Heredity **78** (1997) 583–589 *Received 3 July 1996*

Significant genetic correlations among Caucasians at forensic DNA loci

DAVID J. BALDING*† & RICHARD A. NICHOLS‡

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

Although the effect of population differentiation on the forensic use of DNA profiles has been the subject of controversy for some years now, the debate has largely failed to focus on the genetical questions directly relevant to the forensic context. We re-analyse two published data sets and find that they convey much the same message for forensic inference, in contrast with the dramatically differing conclusions of the original authors. The analysis is likelihood-based and combines information across loci and across populations without assuming constant genetic differentiation. Our results suggest that the relevant genetic correlation coefficients are too large to be ignored in forensic work: although DNA profile evidence is typically very strong, the effect of genetic correlations can be important in some cases. Such correlations can, however, be accommodated in an appropriate assessment of evidential strength so that population genetic issues should not present a barrier to the efficient and fair use of DNA profile evidence.

Keywords: DNA profile, F_{ST}, forensic science, Metropolis algorithm, population genetics.

Introduction

Human populations differ in their genetic composition, including in some cases populations which are geographically close. Such differentiation can be helpful in inferring plausible historical patterns of migration and interbreeding (see for example Barbujani *et al.*, 1994; Cavalli-Sforza *et al.*, 1994). In the use of DNA profiles for forensic identification, however, genetic differentiation is potentially problematic: although a particular DNA profile may be rare overall, it might be substantially more common in an ethnic group which contains both the true perpetrator of a crime and an innocent defendant. It follows that, if it is possible that the defendant is innocent but has an ethnic background similar to that of the actual culprit, then forensic assessments which ignore population differentiation may overstate evidential strength.

This effect is generally acknowledged in principle, but there has been disagreement over whether or not its magnitude is sufficient to warrant concern. Levels of genetic differentiation may well be smaller for the molecular genetic markers in forensic use than at traditional loci, because of their higher mutation rates and, possibly, less intense selection. Although there has been substantial theoretical discussion and some presentation of data, very little data analysis to date bears directly on the issues relevant to forensic identification, which differ somewhat from the usual interests of population genetics.

Two recent studies investigating genetic differentiation among Caucasians have drawn dramatically differing conclusions for the forensic debate. In a controversial study (Krane *et al.*, 1992; henceforth KASPH), DNA profile frequency estimates for Finnish and Italian individuals tended to be substantially smaller when obtained from a mixed Caucasian sample than when based on cognate Finnish or Italian samples. These authors concluded that '...we would not endorse the use of ethnically mixed racial databases (e.g. mixed Caucasians, ...)'. Their study has been criticized (Budowle *et al.*, 1994), although these criticisms have been rebutted (Sawyer *et al.*, 1996). In contrast, the authors of a different study also comparing a mixed Caucasian sample with samples of European origin (from Norway, Spain and Turkey) concluded that '...there should be little

*Current address and correspondence: Department of Applied Statistics, University of Reading, PO Box 240, Reading RG6 6FN, UK. E-mail: d.j.balding@reading.ac.uk.

chance of wrongful bias in forensic identity cases if...general population databases were employed' (Budowle & Monson, 1994; henceforth BM).

The two studies employed differing statistical methodologies and it is consequently difficult to make a direct comparison of their results and identify the reasons behind their conflicting conclusions. Correlation coefficients quantifying population differentiation were not estimated in either study. Correlations between the DNA profiles of distinct individuals are crucial to forensic identification (Balding & Donnelly, 1995a,b). This is because, after observing that a defendant has a DNA profile which matches a crime-scene profile, the question of central interest is whether or not other particular individuals might also have a matching profile. This question could in principle be answered in terms of the appropriate allele frequencies, but these are generally not known. Approximations of the appropriate frequencies by those obtained from forensic databases tends to be unfair to defendants (see the discussion of Fig. 2 below). This bias against defendants can, however, be compensated for using genetic correlation coefficients.

In this paper we estimate the genetic correlation coefficients from the KASPH and BM datasets, and thus provide a direct comparison of the two apparently contradictory studies. We do not consider other important, but nongenetical issues, such as the possibility that errors occur in the collection and evaluation of forensic samples, although we note that courts must consider such alternative explanations to assess fully DNA profile evidence. For a discussion of the role of laboratory errors in some particular cases, see Thompson (1995).

Estimating genetic correlations

Forensic applications

Once a particular allele has been observed in a locality then, because of shared ancestry, it becomes more likely that other individuals in that locality also have the allele. The strength of this effect is measured by a correlation coefficient: if p_A denotes the probability that the first gene sampled is allele A, then the probability that a second gene sampled in the locality is also allele A can be written as $p_A+(1-p_A)F$, where F denotes the correlation coefficient, similar to that known to population geneticists as Wright's F_{ST}.

Forensic calculations require the probability of matches involving four genes: two from the defendant and two from an alternative possible source of the crime sample. The correlations among all four genes must therefore be assessed, and these can be approximated in terms of F only (Nichols & Balding, 1991; Morton, 1992; Weir, 1994).

Established methods for estimating F_{ST} often reflect the traditional interests of population genetics rather than the requirements of forensic work. In particular, the value of F_{ST} usually measures differentiation among populations rather than the differentiation of populations away from the allele frequencies in a forensic database, which is the comparison required in forensic applications. Moreover, F_{ST} is often equated to the standardized variance over populations, which requires the populations to be comparable in terms of sizes, migration rates and evolutionary history. Such an assumption is inappropriate for the diverse human populations which are encountered in forensic work. These problems can be overcome in the flexible likelihood framework which we now describe.

Likelihood-based inference for F

For a wide range of structured populations, Balding & Nichols (1995) obtain a formula for the likelihood of a sample of genes from a particular locus and population. By viewing the sample as a sequence of genes drawn one by one, the formula can be expressed in a simple, recursive form as follows. If, after n genes have been drawn, n_A have been observed of allele A, then the probability that the next gene sampled is of allele A is

$$P_n(A)=\frac{n_AF+(1-F)p_A}{1+(n-1)F}. \qquad (1)$$

By successively applying eqn (1) to each gene in the sequence, and multiplying together the resulting expressions, a formula for the joint likelihood of the entire sample is obtained. This likelihood, technically a special case of the multinominal-Dirichlet likelihood, was also used by Rannula & Hartigan (1996), although these authors did not make use of the recursive formula (1).

Some intuitive insight into the likelihood formula may be obtained by considering a sample of n genes each of which reproduces itself at rate F, while migrant genes arrive at rate $1-F$. The migrants are allele k with probability p_k, and the migration and reproduction processes are mutually independent. Then the rate at which k-genes are generated is $n_kF+(1-F)p_k$, whereas the total rate at which genes are generated is $nF+(1-F)$. These two expressions are, respectively, the numerator and

denominator of (1). Note that when n_A and n are both large, (1) depends only weakly on F, reflecting the fact that little additional information about the value of F can then be obtained from further observations. Effort expended in the collection of very large sample sizes may, therefore, not be rewarded.

Combining information

Likelihood curves based directly on (1) are usually not sharply peaked and the resulting estimates of F are imprecise. One approach to improving the estimation is to combine information across loci (within populations) by multiplying together the likelihoods at distinct loci. This procedure would, however, be appropriate only if F were constant across loci and we will see (Fig. 1) that such an hypothesis is not supported by the data (see also Balding *et al.*, 1996). Possible reasons include differing mutation rates or selection processes at distinct loci. Similarly, it is not possible to combine directly the information from different populations, because of varying population sizes and demographic histories.

A method for obtaining more precise estimation without inappropriate assumptions of constancy was introduced by Balding *et al.* (1996). The value of F in the ith population at the jth locus is modelled by the formula

$$F_{ij} = \frac{1}{1+\alpha_i+\beta_j}, \qquad (2)$$

in which α_i and β_j are non-negative parameters which incorporate, respectively, a population and a locus effect. This formulation reduces the number of parameters to be estimated from $l \times m$ to $l+m$, where l and m are the numbers of populations and loci.

Model (2) reflects the underlying biological processes in that, for example, if the migration rates into subpopulation j are high then the value of β_j will be large and thus F_{ij} will be small for every locus i. Similarly, a high mutation rate at locus i can lead to a small value of F_{ij} for each population j. Eqn (2) holds exactly in the so-called island model of population subdivision and is approximately valid for a variety of population structures (Takahata, 1983; Slatkin & Barton, 1989).

To investigate the robustness of the likelihoods based on (1) and (2) for the data discussed here, we also examined the general model in which an additional parameter γ_{ij} was added to the denominator of (2). The postdata distributions of the γ_{ij} were all centred near the value corresponding to the simpler model, thus giving no indication that the model is inadequate.

Results

Estimates

The data consist of measured restriction fragment lengths at five VNTR (variable number tandem repeat) loci reported by KASPH and BM, except that some errors present in the KASPH data were subsequently corrected by those authors. For direct comparability, the data from KASPH were reclassified into the same bins as those used by BM. The posterior densities shown are affected by the arbitrary binning, but broad conclusions are unchanged. Samples were taken from a number of different European countries: Finland and Italy (KASPH) and Norway, Spain and Turkey (BM), in addition to mixed databases of US individuals of European ancestry. Because the databases are large, the Laplace estimate $(n_A+1)/(n+k)$, where k is the number of alleles (bins), can be used to estimate p_A (Balding & Nichols, 1994). Sample sizes are indicated on the plots (number of chromosomes scored). Combining data from the locus in common between the studies (D2S44) created a database of size 3326 at this locus. This required adjustment for the lengths of the flanking regions excised by the different restriction enzymes, after which the bin frequencies from the two databases were summed. For the other loci the database sizes were 634 (D16S85 and D10S28) and 2706 (D12S11 and D7S21).

Previous human genetic studies of F_{ST} give some guidance to the values of F appropriate here. Such studies (e.g. Cavalli-Sforza *et al.*, 1994) have rarely found values of F_{ST} in excess of 5 per cent among Caucasians, and in large populations they are usually less than 1 per cent. For VNTR loci, mutation rates in some cases appear to be high enough to obscure the differentiation between populations, but in other cases the differentiation is of the same magnitude as at traditional loci (see, for example, Buffery *et al.*, 1991). To encompass these observations, we chose independent, lognormal distributions with parameters 3.5 and 1.5 to model the pre-data uncertainty about the α_i and β_j (i.e. $\log(\alpha_i)$ and $\log(\beta_j)$ each initially have the $N(3.5, (1.5)^2)$ distribution). These values imply a prior probability density for F which has a mode at about 0.25 per cent and has density at least half the modal value between 0.05 per cent and 1.05 per cent. There are sufficient data that the results are insensitive to a wide range of alternative prior densities for the α_i and β_j. As an illustration,

setting $\sigma = 1$ and $\sigma = 2$ leads to substantial differences in the prior density for F: the prior modes are 0.6 per cent and 0.05 per cent, respectively (cf. 0.25 per cent with $\sigma = 1.5$). The posterior mode for Finland at locus D16S85, for example, varies only between 0.58 per cent and 0.62 per cent over this range of σ-values.

Figure 1 shows probability densities for F based on (1) and (2) for the KASPH and the BM data. The curves were obtained from 10000 iterations of a Metropolis simulation algorithm (Metropolis *et al.*, 1953) using the likelihoods specified by (1) and (2) and the lognormal(3.5, 1.5) prior densities. Metropolis algorithms generate samples from a probability distribution via a Markov chain with that stationary distribution. They are particularly valuable for high dimensional distributions for which conventional methods are infeasible. The methodology for implementing the algorithm is described by Smith & Roberts (1993). The first 1000 simulations were used to allow the Metropolis process to equilibrate, and were discarded.

Implications for forensic casework

A further simulation was used to illustrate the implications of genetic correlations for forensic match probabilities. The simulation modelled four loci each with $k = 15$ alleles. The 'global' frequencies (p_A), from which the forensic database is assumed to be drawn, were generated from the uniform distribution with $E(p_A) = 1/k$. The distribution of subpopulation gene frequencies in a range of simple genetic models follows a Dirichlet distribution with the variance specified by F and each expected frequency given by the global frequency p_A (see Balding & Nichols, 1995). 'Cognate' frequencies ($\tilde{p}_A$) were therefore drawn from the Dirichlet distribution with $E(\tilde{p}_A) = p_A$ and variance $p_A(1-p_A)F$. Finally, 100 four-locus DNA profiles were generated independently with the probabilities specified by the $\tilde{p}_A$.

Weight of evidence in forensic applications is measured by the likelihood ratio (see, for example, Evett, 1992; Brookfield, 1995), which in identification cases can usually be interpreted as a match probability. If the cognate frequencies ($\tilde{p}_A$) are known, the match probability can be obtained directly from them by multiplication. If they are not known, which is usually the case in practice, the match probability can be derived from eqn (1) in terms of the database frequencies (p_A) and F:

$$P(AB \mid AB) = 2\frac{(F+(1-F)p_A)(F+(1-F)p_B)}{(1+F)(1+2F)}. \quad (3)$$

Ignoring genetic correlations is equivalent to setting $F = 0$. For each of the simulated profiles, Fig. 2 compares the uncorrected match probability (using $F = 0$ in eqn 3) with the appropriate value (using $F = 1$ per cent; see discussion), both relative to the cognate value. The homozygote case is complicated by the possibility of null alleles, and is treated elsewhere (Balding & Nichols, 1994).

Discussion

Both the KASPH and BM data sets display broadly the same magnitude of genetic correlations, with the most likely values for F typically between 0.2 per cent and 1 per cent, but values in excess of 3 per cent are plausible at one locus (Fig. 1e), even though such large values are very unlikely *a priori* under our modelling assumptions. The report of the US National Research Council (1996) describes the value of 1 per cent as 'conservative' for the US population. This conclusion is based on unpublished analyses which seem both to have assumed constancy of F over subpopulations and to have ignored the role of forensic databases. Moreover, it is unclear what the report means by 'conservative' or how a single F value should be interpreted in the context of the many diverse ethnic groups which make up, say, the US Caucasian population. Our results suggest that further studies and appropriate analyses are required before firm conclusions can be drawn.

In view of our results, it is worthwhile asking how the original authors came to such differing conclusions. The KASPH analysis was based on plots of cognate versus database frequencies of the sample genotypes. This gave an indication of the differentiation between populations, but did not quantify it in a way that could be used directly in match probability calculations. The BM analysis also made comparisons between sample and database frequencies for genotypes. Their analysis did not, however, address the issue of genetic correlations directly because the genotypes were from a set of 'target profiles' of individuals classified as Caucasian, South Asian and Afro-Caribbean instead of appropriate targets from the cognate subpopulation. In their conclusion, BM quote estimates of F in the range -0.2 per cent to 0.2 per cent, based on other data and obtained using methods suffering from the problems outlined in the introduction. The importance of estimating correlations directly can be exemplified by the observation that our estimates, from their own data, imply F values outside this range and possibly over 10 times greater than their upper value.

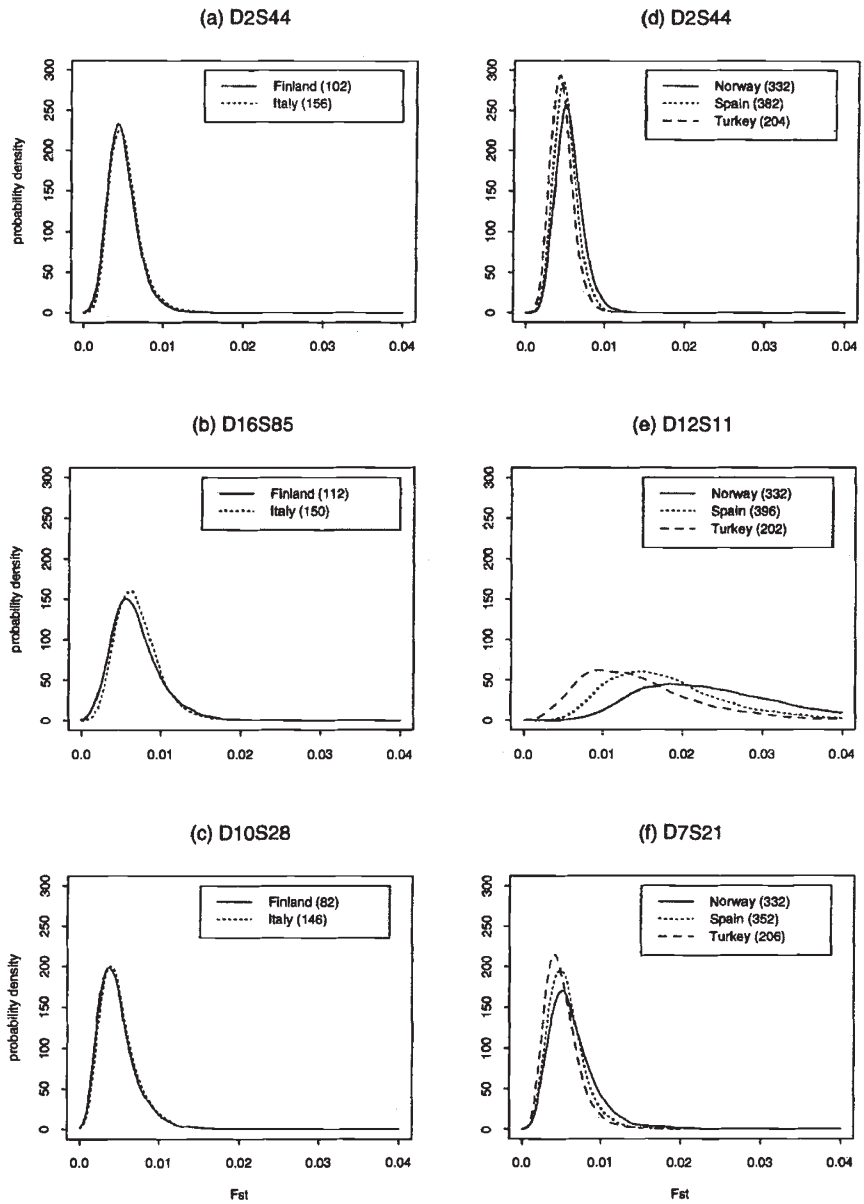

Fig. 1 Probability densities for the genetic correlation coefficient F at various loci based on: (a) to (c), a Finnish and an Italian sample compared with a mixed Caucasian database (the KASPH samples); (d) to (f) a Norwegian, a Spanish and a Turkish sample compared with a mixed Caucasian database (the BM samples).

Figure 1 displays only the marginal distribution of each F_{ij}, whereas forensic calculations require the joint distribution over loci for each subpopulation. Variation in the value of F across loci, in addition to positive correlations in these values, means that employing an average value of F will tend to understate forensic match probabilities.

If an F value in the order of 1 per cent were appropriate, what effect would this have on forensic calculations? Figure 2 illustrates the effect on forensic calculations both of ignoring genetic correlations and of allowing for them using eqn (3). The simulation mimics the situation which often arises in practice: profile frequency estimates are available from a database population, distinct from a cognate population which includes the defendant and some of the alternative possible sources of the crime stain. Use of database population frequencies in place of the correct (but generally unavailable) cognate frequencies tends to be unfair to defendants (70 points out of 100 have cognate values greater than database values, i.e. $x>0$, illustrating that use of database frequencies tends to overstate evidential strength). Appropriate allowance for genetic correlations eliminates this bias against defendants, even when the cognate frequencies are unavailable: only 38 points have $y>0$. Note that the x-values typically exceed the y-values by about 0.5, indicating that the

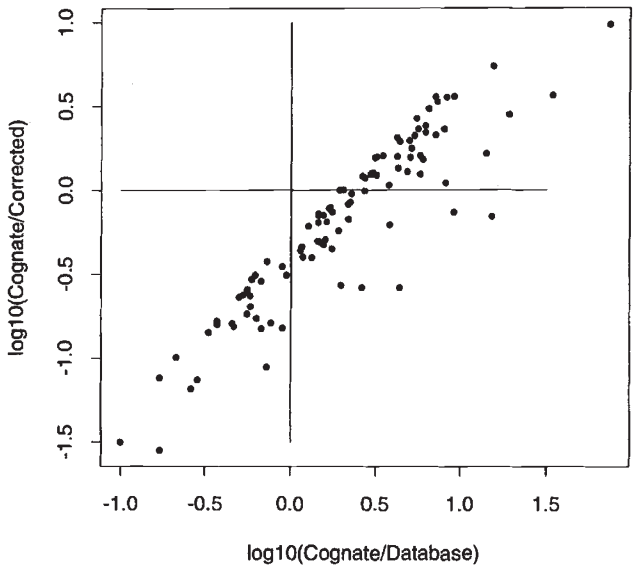

Fig. 2. A comparison of relative errors in match probabilities calculated ignoring genetic correlations (i.e. using eqn 3 with $F = 0$) and allowing for the appropriate level of genetic differentiation (here, $F = 1$ per cent). Each point on the scatter plot corresponds to a DNA profile, heterozygous at each of four loci, simulated from a subpopulation differentiated from a global population with $F = 1$ per cent. The x-coordinate is the ratio of the correct (cognate) profile frequency to its frequency in the database population. The y-coordinate is the ratio of the cognate frequency to the match probability obtained using eqn (3), with the appropriate value of 1 per cent for F.

relative error in using the database frequencies and ignoring a value of F of 1 per cent is about 1/2 an order of magnitude. When $F = 5$ per cent, the relative error is typically about two orders of magnitude. An apparent tendency to overestimation of the corrected match probabilities can be attributed to the logarithmic transform: it follows from $E[\text{Cognate/Corrected}] = 1$ that $E[\log_{10}(\text{Cognate/Corrected})] < 0$.

In practical casework, not only are cognate frequency estimates unavailable, but also it is usually impossible to specify the appropriate value of F. Following these and similar studies, however, the range of plausible values for actual human populations can be assessed. In many cases, DNA evidence is so powerful that even making generous allowance for possible genetic correlations still allows very strong statements of evidential value. In some marginal cases, involving for example partial profiles, little or no corroborating evidence, or close relatives of the defendant, consideration of plausible levels of genetic correlations may, appropriately, permit reasonable doubt about the source of the crime stain DNA (Balding & Donnelly, 1995a). Values of F suggested by the present analyses are not necessarily those appropriate in any particular case: other sources of uncertainty must be taken into account (Balding & Donnelly, 1995b) and the other evidence will have implications for the geographical scale on which genetic correlations should be assessed. To permit such assessments, further studies are needed to investigate genetic correlations on a range of demographic scales, including for example isolated rural communities and close-knit migrant and religious groups.

Acknowledgements

We thank R. Allen, S. Sawyer and D. Hartl for providing the data for Fig. 1(a–c) and Peter Donnelly for helpful comments on an early draft of the manuscript. D.J.B. was supported in part by the Science Research Fellowship Scheme of the Nuffield Foundation and R.A.N. by SERC grant GRG11101.

References

BALDING, D. J. AND DONNELLY, P. 1995a. Inference in forensic identification. *J. R. Statist. Soc. A*, **158**, 21–53.

BALDING, D. J. AND DONNELLY, P. 1995b. Inferring identity from DNA profile evidence. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 11741–11743.

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.

BALDING, D. J. AND NICHOLS, R. A. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

BALDING, D. J., GREENHALGH, M. AND NICHOLS, R. A. 1996. Population genetics of STR loci in Caucasians. *Int. J. Leg. Med.*, **108**, 300–305.

BARBUJANI, G., NASIDZE, I. S. AND WHITEHEAD, G. N. 1994. Genetic diversity in the Caucasus. *Hum. Biol.*, **66**, 639–668.

BROOKFIELD, J. F. Y. 1995. The effect of relatedness on likelihood ratios and the use of conservative estimates. *Genetica*, **96**, 13–19.

BUDOWLE, B. AND MONSON, K. L. 1994. Greater differences in forensic DNA profile frequencies estimated from racial groups than from ethnic subgroups. *Clin. Chim. Acta*, **228**, 3–18.

BUDOWLE, B., MONSON, K. L. AND GIUSTI, A. M. 1994. A reassessment of frequency estimates of PVUII-generated VNTR profiles in a Finnish, an Italian, and a

general US Caucasian database — no evidence of ethnic subgroups affecting forensic estimates. *Am. J. Hum. Genet.*, **55**, 533–539.

BUFFERY, C., BURRIDGE, F., GREENHALGH, M., JONES, S. AND WILLOT, G. 1991. Allele frequency distributions of four variable number tandem repeat (VNTR) loci in the London area. *Forensic Sci. Int.*, **52**, 53–64.

CAVALLI-SFORZA, L. L., MENOZZI, P. AND PIAZZA, A. 1994. *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

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

KRANE, D. E., ALLEN, R. W., SAWYER, S. A., PETROV, D. A. AND HARTL, D. L. 1992. Genetic differences at four DNA typing loci in Finnish, Italian, and mixed Caucasian populations. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 10583–10587.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. AND TELLER, E. 1953. Equation of state calculations by fast computing machines. *Chem. Phys.*, **21**, 1087–1092.

MORTON, N. E. 1992. The genetic structure of forensic populations. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 2556–2560.

NATIONAL RESEARCH COUNCIL 1996. *The Evaluation of Forensic DNA Evidence*. Natl. Acad. Press, Washington DC.

NICHOLS, R. A. AND BALDING, D. J. 1991. Effects of population structure on DNA fingerprint analysis in forensic science. *Heredity*, **66**, 297–302.

RANNALA, B. AND HARTIGAN, J. A. 1996. Estimating gene flow in island populations. *Genet. Res.*, **67**, 147–158.

SAWYER, S. A., PODLESKI, A., KRANE, D. E. AND HARTL, D. L. 1996. DNA fingerprinting loci do show population differences. *Am. J. Hum. Genet.*, **59**, 272–274.

SLATKIN, M. AND BARTON, N. H. 1989. A comparison of three indirect methods for estimating average levels of gene flow. *Evolution*, **43**, 1349–1368.

SMITH, A. F. M. AND ROBERTS, G. O. 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc. B*, **55**, 3–24.

TAKAHATA, N. 1983. Gene identity and genetic differentiation of popluations in the finite island model. *Genetics*, **104**, 497–512.

THOMPSON, W. C. 1995. Subjective interpretation, laboratory error and the value of forensic DNA evidence: three case-studies. *Genetica*, **96**, 153–168.

WEIR, B. S. 1994. The effect of inbreeding on forensic calculations. *Ann. Rev. Genet.*, **28**, 597–621.