*Heredity* **75** (1995) 26–34

*Received 15 September 1994*

# Nonindependence of matches at different loci in DNA profiles: quantifying the effect of close relatives on the match probability

PETER DONNELLY*

*School of Mathematical Sciences, Queen Mary and Westfield College, Mile End Road, London E1 4NS, U.K.*

In comparing a particular DNA profile with that from an unknown (but distinct) individual, matches at different loci between the profiles will not be independent, even in a randomly mating population, because of the presence in the population of relatives of the individuals. The paper contains a theoretical analysis of the extent of this effect on the match probability, for profiling techniques which separately probe different loci. Naive calculation using the product rule could substantially understate the match probability. Past a certain point, the testing of additional loci provides no more information than would be available in discriminating between sibs. The correlation effect described here would be unimportant in criminal casework if close relatives of the suspect, and in particular full-sibs, were excluded as possible culprits. In the absence of such exclusions the current practice of effectively ignoring such relatives in presenting match probabilities could be extremely prejudicial to a suspect, even in cases in which there is no direct evidence to incriminate his/her relatives.

**Keywords:** DNA profiles, forensic science, minisatellites, population genetics, product rule.

## Introduction

Forensic applications of DNA profiling typically involve the comparison of a 'crime profile', obtained from DNA taken from the scene of the crime, and a 'suspect profile', obtained from a particular suspect. Differences between these two profiles that cannot be attributed to experimental effects will usually result in the exclusion of the suspect from further investigation. On the other hand, if the two DNA profiles are judged to match, this will be interpreted as evidence tending to associate the suspect with the crime. The strength of this evidence against the suspect depends on how common the profile in question is. This is usually measured by the match probability, the probability that an unknown individual chosen from an appropriate population will match the crime profile.

The calculation of the match probability, and the validity of the assumptions which underpin the so-called product rule, have attracted considerable controversy. For a discussion, see the report of the National Research Council (1992) and references therein. For a more recent review, and references, see Roeder (1994). Much of the concern has centred on the extent of nonindependence of allele possession at particular loci (violations of Hardy–Weinberg equilibrium) and across loci (linkage disequilibrium), owing to the presence of stratification, and hence non-random mating, within the human population. In such circumstances the product rule, which assumes independence within and between loci, is likely to understate the chance of a match and hence to overstate the strength of the evidence against a suspect. Here we consider a separate issue. The assumption of independence may overstate the chance of a match even if the population is randomly mating.

In this paper we provide a quantitative analysis of the qualitative argument stated briefly in Donnelly (1992). Loosely speaking, it will initially be unlikely that the chosen individual will be related to the suspect. However, after the observation of matches at some loci, it is relatively much more likely that the individuals involved are related (precisely because matches between unrelated individuals are unusual) in which case matches observed at subsequent loci will be less surprising. That is, knowledge of matches at some loci will increase the chances of matches at subsequent loci, in contrast to the independence assumption. Our

*Current address: Departments of Statistics and Ecology and Evolution, University of Chicago, 5734 University Avenue, Chicago, IL 60637, U.S.A.

purpose in this paper is to investigate the magnitude of this effect. We do so by a theoretical analysis. The small magnitude of the probabilities involved mitigates against a direct empirical study.

Throughout, we will phrase the discussion in the terms appropriate to criminal casework, describing the individuals concerned as criminal and suspect. Nonetheless, the analysis has implications more broadly. The match probability concerns the matching of distinct individuals. There has been considerable discussion in the forensic context of the appropriate choice of reference population (the population to which the match probability applies); see for example Roeder (1994). We do not address this issue here (for a discussion, see Balding & Donnelly, 1995) assuming instead that the population in question is specified.

Most current applications of DNA profiling involve single locus probes of minisatellite or microsatellite loci. Each probe used binds to a different locus, the loci being chosen because of the variability they exhibit. In current practice fragments are separated by electrophoresis and the alleles are located by probing a Southern blot, or for some microsatellite-based techniques, by flourescent dyes and an ABI sequencer after PCR amplification. Differences between alleles primarily reflect variation in the number of a tandemly repeated sequence (see for example Jeffreys *et al.*, 1991a).

Quantification of the nonindependence effect at issue here requires a rather intricate analysis of genealogical history at each locus. The more technical details are given in the Appendix. In the next section we describe the framework for the analysis and present numerical comparisons which illustrate the effect. The final section of the paper discusses various practical consequences.

We show below that the presence of close relatives of the suspect in the relevant population can substantially increase the match probability. Most of this effect relates to the possibility that the criminal and suspect are full-sibs. As a consequence, the concerns raised here could be substantially avoided if full-sibs of the suspect were excluded as possible culprits, either through further DNA profiling or through more conventional scientific or investigative means. It would thus be helpful if attempts at such exclusion were to become routine in cases involving DNA evidence. We note, however, that in some legal jurisdictions, lack of co-operation from the relatives concerned could severely hamper such attempts.

At present, cases regularly come to trial in which close relatives of the suspect have not been excluded. In some current practice, forensic scientists are explicit that a reported match probability relates only to unrelated individuals from the relevant population. They may add that related individuals are 'more likely' to match, but do not usually try to allow for this quantitatively. Unless close relatives of the suspect have been excluded, the analysis below shows that a match probability calculated using the product rule could overstate the strength of the DNA evidence, possibly substantially.

Of course it is well known that close relatives are much more likely than unrelated individuals to have matching DNA profiles. The NRC report (National Research Council, 1992, p. 87) notes as a consequence that 'whenever there is a possibility that a suspect is not the perpetrator but is related to the perpetrator, this issue should be pointed out to the court' and Evett (1992) discusses the calculation of match probabilities in cases in which the defence specifically makes such a claim. An important consequence of the analysis here is that the effect can be important *even in cases in which there is no evidence to incriminate relatives of the suspect*, a point which does not seem to be fully appreciated in practical casework.

## Effect on the match probability

In criminal casework, the match probability relates to the probability that a member of the reference population will match the profile obtained from the crime sample, and the probability will in general depend on the multilocus genotype of the crime profile. Here, to obtain a general conclusion, we will consider the related probability (which we also call the match probability) that two distinct individuals will have matching DNA profiles, without the additional information as to the genotype of one of the profiles.

We consider $k$ unlinked loci which contribute to a DNA profile, in a randomly mating population. Throughout, we assume that the profiling technique distinguishes the alleles at the different loci. This is the case for the single locus probes of minisatellite loci in current use, and for probes of particular microsatellite loci. The quantitative analysis which follows does not apply directly to multilocus probes, but the qualitative argument for nonindependence will still apply.

We denote by $g$ the probability that a pair of paternal genes, taken from the same locus in different individuals, is descended from the same individual in the previous generation. This is the probability that the individuals concerned have the same father. For ease of exposition, we will assume that the probability that two different individuals have the same mother is also $g$. The value of $g$ will depend on the demographic history of the population. We will not make detailed assumptions about the demography, instead expressing our results in terms of the parameter $g$. It can be shown (Kingman, 1982) that under quite general demographic

models, $g$ will be of order $N^{-1}$, where $N$ is the population size. We assume that the value of $g$ is the same over the recent history of the population. In contrast to most population genetic studies, the analysis here depends on segregation events only over the past few generations, so this assumption may be reasonable for many human populations.

Some of the loci used for DNA profiling are known to have extremely high mutation rates. In addition, the mutation mechanism (that is, the way in which mutation is likely to change repeat copy number) at these loci is not well understood. We will denote by $\mu_T$ the sum of the mutation probabilities (per gamete per generation) at the $k$ loci. We will assume that if there is a mutation event since the most recent common ancestor gene of two current genes, then those two genes will not match. Ignoring in this way the possibility that the effects of several mutations on repeat copy number will cancel out, or that a mutation may change repeat copy number in such a way that the resulting alleles still fall within match guidelines, means that the calculations below will understate the match probability. As noted above, the nonindependence we discuss is the consequence of ancestral history over only the past few generations, so the error induced by the assumption will be small.

For ease of exposition, we will assume that the probability, $M_1$, that a pair of alleles matches, is the same at each of the $k$ loci under consideration. Extension of the analysis to allow for different match probabilities at each locus is straightforward. The analysis proceeds in two stages. First we consider only the paternal alleles at each of the loci in two distinct individuals. We write $M_k$ for the probability that the parental alleles in these individuals match at each locus. In view of our assumptions, this will also be the probability that the maternal alleles at each locus in two distinct individuals match. The second stage uses these probabilities to calculate match probabilities.

The analysis itself involves a rather intricate study of the joint genealogical histories of the loci in question in the two individuals. We present the conclusions here. The details are given in the Appendix.

Equations 1 and 2 allow lower bounds on the probabilities $M_k$ (relating to the match of paternal alleles at each locus between two individuals) to be calculated in terms of the probability $M_1$ for a single locus. These may then be evaluated numerically.

The product rule would calculate $M_k$ as $M_1^k$. Figure 1 illustrates the dependence of the bounds for $M_3$ and $M_4$ on $g$ for various values for $M_1$. In each case, the value of the product rule calculation is effectively that at the left edge of the appropriate curve. For example, at four loci with $M_1 = 0.01$ the product rule calculation would give $10^{-8}$. The differences between the various curves (which do allow for the presence of close relatives in the population) and the product rule calculation (which does not) is most marked for larger values of $g$, corresponding to relatively smaller population sizes. The calculations in Fig. 1 (and in Fig. 2 below) use $\mu_T = 0.06$, which appears plausible for four-locus systems in current use (Brinkmann *et al.*, 1991), but their dependence on the exact value of $\mu_T$ is slight.

The probability, $\mathcal{M}_k$ say, that the paternal alleles match and the maternal alleles match between two different individuals at the $k$ loci may be calculated from the probabilities $M_k$. There is an additional correlation effect because in human populations the conditional probability, $c$ say, that individuals who share one parent in fact share both parents, is appreciable. Equation 3 in the Appendix allows for this effect in arriving at a lower bound for $\mathcal{M}_k$ in terms of the probabilities considered above.

In practice, it is not usually known which alleles at each locus are maternal and which are paternal. The probability that the $k$-locus genotype of two distinct individuals will match is the sum over the probabilities associated with each of the $(2^k)$ possible assignments. See eqn 4 in the Appendix, which then allows lower bounds for the match probability to be evaluated numerically. The results are illustrated in Fig. 2, where the probability of a match is plotted as a function of $g$ for various values of $M_1$.

Figure 2 also shows (slanted arrows on the $y$-axis) the value which would be calculated for the match probability if the product rule were applied $(2^k M_1^{2k})$. As an example of the magnitude of the effect, consider a four-locus genotype involving alleles with 5 per cent frequency in a population with effective size of about 10000 (i.e. $k=4$, $M_1 = 0.05$, $g = 10^{-4}$). The product rule would calculate the probability of a match between two distinct individuals as $6.25 \times 10^{-10}$ whereas it actually exceeds $2 \times 10^{-7}$. The difference is even more pronounced for rarer alleles: $1.6 \times 10^{-15}$ vs. $10^{-7}$, for alleles with a frequency of 1 per cent. Particularly for larger values of $g$ (smaller population sizes) and/or small values of $M_1$ (and hence extremely small match probabilities using the product rule), the difference between the match probability allowing for the presence of relatives, and the calculation given by the product rule, can be very substantial.

For the purposes of illustration, Fig. 2 uses the value of 0.7 for $c$. The appropriate value will depend on, for example, cultural aspects of the population in question and will vary from population to population. (The value of 0.7 may be reasonable, or possibly an underestimate, for UK or US populations; see for example Ferri (1984) or Spanier & Furstenberg (1987) and

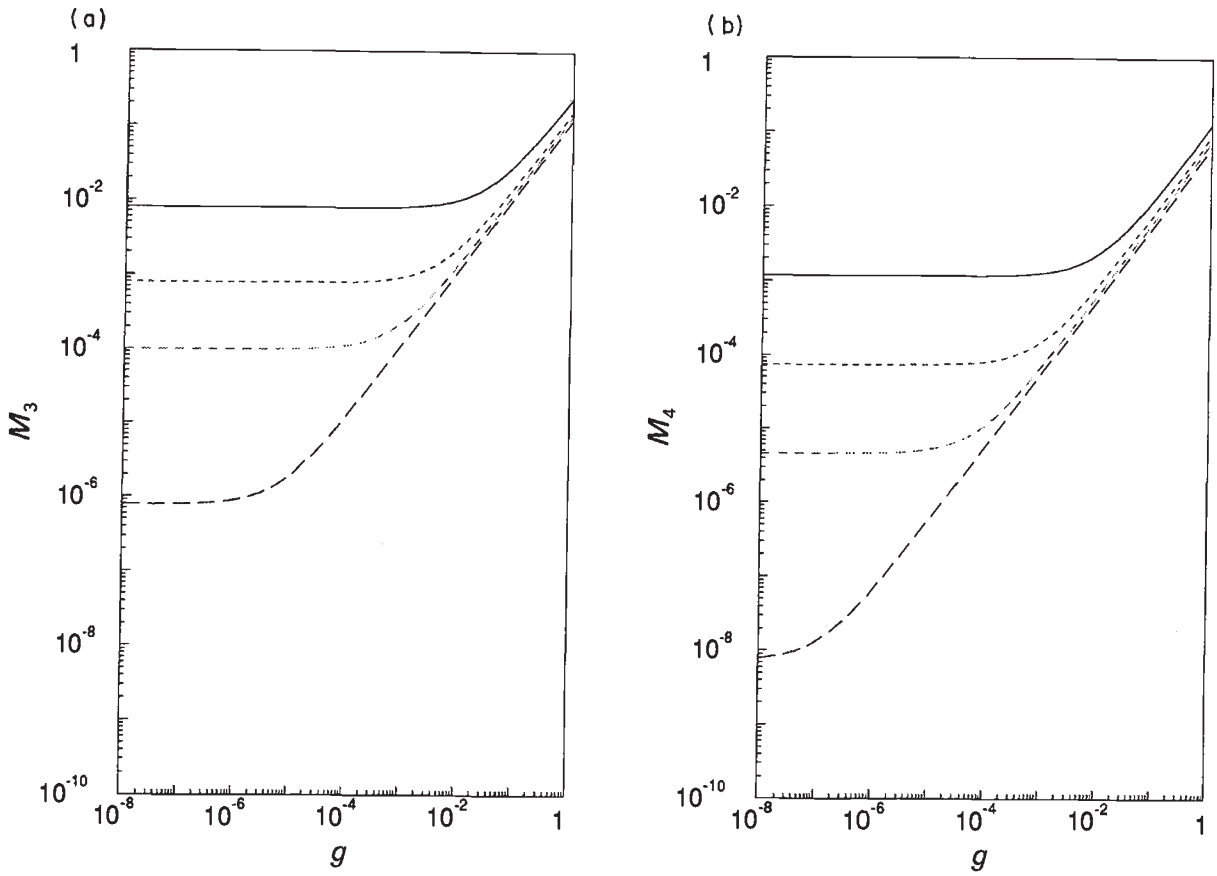


**Fig. 1** The lower bound for $M_k$, the probability that paternal (or maternal) criminal and suspect alleles match at $k$ loci, as a function of the genealogical parameter $g$, for various values of the unconditional match probability $M_1$: $M_1 = 0.2$ (——), $M_1 = 0.1$ (- - -), $M_1 = 0.05$ (· · · · · ·), $M_1 = 0.01$ (— —). Calculated from eqns 1 and 2 with $\mu_T = 0.06$. With the conventional independence assumption $M_k$ is taken to be $M_1^k$. (a) $k = 3$.(b) $k = 4$.

references therein.) Increasing the value of $c$ will increase the match probability, and hence increase the extent to which the product rule understates that probability. The effect (in our model) is roughly linear so that, for example, an increase from $c = 0.7$ to $c = 0.84$ will increase the lower bound in Fig. 2 for the match probability by about 20 per cent.

The most extreme possibility, as far as the correlations discussed here are concerned, occurs when the criminal and suspect are full-sibs. In our model, the probability of this is $gc$ and in this case the chance that all pairs of alleles at the loci in question are descended from the same parental allele is $4^{-k}$. Thus the match probability is at least $(1-2\mu_T)gc/4^k$. This term is actually quite close to the lower bounds evaluated in Fig. 2, unless $M_1$ is relatively large.

## Discussion

Most of the differences between the product rule calculation and the calculation which allows for close relatives result from the possibility that the criminal and suspect are full-sibs. The problems arising from such correlations could be largely avoided in criminal casework if full-sibs of the suspect were excluded as possible culprits. Attempts to obtain such exclusions do not seem to be routine in current investigative procedures. Our analysis has shown that if full-sibs of the suspect have not been excluded, presentation of a match probability based on the product rule could overstate, possibly substantially, the strength of the DNA evidence.

It must also be remembered that the match probability is *not* the probability that the suspect is innocent. The relationship between the match probability and the question of interest to a court, namely what is the chance that this particular suspect is guilty of this crime, involves a number of subtleties. For a detailed analysis, see Balding & Donnelly (1995). One consequence of such an analysis is that concerns about assumptions underlying the product rule cannot be dismissed because they 'do not make a small probabi-

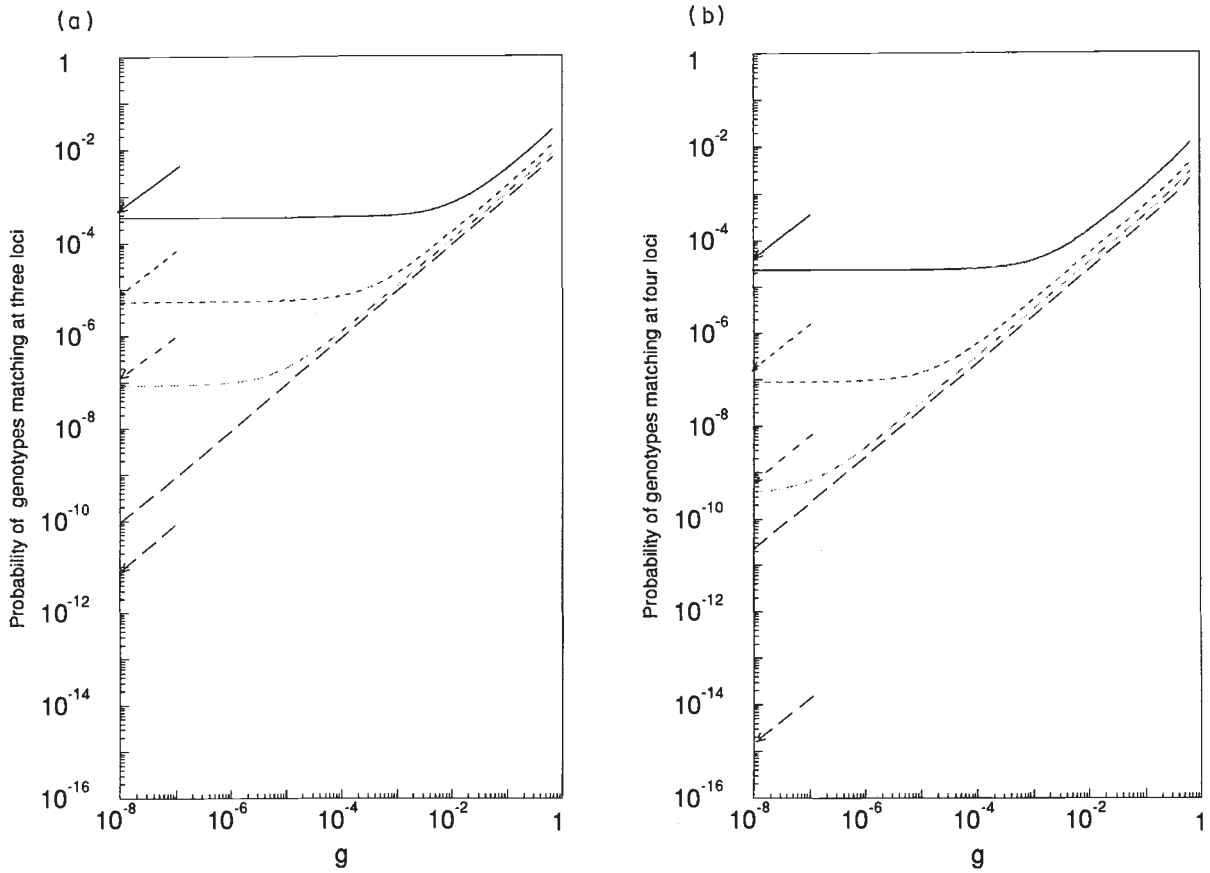


**Fig. 2** Lower bounds for the probability of observing matches between criminal and suspect genotypes at the $k$ loci as a function of the genealogical parameter $g$, for various values of the unconditional match probability $M_1$. $M_1 = 0.2$ (——), $M_1 = 0.1$ (- - -), $M_1 = 0.05$ (··· ··· ···), $M_1 = 0.01$ (— —). Calculated from eqns 3 and 4 with $\mu_T = 0.06$, $c = 0.7$. The arrows on the $y$-axis indicate the match probabilities calculated for each value of $M_1$ using the product rule (i.e. $(2M_1^2)^k$). (a) $k = 3$, (b) $k = 4$.

lity large'. Concerns about the correlation effect with relatives cannot thus be ignored as not of practical relevance simply because the actual values obtained in Fig. 2 are still rather small for values of $g$ which might be thought relevant to particular cases. Indeed, Balding & Donnelly (1995; see also Balding & Nichols, 1994) show that *regardless of the population size* (and hence of the value of $g$), in cases in which the evidence other than the DNA match does not distinguish between the suspect and several of her/his sibs for example, the DNA evidence may not be sufficient for a conviction. In other cases the effect of relatives will depend on the other evidence, but it will often be seriously prejudicial to the suspect to ignore them (Balding & Donnelly, 1995). Suspects themselves, and their defence lawyers at trial, may be reluctant for obvious reasons to introduce the possibility that one of their relatives is the true culprit. In view of this it seems particularly important for forensic scientists to allow for this effect when originally presenting their evidence.

Our analysis has considered only randomly mating populations. The correlation effect we describe will be less marked in populations in which marriages between close relatives occur less frequently than under random mating, and more marked when close relatives marry more frequently than under random mating. The numerical difference between match probabilities (which allow for relatedness effects) in these cases and the random mating case we consider here is likely to be small. As noted above, much of the effect here results from the possibility that two members of the population are full-sibs. This is not changed by inbreeding or outbreeding.

The presence in the population of close relatives of the suspect will induce correlations in structured populations for the same reasons as in the unstructured case. We do not give a full analysis, but note that Balding & Nichols (1994) discuss the calculation of the probability that the DNA profiles of two related individuals will match, for various particular

relationships, under certain assumptions about the effect of population structure.

Several empirical studies (for example Weir, 1992; Risch & Devlin, 1992a,b; see also Roeder, 1994) have failed to detect evidence of nonindependence of matches at various single locus probes. Detection of the effects described here is unlikely to be feasible using current data: the small magnitude of the probabilities which must be compared, and the sizes of current databases, make accurate comparison of, say, four-locus match probabilities difficult. Such comparisons would also be extremely sensitive to the numbers of close relatives actually included in databases, an aspect in which, for various reasons, the databases may not be representative of the populations in question. That the product rule calculation could differ from the match probability in spite of the failure of hypothesis tests to reject the independence assumption reinforces a point made previously by Evett *et al.* (1993) that performing statistical tests of various assumptions underlying the product rule is not directly relevant. It is more appropriate to attempt to assess the magnitude and effect of departures from the assumptions. As we noted above, apparently 'small' discrepancies could be important in particular cases.

It has been suggested that one way of overcoming current criticisms in legal applications of DNA profiling is simply to probe more loci (see for example the report in Roberts, 1991). It follows from the analysis above that past a certain point the additional probes provide no more information than they would in distinguishing between full-sibs. While they are thus not uninformative, they may be substantially less discriminating than initially thought.

The point at which the discriminating power of additional probes diminishes depends on how informative the loci already tested are. The sharing of rare alleles between profiles decreases the match probability but it *increases* the correlation effect discussed here. For example, if one were to exploit variability within the repeat units at minisatellites (Jeffreys *et al.*, 1991b) in addition to the variation in repeat number, a single locus would be very informative, but a second locus may convey vastly less additional information than might be expected. In contrast, microsatellite loci typically exhibit less variation in repeat copy number than do minisatellites. More loci may be needed for discriminating profiles, but the reduction in the information content of subsequent probes will be less. This may be a factor in the adoption of new profiling technologies. There are of course other issues involved, including the cost of, and time required for, various procedures, their robustness to the conditions of crime samples, and variability and measurement errors in the techniques used.

Our purpose here has been to investigate quantitatively the effect on match probabilities of correlation effects caused by the presence in the relevant populations of relatives of the suspect. The analysis has of necessity been a theoretical one. The conclusion is that the effect can be substantial.

It is not suggested that such an analysis should be undertaken in connection with particular cases of forensic identification. There are several reasons for this. The first is that real populations are not randomly mating and in any case the parameter $g$ may not be easy to assess. Furthermore, in a particular case the relevant questions are different in an important respect from those addressed above. Instead of asking 'what is the probability that two individuals from the population will match', we must ask the conditional question, 'what is the probability that a person chosen from the population will match the particular profile left at the scene of the crime, or equivalently, the particular profile of the suspect'. In addition, information will typically be available about the numbers of relatives of various degrees possessed by the suspect, in contrast to the analysis above which effectively averages these numbers over the population as a whole. In this setting, probabilities that particular relatives will match the DNA profile of the suspect are relevant. These are given in Balding & Nichols (1994). The method for incorporating the effect of the other evidence (in a manner consistent with the laws of probability) in particular cases, including available information about relatives of the suspect who have not otherwise been excluded, and for combining specific probabilities that relatives of particular degrees will match, is described in Balding & Donnelly (1995).

## Acknowledgements

I acknowledge helpful discussions with David Balding and Richard Nichols, and thank Adrian Roe for the preparation of the figures. This work was supported in part by SERC grants GR/F 98727, GR/G 11101, and B/AF 1255.

## References

BALDING, D. J. AND DONNELLY, P. 1995. Inference in forensic identification. *J. R. Statist. Soc. A* (in press).

BALDING, D. J. AND NICHOLS, R. A. 1994. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci. Int.*, **64**, 125–140.

BRINKMAN, B., RAND, S. AND WEIGAND, P. 1991. Population and family data RFLP's using selected single- and multi-locus systems. *Int. J. Leg. Med.*, **104**, 81–86.

DONNELLY, P. 1992. Discussion of statistical inference in crime investigation using deoxyribonucleic acid profiling (by D.

A. Berry, I. W. Evett and R. Pinchin). *Appl. Stat.*, **41**, 524–525.

EVETT, I. W. 1992. Evaluating DNA profiles in the case where the defence is It was my brother. *J. Forensic Sci. Soc.*, **32**, 5–14.

EVETT, I. W., SCRANAGE, J. AND PINCHIN, R. 1993. An illustration of the advantages of efficient statistical methods for RFLP analysis in forensic science. *Am. J. Hum. Genet.*, **52**, 498–505.

FERRI, E. 1984. *Stepchildren: A National Study.* Eastern Press, London.

JEFFREYS, A. J., ROYLE, N. J., PATEL, I., ARMOUR, J. A. L., MACLEOD, A., COLLICK, A., GRAY, I. C., NEUMANN, R., GIBBS, M., CROSIER, M., HILL, M., SIGNER, E. AND MONCKTON, D. 1991a. Principles and recent advances in human DNA fingerprinting. In: Burke, T., Dolf, G., Jeffreys, A. J. and Wolff, R. (eds) *DNA Fingerprinting: Approaches and Applications*, pp. 1–19. Birkhäuser, Basel.

JEFFREYS, A. J., MACLEOD, A., TAMAKI, K., NEIL, D. L., MONCKTON, D. G. 1991b. Minisatellite repeat coding as a digital approach to DNA typing. *Nature*, **354**, 204–209.

KINGMAN, J. F. C. 1982. Exchangeability and the evolution of large populations. In: Koch, G. and Spizzichino, F. (eds) *Exchangeability in Probability and Statistics*, pp. 97–102. North Holland, Amsterdam.

NATIONAL RESEARCH COUNCIL OF THE USA 1992. *DNA Technology in Forensic Science.* National Academy Press, Washington, DC.

RISCH, N. J. AND DEVLIN, B. 1992a. On the probability of matching DNA fingerprints. *Science*, **255**, 717–720.

RISCH, N. AND DEVLIN, B. 1992b. DNA fingerprint matches. *Science*, **256**, 1744–1746.

ROBERTS, L. 1991. Fight erupts over DNA fingerprinting. *Science*, **254**, 1721–1723.

ROEDER, K. 1994. DNA fingerprinting: a review of the controversy, *Stat. Sci.*, **9**, 222–278.

SPANIER, G. B. AND FURSTENBERG, F. F. 1987. Remarriage and reconstituted families. In: Sussman, M. B. and Steinmetz, S. K. (eds) *Handbook of Marriage and the Family*, pp. 419–434. Plenum Press, New York.

WEIR, B. S. 1992. Independence of VNTR alleles defined as fixed bins. *Genetics*, **130**, 873–887.

## Appendix

Our basic argument is a recursive one; we ask about the ancestry, in the previous generation, of each of the alleles in the profiles of the two individuals concerned.

As an illustration, consider the case of two loci, and focus attention on the paternal alleles in each individual at these loci. For definiteness, we will call the two individuals criminal and suspect. There are various possible configurations for the ancestral alleles in the previous generation.

1 It may be the case that suspect and criminal have distinct fathers and that both paternal alleles in the criminal originate from the same grandparent and both paternal alleles in the suspect originate from the same grandparent.

2 A second possibility is that both individuals have distinct fathers but that the paternal alleles at the two loci in the current generation in both individuals are descended from more than two of their four grandparents.

3 Both individuals may have the same father and the pairs of alleles in the two individuals at each locus are descended from the same allele in the father.

4 Both individuals may have the same father but at one or both loci, the relevant criminal allele is descended from the father's maternal allele and the suspect's allele from the father's paternal allele (or conversely).

Figure 3 illustrates a particular example of each case.

Now consider the effect of each of these possibilities on the possible matching at the two loci of the two pairs of paternal alleles, one member of each pair being from the criminal and one from the suspect. Throughout this discussion we assume that identical alleles will match and we focus on the case in which all of the alleles in the current generation are descended without mutation from the respective alleles in the previous generation. Case 1 actually covers four different subcases corresponding to whether the paternal criminal alleles are both descended from his/her (paternal) grandfather or both from his/her (paternal) grandmother and whether the paternal suspect alleles are both descended from the relevant grandfather or the relevant grandmother. If criminal and suspect alleles are both descended from the relevant grandfather, then the pairs of paternal alleles in criminal and suspect will match if and only if the pairs of paternal alleles in the criminal's father and the suspect's father match. In other words, for this genealogical history the situation is the same in the previous generation and, conditional on this genealogy and no mutation, the match probability is still $M_2$. If criminal and suspect alleles are both descended from the relevant grandmother, then the pairs of paternal alleles in criminal and suspect will match if and only if the pairs of maternal alleles in the criminal's father and the suspect's father match. Under our assumptions this last event also has probability $M_2$. In the other two cases, the paternal alleles in one of the individuals will be copies of their father's paternal alleles while those from the other individual will be copies of their father's maternal alleles. Conditional on this genealogy and no mutation, they will match if and only if the paternal alleles at the loci in question in one particular individual in the previous generation match the maternal alleles at the loci in question in another particular individual in the previous generation.

In case 2, the matching of the alleles at one locus will depend on the genealogical history of one pair of grandparents, while matching at the other locus will

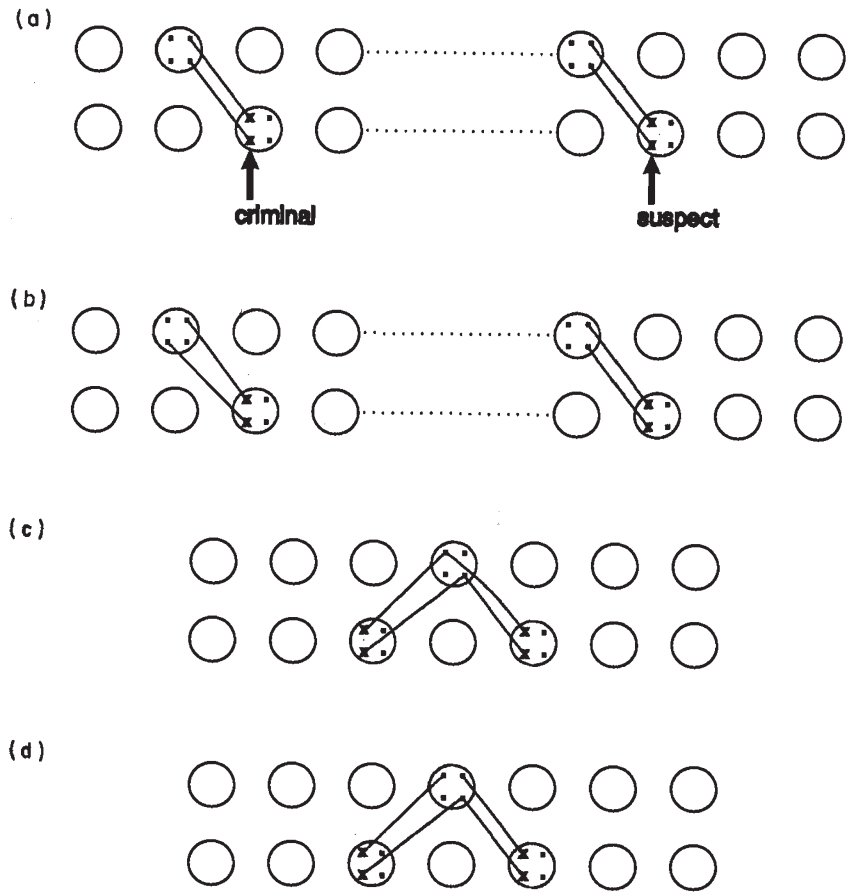


**Fig. 3** Some possibilities in tracing lineages back one generation with $k = 2$ loci. The parental generation is represented above the current generation. Pairs of alleles at each locus are shown in rows (paternal on the left, maternal on the right) and the alleles under consideration in suspect and criminal are marked with a cross. Lines link alleles with the allele from which they are descended in the previous generation. (a) Chosen (paternal) criminal alleles are all descended from the maternal allele of the criminal's father and the chosen (paternal) suspect alleles are all descended from the maternal allele of the suspect's father. (b) At one locus the alleles in question are descended from both paternal grandmothers, at the other locus they are descended from the criminal's paternal grandfather and the suspect's paternal grandmother. (c) Criminal and suspect have the same father and the pairs of alleles in question at the two loci are descended from the same allele in the previous generation. (d) Criminal and suspect have the same father, but at one locus the criminal allele in question is descended from the father's paternal allele while the suspect allele is descended from the father's maternal allele.

depend on the genealogical history of a distinct pair of grandparents. (Here distinct means that the pairs of grandparents are not identical. It may be, however, that one individual appears in both pairs.) We will see below that in this case, matching events at the two loci are effectively independent, so that conditional on this genealogy and no mutation, the match probability is $M_1^2$, as given by the product rule.

In case 3 both pairs of alleles will be identical. That is, conditional on this genealogy and no mutation, the match probability is 1.

Treatment of case 4 also involves consideration of subcases. It may be that at one of the two loci, the alleles in criminal and suspect are descended from the same allele in the father, but that at the other locus, one is descended from the maternal allele in the father and one from the paternal allele. Conditional on this genealogy and no mutation, the alleles at one locus are sure to match, and those at the other will match with probability $M_1$, by the definition of $M_1$ and the assumption of random mating. In the other subcase, at both loci the alleles in criminal and suspect are descended one from the paternal allele in the father and one from the maternal allele in the father. Conditional on this genealogy and on no mutation, the match probability is the same as that in the final two subcases treated under (1) above, namely the probability that the paternal alleles at the loci in question in one particular individual in the previous generation match the maternal alleles at the loci in question in another particular individual in the previous generation. (Our assumption of random mating means that this probability is the same as the probability that the paternal alleles match the maternal alleles at the loci in question in a single individual.)

We now extend the consideration of possible ancestry for two loci to all $k$ loci under consideration and calculate the probabilities associated with the various events of interest.

Consider the paternal pairs of alleles at each of the $k$ loci in suspect and criminal. Recall that we are writing $M_k$ for the probability that the alleles in each pair match each other. We will write $M_l$, $l = 1, 2, \ldots, k$, for the probability that the paternal pairs of alleles match at $l$ loci. By assumption, this probability is the same for any subset of $l$ of the $k$ loci. (Of course, these probabilities also apply to matches of the maternal pairs of alleles at each locus in the two individuals.)

For a particular one of the alleles under consideration there are two possibilities as to the paternal grandparent from which it is descended. Thus at a particular locus there are four possible choices for the paternal grandparents from which the pair of alleles (one in each of suspect and criminal) is descended. Writing the grandparent from which the suspect allele is descended first, these are {grandfather, grandfather}, {grandmother, grandmother}, {grandfather, grandmother} and {grandmother, grandfather}. Denote by $K_1$, $K_2$, $K_3$ and $K_4$, respectively, the number of the $k$ loci for which each of these choices of grandparents applies. At a particular locus, the choice of grandparent for the suspect allele is independent of that for the criminal allele, and each of the four choices is equally likely. As the loci are assumed unlinked, the choices associated with different loci are independent. It follows that the variables $(K_1, K_2, K_3, K_4)$ are distributed multinomially with parameters $k$ and $(1/4, 1/4, 1/4, 1/4)$.

The key observation is that (under our assumptions, notably random mating) if the suspect and criminal have distinct fathers, for two loci associated with different choices of grandparents the events that the pairs of alleles match at each locus are effectively independent, because they depend on the genealogical history of nonidentical pairs of individuals. (They are still not strictly independent in most population genetics models because the assumption of fixed or regulated population size induces correlations in sibship sizes between different families, but this effect will be of order $g^2$ (for example, Kingman, 1982) in such models and we will ignore it. Also if the two loci make the same choice of grandparent for one of the two individuals, there will be correlations in the mutation processes in their ancestry. This could in principle be incorporated explicitly, but as the dependence on mutation rates is small we will neglect it.) This ('effective') independence applies to sets of loci associated with different choices of grandparents, for the same reasons.

Now suppose $(K_1, K_2, K_3, K_4)$ were known and that suspect and criminal had different fathers. The $K_1$ loci associated with the choice {grandfather, grandfather} will have alleles which match with probability $M_{K_1}$. (We define $M_0$ to take the value 1.) The $K_2$ loci associated with {grandmother, grandmother} will have alleles which match with probability $M_{K_2}$. The $K_3$ loci associated with {grandfather, grandmother} will have alleles which match with probability $\tilde{M}_{K_3}$, where $\tilde{M}_l$ is defined to be the probability that the maternal alleles at $l$ of the loci under consideration in one individual match the paternal alleles at those loci in another individual. (We define $\tilde{M}_0$ to take the value 1.) Finally, the $K_4$ loci associated with {grandmother, grandfather} will have alleles which match with probability $\tilde{M}_{K_4}$. These events for the four sets of loci are independent, for the reasons given above.

Now suppose $(K_1, K_2, K_3, K_4)$ were known and that suspect and criminal had the same father. In the absence of mutation, all the pairs of alleles at the $K_1 + K_2$ loci associated with the first two choices of grandparents will be identical because each member of the pair is descended without mutation from the same allele in the previous generation. At the remaining $K_3 + K_4$ loci, the pairs will match with probability $\tilde{M}_{K_3+K_4}$.

Averaging over the distribution of the random variables $(K_1, K_2, K_3, K_4)$ then gives

$$M_k \geq (1-2\mu_T)\{(1-g)\mathrm{E}(M_{K_1}M_{K_2}\tilde{M}_{K_3}\tilde{M}_{K_4}) + g\mathrm{E}(\tilde{M}_{K_3+K_4})\}. \quad (1)$$

The inequality here, and in eqns 2 and 3 below, arises from our assumption that any mutation to any of the alleles under consideration since the previous generation necessarily implies a nonmatch. This assumption will cause our calculation to understate the true match probability.

Similar arguments apply to the matching of maternal alleles in one individual to paternal alleles in another, except that in this case the alleles in any pair cannot have a common ancestor in the previous generation. Thus

$$\tilde{M}_k \geq (1-2\mu_T)\mathrm{E}(M_{K_1}M_{K_2}\tilde{M}_{K_3}\tilde{M}_{K_4}). \quad (2)$$

Successive application of eqns 1 and 2 allows the probability $M_k$ to be bounded in terms of $M_1$ and the parameters $g$ and $\mu_T$.

In considering matches of paternal *and* maternal pairs of alleles in criminal and suspect at each of the loci there is the additional complication that the individuals concerned may share zero, one or two parents. Recall that we use $c$ to denote the conditional probability that individuals who share one parent also share the other parent. An argument similar to the one above, incorporating this additional complication, then gives

$$\mathscr{M}_k \geq (1-4\mu_T)[(1-2g+gc)(\mathrm{E}(M_{K_1}M_{K_2}\tilde{M}_{K_3}\tilde{M}_{K_4}))^2 + 2g(1-c)\mathrm{E}(\tilde{M}_{K_3+K_4})\mathrm{E}(M_{K_1}M_{K_2}\tilde{M}_{K_3}\tilde{M}_{K_4}) + gc(\mathrm{E}(\tilde{M}_{K_3+K_4}))^2]. \quad (3)$$

Finally, to calculate the probability that the genotypes at each locus match between the individuals we must sum over the probabilities associated with each of the possible assignments of observed alleles as paternal or maternal. This gives

$$\mathrm{P}(\text{observed configuration of matches}) = \sum_{i=0}^{k} \binom{k}{i} \mathscr{M}_i \tilde{M}_{k-i}^2. \quad (4)$$