

# Effects of population structure on DNA fingerprint analysis in forensic science

RICHARD A. NICHOLS & DAVID J. BALDING\*

*School of Biological Sciences and \*School of Mathematical Sciences, Queen Mary and Westfield College, University of London, Mile End Road, London E1 4NS, UK*

DNA fingerprints are used in forensic science to identify individuals. However, current analyses could underestimate the probability of two individuals sharing the same profile because the effect of population structure is not incorporated. An alternative analysis is proposed to take into account population stratification. The analysis uses studies of inbreeding in human populations to obtain an empirical upper bound on the magnitude of the effect.

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

## Introduction

### *Inference from DNA fingerprint data*

Human DNA fingerprinting has become an important tool in forensic science, but doubts have been expressed about the way in which the results are analysed. Here we address the problem of population stratification. For example, Lander (1989) mentions the case of Texas versus Hicks, in which a murder occurred in an isolated town founded by a handful of families. A suspect taken at random from the town is more likely to share the DNA fingerprint of the murderer than an individual taken at random from the whole country. More generally, human populations are stratified so that related individuals tend to associate together. Cohen (1990) has suggested that this stratification may result in a serious underestimation of the probability of two DNA fingerprints matching. We propose a modification of the usual analysis of DNA fingerprint data which takes account of the effect of population structure. We argue that it is possible to use existing knowledge to put an upper bound on the magnitude of the effect.

Forensic DNA fingerprints (or profiles) are usually scored using single-locus probes. The probes are chosen to bind to loci which have a very large number of alleles with different lengths. The differences in length are due to variation in the number of a tandemly repeated sequence (Jeffreys *et al.*, 1985a,b). The essence of the procedure is that DNA is extracted from human tissue in a forensic sample and is digested with a restriction enzyme. The fragments are separated by

electrophoresis and the alleles are located by probing a Southern blot. Allele lengths are measured using triangulation to marker bands in adjacent tracks. Because of the large number of alleles segregating in the population, most individuals are heterozygous and generate two bands. Single bands sometimes occur; they may be due to homozygosity, to distinct alleles whose lengths are too similar to be distinguished, or to an allele too small to be measured. Typically, the procedure is repeated at a total of three to six loci generating 6–12 bands.

We consider the problem of determining whether or not DNA from a known suspect is contained in a forensic sample, which we refer to as the criminal sample. The suspect may volunteer a specimen which is fingerprinted in the same way as the criminal sample. Rather than directly drawing inferences from all the measured allele lengths, the analysis is usually simplified by recording a binary 'match'/'no match' outcome given either by an expert's visual inspection of the profile or by a match rule based on the apparent length difference. If, at any of the loci, a suspect band is judged to differ from both criminal bands by a discrepancy not attributable to measurement error then 'no match' is declared and the suspect is exonerated; otherwise 'match' is declared. We wish to quantify the credence which a declared match gives to the hypothesis  $G$  of guilt, i.e. the criminal sample does contain the suspect's DNA, rather than the complementary hypothesis  $I$  of innocence. A convenient quantity is given by the ratio  $R$  of the likelihoods of match under the two hypotheses given  $c$ , the observed criminal

genotype. That is, we calculate

$$R = \frac{\text{Pr}[\text{match}|I,c]}{\text{Pr}[\text{match}|G,c]} \quad (1)$$

In order to define the numerator of (1) uniquely, hypothesis  $I$  is usually interpreted as the assumption that the suspect has been chosen (uniformly) randomly from some population  $Q$ .

Small values of  $R$  provide evidence against hypothesis  $I$  and in favour of hypothesis  $G$ . If, perhaps on the basis of other evidence, it is reasonable to assign prior probabilities  $\pi(G)$  and  $\pi(I)$  to the two hypotheses, with  $\pi(G) + \pi(I) = 1$ , then the posterior probability of guilt is given by Bayes' rule

$$\text{Pr}[G|\text{match},c] = \frac{\pi(G)}{R\pi(I) + \pi(G)} \quad (2)$$

More commonly, explicit assignment of prior probabilities is avoided and the value of  $R$  is subjectively assessed in conjunction with any other evidence.

To calculate  $R$  we assume that the loci are unlinked so that we need only consider a single locus and multiply probabilities across loci. We will write  $R_1$  for  $R$  in the single-locus case. The value of  $R_1$  depends on assumptions about the accuracy of the method of match detection. For true allele lengths  $x$  and  $y$ , let  $M(x,y)$  denote the probability that a match is declared. If we suppose that a match is detected whenever allele lengths are the same, i.e.  $M(x,x) = 1$  for all  $x$ , then the denominator of  $R_1$  is one. The most common assumption is that  $M$  is the step function

$$M(x,y) = \begin{cases} 1 & \text{if } |x-y| < \delta \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where the cut-off value  $\delta$  is typically  $k(x+y)$  for some constant  $k$ . Although this choice of  $M$  does not precisely model the accuracy of real match rules, it may provide a reasonable approximation. Further discussion of allele length measurement error and its effects on estimates of allele frequencies is given by Gill *et al.* (1990). An important feature of assumption (3) is that it facilitates the analysis because it implies that  $R_1$  is simply the relative frequency in the population  $Q$  of all allele pairs whose lengths both match (i.e. are within  $\delta$  of) criminal alleles. Unfortunately, this relative frequency is typically too small for accurate direct estimation. A further simplification is obtained by assuming that the two allele lengths are independent, which is reasonable in the absence of assortative mating and provided that selection is weak. Under this assumption we have

$$R_1 = 2p_f p_m \quad (4)$$

where  $p_f$  and  $p_m$  are the relative frequencies in  $Q$  of alleles which match, respectively, the paternal and the maternal criminal alleles. It is convenient to distinguish the paternal from the maternal allele in the notation, but the symmetry of Mendelian inheritance makes it unnecessary to identify which is which. The factor of two appears in (4) because the order of the two alleles is unimportant; for example, the criminal maternal allele may be matched by either the maternal or the paternal allele of the suspect.

There are several difficulties with formula (4). Perhaps the most obvious difficulty is the specification of the population  $Q$ . An appropriate choice will depend on the nature of the evidence which led to the identification of the suspect. For example, there may be evidence which restricts suspicion to residents of a particular region and/or members of a particular racial group. If the suspect has been identified in such a way that his or her close relatives are also likely suspects, then hypothesis  $I$  is inappropriate as an alternative to  $G$  and a different assumption which takes precise relationships into account should be considered. There are additional problems associated with the variability of sample estimates of small relative frequencies and the assumption that measurement error can reasonably be characterized by a step function (Gill *et al.*, 1990). These difficulties cannot be entirely overcome to obtain a satisfactory expression for  $R_1$ . However, it suffices to approximate  $R_1$  provided that the approximation is conservative. Here, 'conservative' means erring in favour of the suspect so that the probability of accepting hypothesis  $G$ , when it is false, is not increased. On the other hand, an excessively conservative approximation leads to a small probability of accepting  $G$  when it is true, so that many criminals will be incorrectly exonerated. Hence, we would wish to replace  $R_1$  with the smallest value which can be confidently asserted to be larger than the true value.

The problems associated with (4), mentioned above, are discussed elsewhere (D. J. Balding & R. A. Nichols, in preparation). Here we focus on a further problem:  $Q$  is typically a large population with a complex structure, which may include a distinct sub-population  $Q'$  from which both suspect and criminal are drawn. This situation is likely to be common because, if a suspect is innocent, the police search may have concentrated either on associates of the actual criminal or on people fitting a witness' description. Allele frequencies in  $Q'$  may differ from those of  $Q$  due to a variety of causes such as genetic drift and migration patterns. The observation of suspect and criminal sharing bands which are rare in  $Q$ , usually interpreted as providing strong support for hypothesis  $G$ , also suggests that such a structure exists and the bands are more common

in  $Q'$  than in the whole population  $Q$ . Hence  $R$  overstates the evidence for  $G$ . We claim that a more appropriate measurement of the strength of the evidence for  $G$  is given by  $R'$ , obtained by replacing  $I$  in (1) with  $I'$ , the hypothesis that both suspect and criminal are chosen randomly from the sub-population  $Q'$ . In practice,  $R'$  is larger than  $R$  and hence assuming  $I'$  when, in fact, there is no distinct sub-population including both suspect and criminal leads to an error in the conservative direction, unlike the error incurred by inappropriately assuming  $I$ . Further, Slatkin & Barton (1989) show that models with discrete sub-populations give equivalent results to models with spatially continuous sub-population distributions and hence we believe that conclusions based on  $I'$  will also apply to more general population structures.

In this paper we draw on the current knowledge of the genetics of human populations to seek a reasonable upper bound on  $R'$ . First, we discuss sampling from a population with the structure specified by hypothesis  $I'$  and we introduce three parameters commonly used to summarize population structure. The relevant value of the key parameter  $F_{ST}$  cannot be directly estimated because, under  $I'$ , the sub-population  $Q'$  is unidentified. This is appropriate because it is, in practice, impossible to specify the correct sub-population to consider and, in any case, estimates of allele frequencies are typically available for  $Q$  but not for any sub-population. Current forensic practice is to estimate frequencies for broadly defined racial groups (Caucasian, Afro-Caribbean, South Asian, Arabic, etc.), but not for any sub-populations generated by, for example, geographic, religious and social divisions within these groups. Next we argue that it is possible to make use of studies which have appeared in the literature to identify an empirical upper bound on  $F_{ST}$  and derive an expression for  $R'$  in terms of  $F_{ST}$ . We conclude by comparing values of  $R$  with  $R'$  for some particular cases and suggesting avenues to improve the method of analysis.

### Sampling from a simple, structured population

Consider a simple population structure in which the population  $Q$  includes a discrete, partially isolated sub-population  $Q'$ . Initially nothing is known about  $Q'$  so the probability that a random choice from  $Q'$  results in an allele  $A$  is simply  $p$ , the population frequency of  $A$ . However, given this outcome, the probability that a second, independent selection in  $Q'$  also results in  $A$  is greater than  $p$ , as the first draw suggests that the allele is more common in  $Q'$  than in  $Q$ . The magnitude of this effect depends on the variability of the relative fre-

quencies in  $Q'$  about the corresponding values in  $Q$ . A common assumption (Wright, 1969) is that the relative frequency of  $A$  in  $Q'$  is random with mean  $p$  and variance  $p(1-p)F_{ST}$ . This assumption holds exactly in the simple case that the members of  $Q'$  are chosen randomly from  $Q$ . The parameter  $F_{ST}$  is constant over alleles and is interpreted as a measure of the difference between the allele distributions in  $Q$  and  $Q'$ , and is sometimes referred to as the coancestry (Weir & Cockerham, 1984). The probability of drawing a second  $A$  under this assumption, given that the first draw results in  $A$ , is

$$Pr[AA|A] = p + F_{ST}(1-p). \quad (5)$$

The derivation of (5) is explained in the Appendix. We will follow the custom of loosely referring to  $F_{ST}$  as a correlation since it can also be interpreted as an expectation (with respect to the allele distribution in  $Q'$ ) of the correlation of the two selections from  $Q'$ , expressed in terms of the allele distribution in  $Q$ . Of course, if  $Q'$  is not identified then  $F_{ST}$  is unknown, but it has been estimated for many known human populations. These estimates suggest the range of realistic values of  $F_{ST}$  and hence of the probability (5). We argue that a formula involving  $F_{ST}$  appropriately reflects our knowledge of the forensic situation: we do not know the particular sub-population relevant to a given case but we do have, in the population genetics literature, an accumulated reservoir of knowledge of the structure of many actual human populations and this is conveniently formulated as estimates of the parameter  $F_{ST}$ .

As the identification of human sub-populations and hence the estimation of  $F_{ST}$  are difficult, another measure is useful for direct estimation: the correlation between the two alleles within an individual. If the alleles are the same, the individual's parents had a matching band. This correlation can be calculated with respect to the allele's frequency either in  $Q$  or in  $Q'$ ; the quantities are designated  $F_{IT}$  and  $F_{IS}$  respectively. The subscript notation should be interpreted as I = individual; S = sub-population; T = total population (Wright, 1969). In the presence of random mating,  $F_{IT}$  equals  $F_{ST}$ . Breeding between close relatives is avoided in almost all human populations, so  $F_{IT}$  is less than  $F_{ST}$ . However, conventional police work or further testing can usually be used to exclude close relatives of the suspect from consideration, so that it may be appropriate to consider  $F_{IT}$  rather than  $F_{ST}$ .

Devlin *et al.* (1990) found no evidence for an excess of homozygotes over the Hardy-Weinberg expectations in one of the larger databases of DNA fingerprint records. This suggests that  $F_{IT}$  was not significantly greater than zero. However, human populations are known to be composed of large relatively outbred

populations, typically in cities, and smaller inbred populations. In a random sample from both, the effects of inbreeding in the small populations will be swamped. We must examine studies of inbred populations for estimates of the maximum values of  $F_{IT}$  and  $F_{ST}$ .

### Estimating genetic correlations

For a well-defined sub-population of a human population, genetic correlations can be estimated using records of ancestry movement, or directly from genetic data. The records have been used to construct analyses of genealogies, the distribution of family names, and patterns of movement. Genetic surveys are usually restricted to smaller samples for practical reasons, but they can be used to corroborate other methods.

Genealogies can be used to predict genetic correlations because for any two individuals, the number of generations back to a common ancestor can be counted. Pairs with more recent common ancestors are more likely to share an allele identical by descent. The distribution of common ancestors can therefore be used to calculate  $F_{ST}$  (Wright, 1922). It is not possible to account directly for common ancestors that precede the records; however, calculations based on the frequency of shared family names (isonymy) will tend to overestimate  $F_{ST}$  (e.g. Jorde, 1989) because unrelated families may share the same name.

Cavalli-Sforza & Bodmer (1971) surveyed the literature in which these techniques have been used. The most extreme cases were correlations of 5 per cent; these correspond to severe inbreeding, such as that associated with a tradition of uncle-niece marriages. The largest values found in Europe were an order of magnitude smaller, and more recent surveys show dramatic reductions associated with increased mobility due to modern transport (Vogel & Motulsky, 1979). More typically, values are another order of magnitude smaller. Hence the value 5 per cent appears to be very conservative for any large population and smaller values would be appropriate in cases where extreme inbreeding is known not to occur.

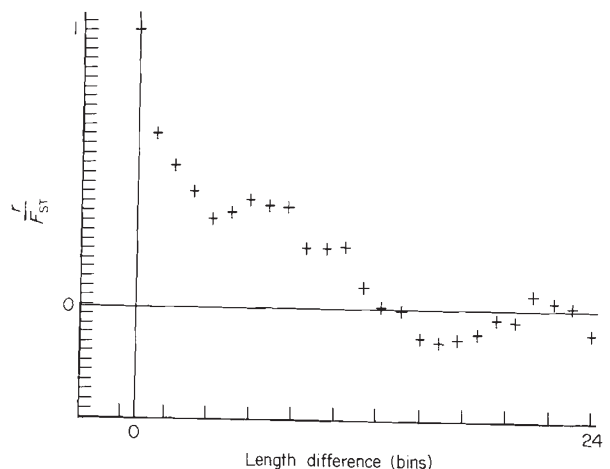
So far we have assumed that the sub-population  $Q'$  is well defined. However, there may be less genetic variation between, for example, a village and the surrounding towns, than between the village and the whole country. In theoretical models this effect is only important over scales less than the characteristic length  $l_c$  (Slatkin & Barton, 1989). In neutral models  $l_c = \sigma/\sqrt{2\mu}$ , where  $\sigma^2$  is the dispersal variance and  $\mu$  is the mutation rate. However, in human populations there is a broader scale variation associated with the history of colonizations (e.g. Sokal *et al.*, 1989). As the

scale of variation depends on the mutation rate and selection, it will vary from locus to locus. Therefore this variation, at a scale intermediate between  $Q$  and  $Q'$ , will affect the genetic correlations in a way that depends on the locus,  $Q$  and  $Q'$ , and could inflate  $F_{ST}$ . At least in some cases the effect is small (Prescittini *et al.*, 1989). The size of the effect needs to be verified for fingerprint loci; but it is expected to be smaller than at other loci because fingerprint loci have particularly high mutation rates.

### The probability of match in a structured population

In the previous two sections we have discussed calculations of match probabilities under hypothesis  $I'$ , ignoring measurement error, and conservative estimates of the parameter  $F_{ST}$ . We now turn to calculating match probabilities under  $I'$ , accounting for the measurement error specified by  $M$  at (3). Hence we need to consider not only the case that two alleles are the same, but also the case that their lengths are similar.

Balazs *et al.* (1989) found substantially different distributions of allele lengths in three US ethnic groups. Their histograms show that the allele lengths have a smoothness property: cells of each histogram tend to have a similar frequency to neighbouring cells. The size and scale of this effect are summarized in Fig. 1, which



**Fig. 1** Correlations among alleles of different lengths. Data from fig. 3 of Balazs *et al.* (1989) for three ethnic groups in the USA ('Black', 'Caucasian', and 'Hispanic') at the D2S44 locus.  $F_{ST}$  is calculated as the weighted average of the variation of the individual bin frequencies for each ethnic group around the combined mean  $p$  divided by  $p(1-p)$ , and has value 0.0042 (a bin is approximately 1.3 per cent of the allele length). The average correlation ( $r$ ) was calculated between the frequencies in each ethnic group and the frequencies in the adjacent 24 bins. The value plotted is the average value of  $r/F_{ST}$  for length differences of 0–24 bins.

plots the sample correlation (with respect to the total sample mean) of frequencies of cells separated by given distances. Some apparent smoothing can be attributed to measurement error, as alleles of similar lengths will be confused. However, Fig. 1 indicates large positive correlations over distances much greater than the range of measurement error. This smoothing may be a result of the mutation process, which appears to generate new alleles of lengths similar to the progenitor allele (Jeffreys *et al.*, 1990).

Hence, given a criminal allele, population structure effects lead not only to a higher probability of observing the same allele in the suspect sample but also the chance of observing alleles of similar length is enhanced compared with that due to a random selection in  $Q$ . However, the correlation of similar-length alleles will be smaller than  $F_{ST}$  if, as we believe, it is due to the interaction of inbreeding (of magnitude  $F_{ST}$ ) and other processes (principally mutation) which propagate, but attenuate, the effect to alleles of adjacent lengths. This belief is supported by the data of Balazs *et al.* (Fig. 1).

Therefore, the probability of match for a single band is conservatively estimated by (5) on replacing the allele frequency  $p$  with the cell frequency  $p_f$  or  $p_m$  and choosing an overestimate of  $F_{ST}$ . Finally, approximating once again the appropriate genotype frequency as the product of the two cell frequencies, the single-locus value  $R'_1$  is given by

$$R'_1 = 2[p_f + F_{ST}(1 - p_f)][p_m + F_{ST}(1 - p_m)]. \quad (6)$$

Let us illustrate the calculation for single-locus data in the case that the criminal genotype consists of two rare alleles, for each of which the frequency of alleles within the assumed range of measurement error is equal to 1 per cent. Using (3), which ignores population structure effects, the likelihood of a match is

$$R_1 = 2 \times 10^{-4},$$

while invoking (6) with a conservative value of 5 per cent for  $F_{ST}$  gives

$$\begin{aligned} R'_1 &= 2(0.01 + 0.99 \times 0.05)(0.01 + 0.99 \times 0.05) \\ &= 7.1 \times 10^{-3} \end{aligned}$$

so that  $R'_1$  is about 35 times  $R_1$ . When multiplied over several loci, the ratio of  $R'$  to  $R$  can be substantial.

We claim that the astronomically small likelihoods which have been reported to juries (Lander, 1989) are misleading because of the relative magnitude of the population genetic effects. Furthermore, inference from matches of rare alleles is difficult because of the relatively high variability of estimates of the population frequencies. In the other direction, choosing the cut-off

value  $\delta$  sufficiently large to reflect actual measurement errors and using a conservative value of  $F_{ST}$  can lead to single-locus likelihoods  $R'_1$  as high as 0.1, so that a larger number of loci than has previously been suggested may be required for reliable discrimination.

### Future work

We have argued here that  $R'$  should be used for inference in preference to  $R$  since it tends to err in a conservative direction. We now suggest two approaches to making more efficient use of the data while remaining conservative: (a) a more accurate estimate of the correlation between alleles of similar length and (b) an analysis which draws inferences directly from the observed allele lengths without considering the event 'match'.

The analysis proposed in this paper employs  $F_{ST}$  as an estimate of the correlation of similar-length alleles from individuals within a sub-population. For non-identical alleles this will be an overestimate. It would be more efficient to estimate the correlation between similar-length alleles empirically. This is possible if a collection of fingerprints can be obtained from populations known to be inbred. The correlation between the lengths of an individual's two alleles would give an estimate of the correlation between the allele lengths in the parents (minus the measurement error between gels, and gel tracks).

The assumption of a step function (3) implies that a match is declared for all pairs of alleles whose difference in length is less than  $\delta$ . This assumption simplifies the analysis but is clearly unrealistic and is also unsatisfactory as several distinct alleles may lie within  $\delta$  of the criminal allele and each of these is weighted equally in the outcome 'match'. A more efficient analysis would be obtained by making inferences directly from the observed allele lengths. A possible approach is to specify a probability model of measurement error as a function of true lengths (a Normal assumption may be appropriate) and then to use Bayes' rule to calculate likelihoods of the true allele lengths under the two hypotheses, given the observed lengths, and employing a prior allele length distribution based on population estimates. Hence, the cell frequencies  $p_f$  and  $p_m$  would be replaced with a weighted sum over alleles near the observed lengths, with the weights summing to unity. A similar analysis, but ignoring population structure, has been advocated by Evett *et al.* (1989), however it appears not to be widely used.

## Acknowledgements

We would like to thank Geoff Watterson of Monash University and Cecilia Buffery, Matthew Greenhalgh and Geoff Willott of the Metropolitan Police Forensic Science Laboratory for valuable discussions.

## References

- BALAZS, I., BAIRD, M., CLYNE, M. AND MEADE, E. 1989. Human population genetic studies of five hypervariable DNA loci. *Am. J. Hum. Genet.*, **44**, 182–190.
- CAVALLI-SFORZA, L. L. AND BODMER, W. F. 1971. *The Genetics of Human Populations*. W. H. Freeman, San Francisco.
- COHEN, J. E. 1990. DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number volatility. *Am. J. Hum. Genet.*, **46**, 358–368.
- DEVLIN, B., RISCH, N. AND ROEDER, K. 1990. No excess of homozygosity at loci used for DNA fingerprinting. *Science*, **249**, 1416–1420.
- EVETT, I. W., WERRETT, D. J., GILL, P. AND BUCKLETON, J. S. 1989. DNA fingerprinting on trial. *Nature*, **340**, 435.
- GILL, P., SULLIVAN, K. AND WERRETT, D. J. 1990. The analysis of hypervariable DNA profiles: problems associated with the objective determination of the probability of a match. *Hum. Genet.*, **85**, 75–79.
- JEFFREYS, A. J., NEUMANN, R. AND WILSON, V. 1990. Repeated unit variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell*, **60**, 473–485.
- JEFFREYS, A. J., WILSON, V. AND THEIN, S. L. 1985a. Hypervariable 'minisatellite' regions in human DNA. *Nature*, **314**, 67–73.
- JEFFREYS, A. J., WILSON, V. AND THEIN, S. L. 1985b. Individual-specific 'fingerprints' of human DNA. *Nature*, **316**, 76–79.
- JORDE, L. B. 1989. Inbreeding in the Utah Mormons: an evaluation of estimates based on pedigrees, isonymy, and migration matrices. *Ann. Hum. Genet.*, **53**, 339–355.
- LANDER, E. S. 1989. DNA fingerprinting on trial. *Nature*, **339**, 501–505.
- PRESCIUTTINI, S., PAOLI, G. AND FRANCESCHI, M. 1990. Surname versus gene structure of a small isolate. *Ann. Hum. Genet.*, **54**, 79–90.
- SLATKIN, M. AND BARTON, N. H. 1989. A comparison of three indirect methods for estimating average levels of gene flow. *Evolution*, **43**, 1349–1368.
- SOKAL, R. R., ODEN, N. L., LEGENDRE, P., FORTIN, M., KIM, J. AND VANDOR, A. 1989. Genetic differences among language families in Europe. *Am. J. Phys. Anthropol.*, **79**, 489–502.
- WEIR, B. S. AND COCKERHAM, C. C. 1984. Estimating  $F$ -statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
- WRIGHT, S. 1922. Coefficients of inbreeding and relationship. *Am. Nat.*, **56**, 330–338.
- WRIGHT, S. 1969. *Evolution and the Genetics of Populations*, Vol. 2, University of Chicago Press, Chicago.
- VOGEL, F. AND MOTULSKY, A. G. 1979. *Human Genetics*, Springer-Verlag, Berlin.

## Appendix

Prior to observing any data, we have assumed that  $p'$ , the relative frequency of  $A$  in  $Q'$ , is random with mean  $p$  and variance  $p(1-p)F_{ST}$ . Hence the probability of selecting allele  $A$  with a random draw from  $Q'$  is  $p$ , the prior mean of  $p'$ . This observation leads to an updated, or posterior, probability distribution for  $p'$  with pdf  $f_A$  given by Bayes' rule

$$f_A(x) = \frac{x\pi(x)}{\int_0^1 y\pi(y) dy} = \frac{1}{p} x\pi(x), \quad x \in [0, 1],$$

where  $\pi$  denotes the prior pdf of  $p'$ . Hence the probability of drawing a second  $A$ , given the first, is the posterior mean of  $p'$  given by

$$\begin{aligned} Pr[AA|A] &= \int_0^1 x f_A(x) dx = \frac{1}{p} \int_0^1 x^2 \pi(x) dx \\ &= p + (1-p)F_{ST}. \end{aligned}$$