Int J Legal Med (1996) 108:300–305

ORIGINAL ARTICLE

David J. Balding · Matthew Greenhalgh
Richard A. Nichols

Population genetics of STR loci in Caucasians

Received: 17 October 1995 / Received in revised form: 3 January 1996

Abstract STR loci are becoming increasingly important in forensic casework. In order to be used fairly and efficiently, the population genetics of these loci must be investigated and the implications for forensic inference assessed. A key population genetics parameter is the "coancestry coefficient", or F_{ST}, which is the correlation between two genes sampled from distinct individuals within a subpopulation. We present analyses of STR data, at geographic scales which range from national to regional, from the UK and other European sources. We implement a likelihood-based method of estimating F_{ST}, which has important advantages over alternative methods: it allows a range of plausible values to be assessed, rather than presenting a single point estimate, and it allows a subpopulation to be compared with a larger population from which a database has been drawn, which is the relevant comparison in forensic work. Our results suggest that values of F_{ST} appropriate to forensic applications in Europe are too large to be ignored. With appropriate allowance, however, it is possible to make use of STR evidence in a way which is efficient yet avoids overstatement of evidential strength.

Key words Short tandem repeat (STR) · DNA profiles · Statistics · Identification · Paternity

D. J. Balding
School of Mathematical Sciences,
Queen Mary and Westfield College, Mile End Road,
London E1 4NS, UK

M. Greenhalgh
Metropolitan Police Forensic Science Laboratory,
109 Lambeth Road, London SE1 7LP, UK

R. A. Nichols (✉)
School of Biological Sciences,
Queen Mary and Westfield College, Mile End Road,
London E1 4NS, UK

Introduction

Short tandem repeat (STR) loci are becoming widely used in forensic identification and paternity testing (Kimpton et al. 1994). Their advantages over alternative DNA typing systems include the existence of distinct allelic classes and small length measurement error. They can be used effectively in conjunction with PCR to type very small amounts of DNA.

In order to make appropriate use of STR data in forensic work, assessments of the relevant genetic correlations are required. Genetic correlations arise because a defendant may share ancestry with other possible culprits. Consequently, the probability that a particular alternative culprit shares the defendant's profile will usually be higher than estimates of the profile frequency obtained directly from forensic databases using the so-called "product rule" (National Research Council 1992). The case that one or more alternative culprits are close relatives of the defendant, and hence may both inherit alleles from known parents or grandparents, can be directly addressed (Balding & Nichols 1994). Here, we are concerned with the shared ancestry, or "coancestry", which is not attributable to known common ancestors.

Because levels of coancestry are typically small, genetic correlations are often neglected in forensic work. This assumption is wrong in fact, because it ignores established population genetics knowledge, and wrong in principle, because it has the effect of exaggerating the strength of the evidence against the defendant. We will show that the exaggeration is not trivial for the size of correlations found within European populations. Moreover, the assumption of no genetic correlation between defendant and other possible culprits is also unnecessary. Levels of coancestry can be assessed and their effects allowed for using established population genetics theory and data (Balding & Nichols 1994). Resulting assessments of evidential strength usually remain strong enough to allow effective prosecutions without unfairness to defendants.

In this paper we present analyses of STR data from a range of European populations listed in Table 1. For each

Table 1 Geographic origins and sizes of the samples analysed

Sample	Category	Size (Individuals)			
		VWA	THO1	F13	FES
UK Caucasians	Database	679	680	680	679
Derbyshire (England)		582	582	582	582
Dundee (Scotland)	UK	239	257	238	167
Northern Ireland	Regions	114	114	114	113
Strathclyde (Scotland)		139	139	139	139
Greece	Other	341	341	341	341
Greek Cypriot	European	26	28	25	25
Italy		100	101	95	93

population we estimate the value of F_{ST} which measures the genetic correlations with respect to a forensic database. Current alternative methods of estimation of F_{ST} reflect the traditional interests of population geneticists and are less suited to forensic applications because they do not measure correlations relative to a forensic database. Every forensic database will have a characteristic genetic composition because of the populations from which it is drawn and hence would generate a specific set of F_{ST} estimates for the samples. We use a combined database of two UK forensic science laboratories for our estimation. Our purpose is to evaluate the range and pattern of F_{ST} values found as an indication of the importance of coancestry in forensic inference.

Further problems arise where estimation methods assume a constant value of F_{ST} over populations and/or over loci. Assuming constancy of F_{ST} over populations is clearly inappropriate as there are substantial discrepancies between human populations in size and in patterns of mating and migration. The resulting differences in F_{ST} values may be important in forensic work because the value(s) appropriate to a particular case can vary according to the relevant population(s). Variations in the value of F_{ST} over loci may arise as a result of differences in selection and/or mutation rates (Slatkin 1985). Note however that the large stochastic variability of genealogies at distinct loci (Donnelly 1996) can produce estimates which vary substantially over loci, even though the mutation rate and selection effects are constant.

The procedure for calculating likelihoods for F_{ST} is justified more fully elsewhere (Balding & Nichols 1995). Here we demonstrate a further development for combining information from different loci and different populations that overcomes the difficulties described above. It allows, for example, a comparison of estimates with, and without the assumption of constancy over loci. The method has the further advantage of providing likelihoods and posterior probability distributions which give the range of plausible values of F_{ST}, rather than just a point estimate (with, possibly, a standard error). This feature is important as point estimates can be misleading in the presence of the highly skew distributions found in the analysis of F_{ST}. The approach is particularly valuable in the interpretation of results from those smaller ethnic groups which may have particularly high levels of coancestry. These cases are often represented by small samples, so the precision of the F_{ST} estimate needs to be established.

Genetic correlations in forensic casework

In forensic casework, allele frequency estimates are typically available from databases which cover a large, heterogeneous population, such as the mixed Caucasian databases of the London Metropolitan Police Forensic Science Laboratory (MPFSL) and the UK Forensic Science Service (FSS). When a suspect is found to have an STR profile which matches a crime profile, the strength of the evidence depends on the (conditional) probabilities that other individuals also have that profile. The role of genetic correlations in evaluating these probabilities can be illustrated by considering the case that the crime stain provides a one-locus, heterozygous, profile and a defendant has a matching profile at this locus. Consider the probability, which we write $P(AB \mid AB)$, that a particular individual, apparently unrelated to the defendant, also matches the crime profile. This probability can be calculated in terms of the frequency of the alleles A and B (denoted p_A and p_B respectively) using the formula

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

where F is a correlation which may be due to a number of factors (Balding & Donnelly 1995) but is often primarily due to coancestry between the defendant and the other possible culprit under consideration (Balding & Nichols 1994). Ignoring correlations, that is setting F to 0, the match probability simplifies to the profile frequency, which in the one-locus heterozygote case is given by

$$P(AB) = 2p_A p_B, \quad (2)$$

which is smaller, and hence overstates the strength of the evidence, in all cases of practical interest (e.g. whenever $p_A + p_B < 2/3$). The importance of this overstatement depends on the magnitude of F, which must therefore be estimated.

It will not usually be possible to specify the exact value of F appropriate to a particular case. Instead, we propose studies of a range of populations at differing demographic scales. In this manner, the range of plausible values of F can be assessed and values appropriate to the circumstances of a particular case can be selected. The analyses of the present paper may be viewed as an initial step in this programme. In practice, it may be found convenient to routinely employ, in all but the most exceptional cases, a common value of F towards the upper end of the plausible range. To allow this, it is particularly important to investigate the extreme cases of differentiation in actual human populations. These are likely to be small and/or isolated populations. Studies of such populations should have a high priority.

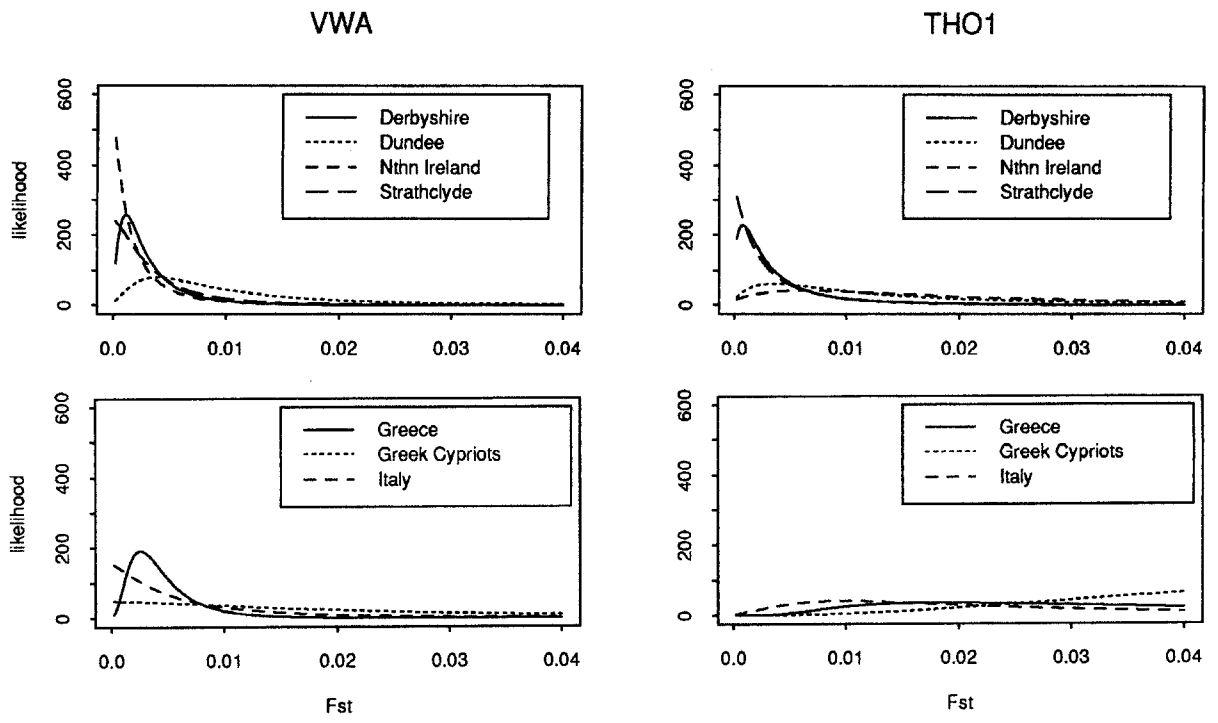

Fig. 1 Likelihood curves for F_{ST}, calculated from (3) using data from the populations listed in Table 1 at loci VWA and THO1. The y-axes have been scaled so that each curve can be interpreted as a probability density (i.e. each curve covers an area of one)

Materials and methods

Data

The STR genotypes at four loci (HUMTHO1, HUMVWA31, HUMF13A1 and HUMFES) had previously been determined for each individual using an automated fluorescence method (Kimpton et al. 1994). The samples consisted of apparently unrelated individuals (in cases involving individuals known to be related then only one from each family was included). Casework samples usually comprise the crime victim and one or more suspects. As these are not planned samples, they may include individuals who do not have local ancestry or who are otherwise atypical. Because of the manner of sampling, the geographic origin of the samples can be specified to the level of the designated region, but not more precisely.

Caucasian database: 423 individuals from FSS casework combined with 257 from MPFSL staff members and casework.

Derbyshire: Police and civil staff from the Derbyshire constabulary.

Dundee: Casework samples drawn from the residents of Tayside, Fife and the central region of Scotland served by the Tayside Police Forensic Science Laboratory.

Northern Ireland: Staff of the Northern Ireland Forensic Science Laboratory.

Strathclyde: Casework samples drawn from the Strathclyde region.

Greece: Greek donors from Athens, excluding individuals with Northern Greek ancestry.

Greek Cypriot: Greek Cypriot individuals attending a thalassaemia clinic in the London area.

Italy: Donors drawn from the residents of the province of Parama in Northern Italy.

Estimation methods

We employ a likelihood-based method of estimating F_{ST} (Balding & Nichols 1995). The joint likelihood of a sample consisting of N genes, of which n_i genes are of type i, $i = 1,..., k$, is obtained from the recurrence relationship

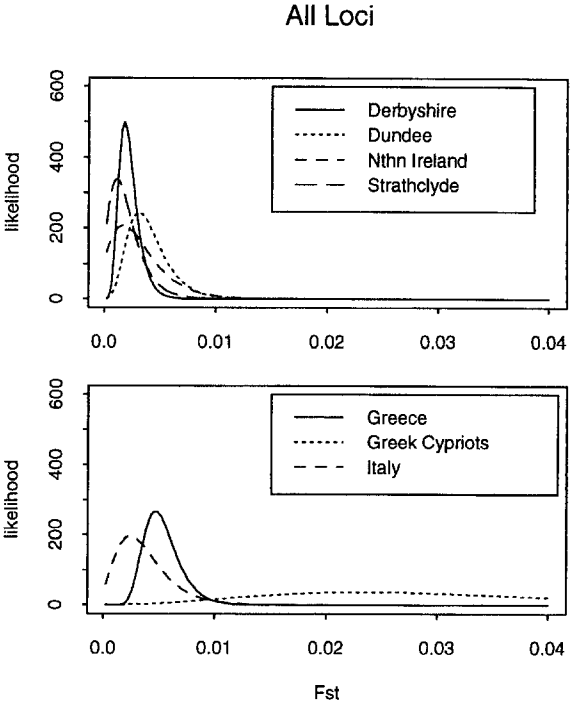

Fig. 2 Likelihood curves for F_{ST} from the populations listed in Table 1 based on assuming that F_{ST} is constant across loci. Each curve is obtained as the product of values of (3) across the four loci. The y-axes have been scaled so that each curve can be interpreted as a probability density

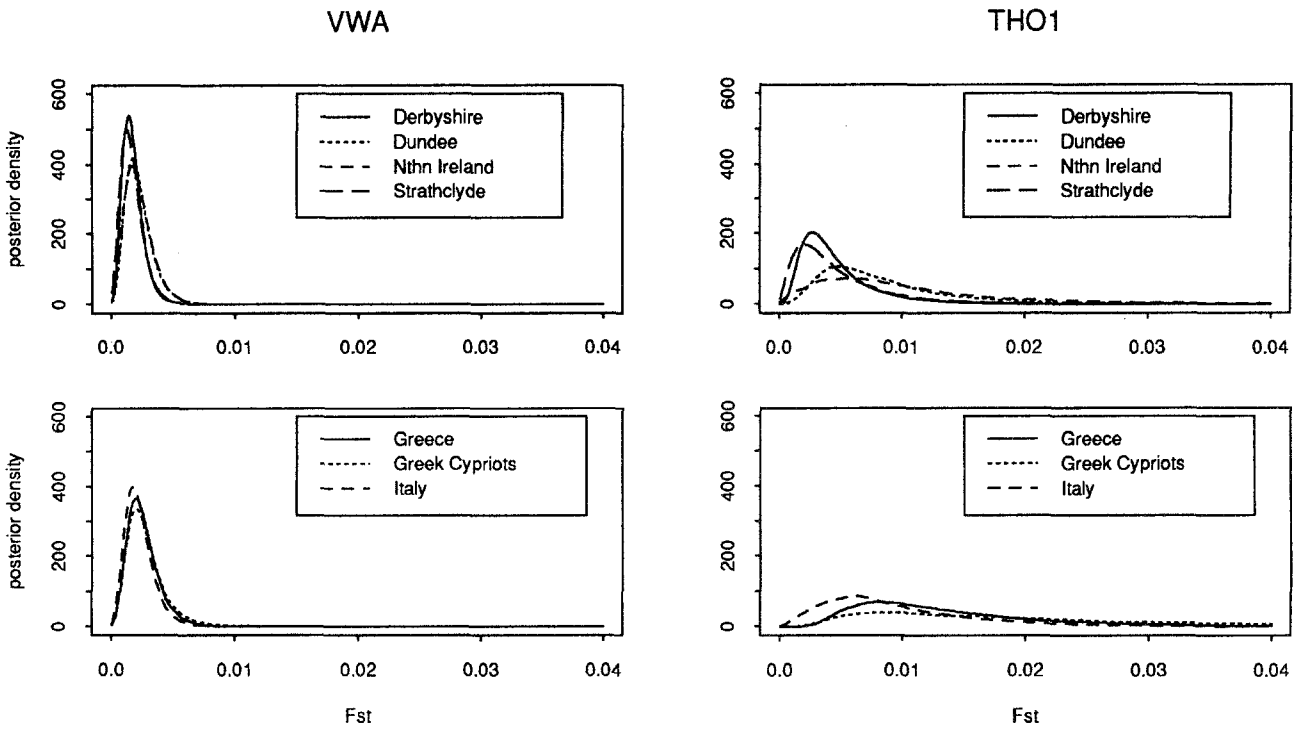

Fig. 3 Probability density for F_{ST} from the populations listed in Table 1 at loci VWA and THO1. The densities are obtained using (3) and (4) together with an assumption of independent, lognormal (3·5,2·0) distributions for each parameter in (4). Gaussian kernel density estimation was used, based on 10,000 iterations of a Metropolis algorithm, of which the first 1,000 were discarded

$$P_N(n_1,n_2,...,n_i,...,n_k) = P_{N-1}(n_1,n_2,...,n_i-1,...,n_k)\times \left(\frac{(n_i-1)F+p_i(1-F)}{1+(N-2)F}\right) \quad (3)$$

for $n_i > 0$. Successively implementing this equation with arbitrary selection of i at each step, together with the initial condition $P_0(0,0,...,0) = 1$, leads to an explicit expression for the joint likelihood as a product of $n_1 \times n_2 \times ... \times n_k$ terms. The p_i were estimated from the mixed Caucasian database using the formula $p_i = (n_i + 1)/(N + k)$ where n_i and N refer to numbers in the database.

The likelihood (3) can be derived in the context of a model involving randomly-mating subpopulations partly isolated from the larger population from which the database has been collected. It can also be shown to apply under more general assumptions (Balding & Nichols 1995). It aims to capture the essential features of the relevant genetic correlations while remaining simple enough to be tractable.

Figure 1 shows likelihood curves obtained from (3) for the loci VWA and THO1, which display respectively the smallest and the largest values of F_{ST} among the four loci. Correlations are generally difficult parameters to estimate, which is reflected by the fact that the curves in Fig. 1 are generally not tightly peaked: there is usually not enough information at a single locus and a single population for precise estimation. Note that the problem is not primarily due to sample size: simulation studies suggest that larger samples would give little further precision, except in the case of the Greek Cypriot sample.

Sharper estimates can, however, be obtained by combining information across loci. Figure 2 shows the joint likelihood of F_{ST}, assumed constant over the four loci. The curves are more tightly peaked than in Fig. 1, but at the cost of a possibly inappropriate assumption.

In an attempt to improve upon Fig. 1 while still allowing variation in F_{ST} across loci, we also modelled F_{ST} at the ith population and the jth locus by the relationship

$$F_{ij}\frac{1}{1+a_i+b_j}, \quad (4)$$

where the a_i and b_j are parameters which reflect, respectively, a locus and a population effect. Equation (4) can be derived under the "Island" model of population structure (Takahata 1983). Direct evaluation of the relevant likelihoods is not possible under (4), and the probability density curves in Fig. 3 were obtained from a method of stochastic simulation known as a Metropolis algorithm (Metropolis et al. 1953; Smith & Roberts 1993), assuming independent lognormal(3.5,2.0) prior distributions for the a_i and b_j. Other pre-data modelling assumptions for the a_i and b_j were investigated and the curves in Fig. 3 were found to be insensitive to a wide range of plausible choices.

Results

Although there is evidence for variation in F_{ST} values across loci, Fig. 2 may be of interest in providing a guide to plausible F_{ST} values for the range of geographic scales investigated. In UK populations, most likely values are around 0.25% and values up to about 1% are supported by the data. Perhaps surprisingly, the data do not suggest a larger value of F_{ST} for Italy than for Dundee. This may be explained in terms of Dundee's history of relative geographic isolation from the bulk of the UK population. Values are somewhat larger for Greece, while remaining under 1%. Cyprus is of interest because of the smaller population size and the isolation imposed by its island status. Although the sample size is small, the estimate of F_{ST} has effectively been resolved to the range between 0.9% and 4.8%. This gives some indication of the range of values appropriate for migrant groups in the UK originating from other small and/or isolated populations.

Looking at the estimates on an individual locus basis and assuming (4), Fig. 3 suggests that the values of F_{ST} for THO1 are substantially larger than for VWA. The curves for F13 and FES (not shown) are intermediate between those for VWA and THO1. In summary, values of F_{ST} between 0.1% and 0.5% are well supported at each locus. Further, values up to 1% and up to 2% are also supported at, respectively, FES and THO1 (up to 2% and 4% for Greek Cypriots).

Discussion

Our analyses suggest values for F_{ST} of up to 1% for the populations investigated, other than the Greek Cypriots. For typical four-locus profiles, equation (1) with $F = 1\%$ leads to a match probability around 2 to 5 times greater than that obtained using equation (2) (Balding & Nichols 1995). The larger values indicated for the (London) Greek Cypriots may be due to factors which are not unique to this group. This community has a distinct cultural identity and a geographically restricted ancestry. There may well be comparable populations within the UK and elsewhere.

The application of DNA typing to forensic identification and paternity testing is relatively recent and studies of the loci used in forensic work are limited. Such studies mainly consist of analyses of heterozygosity within forensic databases, often drawn from large, imprecisely defined geographic areas, and differentiation between these databases (Hammond et al. 1994; Budowle 1995). The estimates of F_{ST} (and related parameters) obtained from these studies are typically smaller than the values we present here, which can be explained in terms of the design of the different studies.

Excess homozygosity within a database may indicate population substructuring, and can be used to estimate a genetic correlation (Morton 1993). The true extent of homozygosity can be difficult to assess at loci used in forensic casework because of technical problems including the possibility of null alleles. More fundamental is a problem with the pattern of human population substructuring. The most genetically distinct populations tend to be small and isolated. The correlations calculated from homozygosity within a database will be an average which combines such populations with less differentiated (usually larger) populations. Even if they are proportionately represented in a database, this averaging is inappropriate. Separate estimates are required from a variety of distinct populations, because it is the distribution of F_{ST} values that is important in genetic inference, not an average (Nichols 1995).

Gill & Evett (1995) have calculated point estimates of F_{ST} at STR loci by comparison of a number of samples, including large databases, with each other. If the databases are drawn from a number of genetically distinct populations, then some of the differentiation between populations will be misinterpreted as differentiation between individuals. In such circumstances the estimates of F_{IT} may be closer to what is required for forensic calculations (Nichols & Balding 1991). It is notable that the F_{IT} estimates of these authors are of the same order as the F_{ST} values calculated here. Using a randomisation test, they could not find significant evidence of geographic differentiation. In contrast, we found little support for $F_{ST} = 0$ in several populations (Fig. 3). This again highlights the advantages of making separate estimates for each population. Combining information across loci (Fig. 2) allows useful estimates from surprisingly small sample sizes. In particular the sample of only 25 Cypriots was sufficient to demonstrate convincingly that the combined F_{ST} for this population was in excess of 0.9%. Conversely, it is in the nature of genetic data that even large samples from a single locus provide imprecise estimates of F_{ST}, as demonstrated by the broad curves for the Derbyshire sample in Fig. 1. This limitation arises because the gene frequency estimates are subject to two sources of variability: sampling error and genetic variability. The genetic variability includes the random action of migration and genetic drift. Increased sample sizes only reduces the sampling error in the F_{ST} estimates. Combining information across loci also reduces the second source of error because the genetic processes have independent effects on each locus. This principle explains why the combined estimates in Fig. 2 are so much more precise (the curves are more peaked) than the individual estimates in Fig. 1.

Another distinct feature of our method is that the genetic correlations are calculated based on differentiation of populations from a specific database (rather than from each other). These are the correlations which are directly relevant to forensic calculations, because they can be used to calculate the probability that an individual with an ethnic background similar to that of the defendant has a matching DNA profile. Such probabilities are relevant even when there is no particular evidence to suggest that the culprit came from the population of the defendant (Balding & Donnelly 1995).

The magnitude of genetic differentiation between human populations has been extensively studied by population geneticists, because it provides clues about human demographic history (Ammerman & Cavalli-Sforza 1984; Cavalli-Sforza et al. 1995). For the loci traditionally studied by population geneticists, mutation rates are very low and consequently, when selection is weak, the expected genetic differentiation is the same at each locus. The loci in use in forensic science are chosen partly because they are highly polymorphic, and this often seems associated with, amongst other differences, higher mutation rates (Weber & Wong 1993). Furthermore, STR loci are typically found in introns of genes that may possibly be subject to geographically varying selection. High mutation rates and some forms of selection can produce differences in F_{ST} across loci (Slatkin 1985; Slatkin & Barton 1989). It follows that, whereas previous work may provide a guide to the magnitude of genetic differentiation, direct studies of forensic loci are required to underpin expert testimony. Despite these differences, existing population genetics theory and data can cast useful light on forensic analyses. The major geographic patterns observed at traditional loci are most readily interpreted as a consequence of the his-

tory of population movement and expansion rather than current patterns of dispersal (Cavali-Sforza et al. 1995; Barbujani & Sokal 1991). The general patterns can guide the design and interpretation of a study. In particular, the previous studies highlight the influence of geographic scale on the magnitude of genetic variation.

In comparisons between pairs of populations, F_{ST} tends to increase with the geographic distance between populations (Cavali-Sforza et al. 1995). The general trend differs between continents in that median values approach 2% at a distance of 3000 km in Africa and Asia whereas in Europe F_{ST} is little over half that. There is considerable variation around these trends. In some areas, populations with separate and ancient histories are in close geographical proximity and retain markedly distinct genetic compositions (Barbujani 1991). The values of F_{ST} found in this study are thus in the lower range of those found at traditional loci, which may be a consequence of higher mutation rates (Takahata 1983).

In this study we have made use of samples collected for a variety of purposes. The advantages of having more precise information about the geographic origin of samples is illustrated by Cavalli-Sforza & Feldman's (1990) survey of villages in the Parama Valley. Their estimates of F_{ST} varied between 0.3% and 2.6% for the same data depending on whether single villages are distinguished, or are clustered into large groups. It would therefore be valuable to have additional surveys at finer demographic scales. Samples from regions which have previously been demonstrated to be genetically distinct at traditional loci would be revealing. Within the UK, towns within mid-Wales, East Anglia, West Cornwall and West Scotland are likely candidates.

Acknowledgements The authors are grateful to all individuals who have contributed data or samples for analysis. These include P. Gill, J. Dunlop, J. Peden, C. Konialis, N. Cucurachi, S. Kilpatrick, H. Butler, K. Way, H. Mygill, R. Piercy, B. Mckeown and J. Andersen. We thank P. Donnelly and L. Foreman for helpful comments on a draft of the manuscript.

References

Ammerman AJ, Cavalli-Sforza LL (1984) The neolithic transition and the genetics of populations in Europe. Princetown University Press, Princetown NJ

Balding DJ, Donnelly P (1995) Inference in forensic identification. J R Stat Soc A 158: 21–53

Balding DJ, Nichols 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

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96: 3–12

Barbujani G (1991) What do languages tell us about human microevolution? Trends Ecol Evol 6: 151–156

Barbujani G, Sokal RR (1991) Genetic population structure of Italy II: physical and cultural barriers to gene flow. Am J Hum Genet 48: 398–411

Budowle B (1995) The effects of inbreeding on DNA profile frequency estimates using PCR-based loci. Genetica 96: 21–25

Cavalli-Sforza LL, Feldman M (1990) Spatial subdivisions of populations and estimates of genetic variation. Theor Popul Biol 37: 3–25

Cavalli-Sforza LL, Menozzi P, Piazza A (1995). The history and geography of human genes. Princeton University Press, Princeton, NJ

Donnelly P (1996) Interpreting genetic variability: the effects of shared evolutionary history. In: Weiss K (ed) Variation in the human genome. Wiley, New York (in press)

Gill P, Evett I (1995) Population genetics of short tandem repeat (STR) loci. Genetica 96: 69–87

Hammond HA, Jin L, Zhong Y, Caskey CT, Chakraborty R (1994) Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am J Hum Genet 55: 175–189

Kimpton CP, Fisher D, Watson S, Adam M, Urquhart A, Lygo JE, Gill P (1994) Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. Int J Legal Med 106: 302–311

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. Chem Phys 21: 1087

Morton NE (1993) Kinship bioassay on hypervariable loci in blacks and caucasians. Proc Natl Acad Sci USA 90:1892–1896

National Research Council (1992) DNA technology in forensic science: theory, techniques and applications. National Academy Press, Washington DC

Nichols RA (1995) How large are the relevant genetic correlations. Comment on "Inference in forensic identification''. D.J. Balding & P. Donnelly. J R Stat Soc A 158: 2153

Nichols RA, Balding DJ (1991) Effects of population structure on DNA fingerprint analysis in forensic science. Heredity 66: 297–302

Slatkin M (1985) Gene flow in natural populations. Ann Rev Ecol Syst 16: 393–430

Slatkin M, Barton NH (1989) A comparison of 3 indirect methods for estimating average levels of gene flow. Evolution 43: 1349–1368

Smith AFM, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J R Stat Soc B 55: 3–24

Takahata N (1983) Gene identity and genetic differentiation of populations in the finite island model. Genetics 104: 497–512

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2: 1123–1128