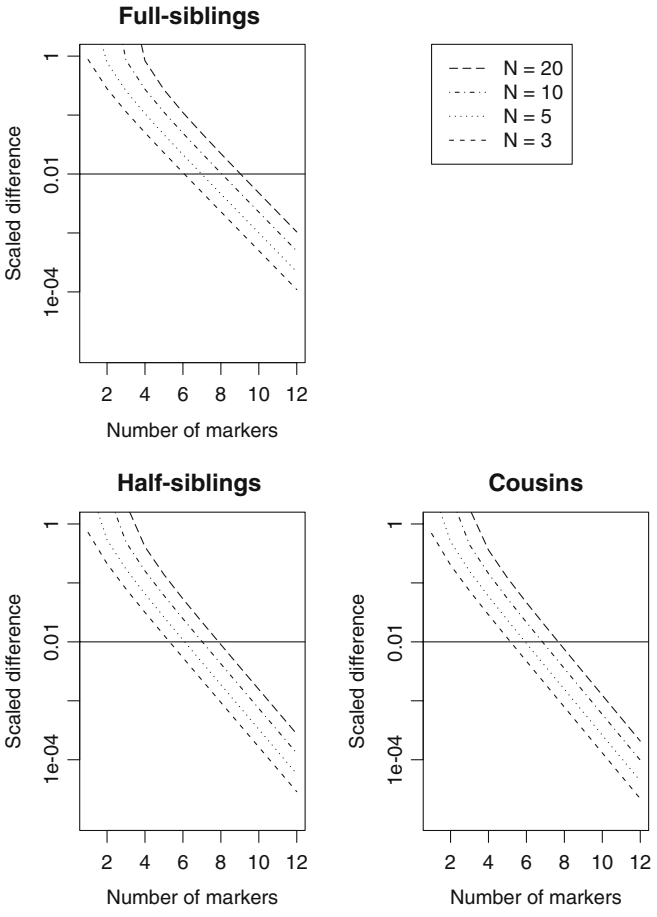

Fig. 5 Scaled differences between upper and lower bounds. In this plot, we show the difference between the upper and lower bounds as a proportion of the lower bound. In each case, these plots represent the situation in which the suspect is heterozygous at each locus and every allele has a population frequency of 0.20

among the relatives is quite small, even when the number of relatives is large.

Continuing with the situation in which the suspect is heterozygous at each locus (and each allele has a population frequency of 0.2), we next investigated the number of loci necessary in order to ensure that the upper and lower bounds on the match probability were close enough together to allow the upper bound alone to be used as the estimated match probability. Recognizing that the bounds will be farthest apart when the probability of two or more matches among the relatives is greatest, we chose to use only pedigree C for the half-sibling case. Since we are looking only at the bounding probabilities, we were able to expand our calculations to include the situation in which the suspect has n first cousins of interest. We chose to examine the case in which the cousins are all full siblings to each other because this is the case in which the upper and lower bounds will be farthest apart.

To evaluate the similarity between the two bounds, we calculated the difference between the upper and lower bounds as a proportion of the lower bound. Figure 5 shows this scaled difference between the bounds for three family types (full siblings, half-siblings, and cousins) for three, five, ten, and 20 relatives. We did not look at the case in which there is just one relative of interest, because in that case the upper bound is exact. Similarly, since the lower bound is exactly the match probability when there are only two relatives being considered, we did not plot that case. The rapid decline in the scaled difference with an increase in the number of loci led us to present our results on the log scale.

One thing to be seen from Fig. 5 is that, when the number of loci is low, there can be substantial differences between the upper and lower bounds. This indicates that these bounds are not helpful in estimating the match probability when few loci have been typed. Hence, in this situation, either exact calculations or Monte Carlo estimations must be used. Perhaps more importantly, the figures show that for all three family types and virtually any reasonable number of relatives of interest, when at least nine loci have been typed, the difference between the upper and lower bounds is at most 1%. This means that the bounding method can be used in these cases and, even better, the upper bound itself will be within 1% of the true match probability.

While Fig. 5 indicates that the upper bound probability may be used as an estimate for the match probability when there are at least nine loci in a "typical" case for which the suspect is heterozygous at each locus and all allele frequencies are 0.20, there may be some question as to whether nine loci is an appropriate cutoff in other situations. To investigate this, we evaluated the scaled difference between the bounds in the "worst reasonable" case scenario for each family situation. The difference between the two bounds is greatest when the probability that a pair of relatives both match the profile is greatest. Since this probability increases when the relatives are most closely related, the family configurations shown in Fig. 5, in which the relatives are all full siblings to each other, are already the worst case possible. Increasing the allele frequencies is the most effective way to increase the probability of multiple matches. Since it is relatively rare for any allele used in these situations to have a frequency of over 0.25, we chose, for our "worst reasonable" case, to use allele frequencies of 0.25. While it is not realistic for the suspect to be homozygous at each locus, the match difference between the bounds (with θ=0.03) is greatest in the full siblings and half-sibling cases for homozygous loci, so we considered the case in which the suspect was homozygous at each locus. When the people of interest are cousins to the suspect (and θ=0.03), the difference between the bounds is slightly greater in the heterozygous case, so for our "worst reasonable" case scenario with cousins, we chose to let $\mathcal{P}$ be heterozygous.

The results of our "worst reasonable" case scenario computations are as follows: With nine loci and ten full siblings, the scaled difference between the bounds is 0.0163, or just over 1.6%. When the suspect has 20 half-siblings, the

scaled difference is 0.0153. Finally, with 20 first cousins, the scaled difference is just 0.00615. Based on these results, we suggest that, as a rule of thumb, if the suspect has no more than ten full siblings (or, in the absence of full siblings, no more than 20 half-siblings and/or cousins) and the suspect has nine (or more) loci at which the alleles both have frequencies of under 0.25, then the match probability can be accurately estimated by simply using the upper bound probability.

Conclusions

We have presented a methodology for calculating the probability that at least one of a suspect's "relatives of interest" shares the suspect's DNA profile. This probability is of interest in forensic settings when a suspect whose profile matches a profile found at a crime scene is known to have a number of relatives who might also share that profile. In cases in which there are few loci and the family configuration is simple, this probability can be found by direct calculation. Alternatively, the match probability can be estimated by using upper and lower bounds to obtain a range of values that includes the actual match probability. When the number of loci is large, these bounds will be close together, allowing the match probability to be well specified. We have shown that, in general, if nine or more loci have been typed, the two bounds are close enough together that the upper bound itself provides a good estimate for the match probability.

In certain circumstances, if the number of loci is neither very small nor large and the family structure is complicated (e.g., the suspect has several half-siblings with a large number of parents between them), direct calculation of the match probability may be infeasible and the bounds described above may be too far apart to provide a useful estimate. In this situation, the match probability may be estimated by Monte Carlo simulation according to the method we described.

Thus, by using one of the three methods presented here, the probability that one of a suspect's full siblings or half-siblings matches the suspect's profile can be evaluated for any number of loci or family configuration.

Acknowledgement This work was supported in part by NIH grant GM45344 and NSF grant DMS9819895.

Appendix

The probability that both individuals in a pair of the suspect's relatives share the suspect's single-locus genotype can be found by summing over all possible genotypes among their parents in a manner similar to that described in Materials and methods. The results for various relative pairs are given as in the following equations, using the notation described previously and the convention that $C_j=(1+j\theta)(1+(j-1)\theta)\ \ldots\ (1+\theta)$.

For the case in which both individuals in the pair are full siblings to $\mathcal{P}$, we derive:

$$\Pr(\text{both match}|P=A_1A_1) = \frac{8M_{1,2}(1+2\theta+M_{1,3})+M_{2,1}M_{2,0}}{16C_2} \tag{10}$$

$$\Pr(\text{both match}|P=A_1A_2) = \left[(1+2\theta)(4M_{1,1}+4M_{2,1}+M_{3,0}) + 12M_{1,1}M_{2,1}\right]/16C_2 \tag{11}$$

The next equations hold when one individual is a full sibling and the other is a half-sibling to $\mathcal{P}$, or (same formula) when both individuals are half-siblings to the suspect but are full siblings to each other.

$$\Pr(\text{both match}|P=A_1A_1) = \frac{M_{1,2}(8M_{1,4}M_{1,3}+6M_{1,3}M_{2,0}+M_{2,1}M_{2,0})}{8C_3} \tag{12}$$

$$\Pr(\text{both match}|P=A_1A_2) = \Big[(1+3\theta)\big(2M_{1,2}M_{1,1}+2M_{2,2}M_{2,1}+10M_{1,1}M_{2,1} + M_{3,0}(M_{1,1}+M_{2,1})\big)+6M_{1,1}M_{2,1}(M_{1,2}+M_{2,2})\Big] /16C_3. \tag{13}$$

For pairs of individuals in which both individuals are half-siblings to the suspect and are half-siblings to each other, the following equations hold:

$$\Pr(\text{both match}|P=A_1A_1) = \frac{M_{1,3}M_{1,2}(3M_{1,5}+1)}{4C_3} \tag{14}$$

$$Pr(\text{both match}|P=A_1A_2) = \Big[(1+3\theta)(M_{1,2}M_{1,1}+M_{2,2}M_{2,1}) + 6M_{1,1}M_{2,1}(M_{1,2}+M_{2,2})\Big]/8C_3 \tag{15}$$

The remaining half-sibling case is the situation in which the two half-siblings are unrelated to each other (i.e., one is related to the suspect through the suspect's mother and the

other is related through the father). The equations for this type of half-sib pair are as follows:

$$\begin{aligned}&\Pr(\text{both match}|P = A_1A_1)\\&= \Big[M_{1,3}M_{1,2}(4M_{1,5}M_{1,4} + 4M_{1,4}M_{2,0})\\&\quad + M_{2,1}M_{2,0}\Big]/4C_4\end{aligned} \tag{16}$$

$$\begin{aligned}&\Pr(\text{both match}|P = A_1A_2)\\&= M_{1,1}M_{2,1}\Big[(1+4\theta)(3M_{1,2} + 3M_{2,2} + M_{3,0})\\&\quad +4M_{1,2}M_{2,2}\Big]/4C_4.\end{aligned} \tag{17}$$

References

1. Ayres KL (2000) Relatedness testing in subdivided populations. Forensic Sci Int 114:107–115
2. Balding DJ, Donnelly P (1995) Inference in forensic identification (with discussion). J R Stat Soc, A 158:21–53
3. Balding DJ, Nicholas RA (1994) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 64:125–140
4. Belin TR, Gjertson DW, Hu MY (1997) Summarizing DNA evidence when relatives are possible suspects. J Am Stat Assoc 92:706–716
5. Brookfield JFY (1994) The effect of relatives on the likelihood ration associated with DNA profile evidence in criminal cases. J Forensic Sci Soc 34:193–197
6. Donnelly P (1995) Nonindependence of matches at different loci in DNA profiles: quantifying the effect of close relatives on the match probability. Heredity 75:26–34
7. Evett IW (1992) Evaluating DNA profiles in a case where the defence is “It was my brother”. J Forensic Sci Soc 31:5–14
8. Fukshanky N, Bär W (2000) Biostatistics for mixed strains: the case of tested relatives of non-tested suspect. Int J Leg Med 114:78–82
9. Griffiths RC (1979) A transition density expansion for a multiallele diffusion model. Adv Appl Probab 11:310–325
10. Hu YQ, Fung WK (2003) Interpreting DNA mixtures with the presence of relatives. Int J Leg Med 117:39–45
11. Fung WK, Chung Y, Wong D (2002) Power of exclusion revisited: probability of excluding relatives of the true father from paternity. Int J Leg Med 116:64–67
12. Lee JW, Lee HS, Park M, Hwang JJ (1999) Paternity probability when a relative of the father is an alleged father. Sci Justice 39:223–230
13. National Research Council (1996) The evaluation of forensic DNA evidence. National Academy Press, Washington DC
14. Weir BS (1994) The effects of inbreeding on forensic calculations. Annu Rev Genet 28:597–621
15. Weir BS (2003) Forensics. In: Balding D, Bishop M, Cannings C (eds) Handbook of statistical genetics, Chap 27. Wiley, Chichester, pp 830–852
16. Weir BS, Hill HG (1993) Population genetics of DNA profiles. J Forensic Sci Soc 33:218–225