Conditional Genotypic Probabilities for Microsatellite Loci

Jinko Graham,*,†,[1] James Curran†,[2] and B. S. Weir†,*

**National Institute of Statistical Sciences,* †*Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203*

Manuscript received August 3, 1999
Accepted for publication April 17, 2000

ABSTRACT

Modern forensic DNA profiles are constructed using microsatellites, short tandem repeats of 2–5 bases. In the absence of genetic data on a crime-specific subpopulation, one tool for evaluating profile evidence is the match probability. The match probability is the conditional probability that a random person would have the profile of interest given that the suspect has it and that these people are different members of the same subpopulation. One issue in evaluating the match probability is population differentiation, which can induce coancestry among subpopulation members. Forensic assessments that ignore coancestry typically overstate the strength of evidence against the suspect. Theory has been developed to account for coancestry; assumptions include a steady-state population and a mutation model in which the allelic state after a mutation event is independent of the prior state. Under these assumptions, the joint allelic probabilities within a subpopulation may be approximated by the moments of a Dirichlet distribution. We investigate the adequacy of this approximation for profiled loci that mutate according to a generalized stepwise model. Simulations suggest that the Dirichlet theory can still overstate the evidence against a suspect with a common microsatellite genotype. However, Dirichlet-based estimators were less biased than the product-rule estimator, which ignores coancestry.

SEVERAL authors (*e.g.*, Balding and Nichols 1994, 1995; Lange 1995) have discussed the need to account for the coancestry of individuals when assessing the evidential strength of matching DNA profiles in forensic identification. Matching profiles could reflect genetic homogeneity of a subpopulation, rather than guilt of the suspect. Hence, a fair assessment of DNA profile evidence should allow for the possibility that the suspect and perpetrator belong to the same subpopulation (Weir 1994). Often, there are no data available on a crime-specific subpopulation. In such instances, standard product-rule estimators of match probabilities (National Research Council 1992) assume that the effects of population subdivision are negligible. However, Balding and Nichols (1997) examined the genetic correlations quantifying population differentiation among Caucasians and concluded that coancestry was too large to be ignored. They found that product-rule estimators of match probabilities can, in many cases, overstate the strength of evidence against the suspect.

Theory has been developed to account for the effects of coancestry on match probabilities. The theory relates the population-wide genotype probabilities to the expected joint allele frequencies within a crime-specific subpopulation for which there are no genetic data. A mutation model is assumed in which the allelic state after a mutation event is independent of the state prior to mutation (Wright 1951; Griffiths 1979). Under this source-invariant mutation model, the joint allele probabilities within a subpopulation may be expressed in terms of the marginal probabilities of the alleles and identity-by-descent measures appropriate to the genetic model. Joint allele probabilities determine match probabilities.

Balding and Nichols (1994) assumed a subdivided population of constant size and used a coalescent argument to arrive at expressions for the joint allele probabilities within a subpopulation. Subpopulations were not necessarily independent because of migration and common history. A genetic replicate was therefore defined as the combined evolutionary history of the subpopulations. These authors showed that when the marginal probabilities of an allele have reached a steady state, the joint allele probabilities within a subpopulation match the moments of a Dirichlet distribution. Since the genetic model was formulated without reference to a base population, measures of identity by descent were defined in terms of the coalescence of ancestral lines, without intervening mutations or migrations. The measures therefore depend on the rates of mutation, migration, and coalescence. Coalescence rates, in turn, depend on population size. Hence, under constant population size and constant mutation and migration

Corresponding author: Bruce Weir, Program in Statistical Genetics, Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27695-8203. E-mail: weir@stat.ncsu.edu

[1] *Present address:* Department of Mathematics and Statistics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada.

[2] *Present address:* Department of Statistics, University of Waikato, Hamilton, New Zealand.

rates, the measures of identity by descent remain constant over time.

Weir and Cockerham (1984) proposed an estimator of the coancestry coefficient under a genetic model in which each subpopulation is constructed by randomly drawing individuals from a base population of infinite size. At the time of the base population, the probability of drawing an allele is assumed to be in steady state. The expected value of an allele frequency in the subpopulation is therefore equivalent to the allele frequency in the base population. Thereafter, subpopulations are assumed to evolve under similar demographic conditions. Inference is conditional on allele frequencies in the base population, and each subpopulation represents an independent genetic replicate. In this prospective genetic model, identity by descent is defined with respect to the base population and therefore decays with the time elapsed since the base population. Under equilibrium of descent measures within subpopulations, higher-order descent measures can be written in terms of the pairwise measure of identity by descent (Li 1996). Then joint allele probabilities within a subpopulation have the Dirichlet form as well, but are defined in terms of parameters in the prospective genetic model. Li (1996) used the resulting expressions to approximate joint allele probabilities early in the history of constant-size subpopulations and found that the approximation performed well.

Both approaches invoke the equilibrium distribution of the frequencies of an allele under the source-invariant mutation model, in populations of constant size. Under the source-invariant mutation model, the equilibrium joint allele frequencies within a subpopulation are approximately Dirichlet (Wright 1951; Griffiths 1979). However, for microsatellite DNA profiles, a stepwise mutation model would more realistically reflect replication slippage (Levinson and Gutman 1987) than a source-invariant model. Although no equilibrium distribution exists under stepwise mutation (Moran 1975), the Dirichlet approximation is increasingly used to account for coancestry in assessing forensic evidence from microsatellite profiles. We therefore investigate the adequacy of the Dirichlet approximation for a hypothetical subpopulation in which alleles mutate according to a generalized stepwise mutation model.

There are a large number of stepwise mutation models, starting with the one- and two-step versions proposed for electrophoretic alleles (*e.g.*, Ohta and Kimura 1973; Wehrhahn 1975; Moran 1975; Li 1976; Weir *et al.* 1976). The generalized stepwise model we have selected is parsimonious and sufficiently flexible to accommodate rare mutations of size larger than a few repeats, a pattern that is suggested by human population samples (DiRienzo *et al.* 1994). However, the model does not accommodate allele-specific mutation rates, such as the higher rates observed for longer repeats in human samples (Brinkmann *et al.* 1998). Nevertheless, the model has been successfully applied previously as a useful first approximation to the complex process of microsatellite evolution (Fu and Chakraborty 1998). Throughout, we rely on a simplified demographic model, with parameters chosen to reflect both a modern population such as New Zealand caucasians and historical human population estimates from the literature (Harpending *et al.* 1998; Kruglyak 1999).

METHODS

Demographic parameters: To simplify the analysis, we assumed the same demographic history for all subpopulations in our simulation study. The current number of 2×10^6 individuals in a subpopulation was chosen to be typical of the effective size of a modern subpopulation such as New Zealand caucasians. Subpopulations were not of constant size over time, but instead underwent exponential growth. Each subpopulation arose 5000 generations before present (gbp) from 500 random individuals in a population of size 10,000 individuals. Subsequently, each subpopulation evolved independently of the others, with no migration. Prior to 5000 gbp, the size of the metapopulation giving rise to the subpopulations was constant at 10,000 individuals, and there was no subdivision. Historical population sizes are based on estimates from the literature (*e.g.*, Harpending *et al.* 1998; Kruglyak 1999), which suggest that the approximate date of the *Homo sapiens* migration out of Africa is $G \approx 5000$ gbp, and that the effective population size prior to the migration was $N \approx 10{,}000$ individuals.

The simple demographic model we have selected reflects the historical expansion of subpopulations after the migration out of Africa at 5000 gbp. However, prior to this migration, a single panmictic population is assumed. The impact of subdivision in the African source population was therefore explored in further simulations by assuming five subpopulations, each of constant size 2000 individuals, during the interval from 5000 to 80,000 gbp. Prior to 80,000 gbp, a single panmictic population of size 10,000 individuals was assumed. All parameters associated with times more recent than 5000 gbp were kept the same as before.

Mutation model: Let π_{ij} be the probability that a mutation causes an allele of size i repeats to change to an allele of size j repeats. Fu and Chakraborty (1998) proposed a generalized stepwise mutation model in which π_{ij} depends on i and j only through $|i - j|$. Their mutation model does not accommodate allele-specific mutation rates, such as the higher rates observed for longer repeats in human samples (Brinkmann *et al.* 1998), nor does it accommodate constraints on allele size. However, a homogeneous distribution was preferred because it was flexible yet parsimonious and because only the relative sizes of alleles were known. Under their generalized stepwise mutation model,

TABLE 1

Allele frequencies of microsatellite marker D8S1179 in New Zealand caucasians

	Allele length in repeat units									
	8	9	10	11	12	13	14	14	16	17
Percentage	2.13	0.45	11.63	7.72	14.54	33.33	17.56	9.40	3.02	0.22

$$\pi_{ij} = \begin{cases} \alpha P(1 - P)^{j-i-1}, & j > i \\ (1 - \alpha)P(1 - P)^{i-j-1}, & j < i. \end{cases}$$

The parameter α describes the probability of an increase in repeat number; the size $|i - j|$ of the resulting change in allelic length has a geometric distribution with probability $P(1 - P)^{|i-j|-1}$.

Other parameters of the model include the mutation rate μ and the length A of the allele of the most recent common ancestor (MRCA) of the sample. We have selected a sample of size 1000 chromosomes. Simulations indicate that with high probability ($\sim$0.998) the sample MRCA coincides with the MRCA of the population. (Even with a more modest sample size of 100 chromosomes, the probability is still very high at $\sim$0.980.) Hence, A may also be viewed as the allelic length of the MRCA of the population. Given A and the realized ancestral tree, microsatellite mutations can be placed on the tree, from the root to the tips, as described by Fu and Chakraborty (1998). Conditional on the length of a segment of the tree, the number of mutation events on the segment is approximated by a Poisson random variable, with mean equal to the product of the mutation rate and the segment length.

For the simulation study, we chose $A = 9$, $\mu = 5 \times 10^{-4}$, $\alpha = 0.720$, and $P = 0.999$. These parameter values produce simulated allele frequencies consistent with observed frequencies for the microsatellite D8S1179 in a sample of 447 New Zealand caucasian offenders, shown in Table 1. The selected parameter values also reflect estimates from the literature. Microsatellites have a high mutation rate of $\sim 10^{-4}$–10^{-3} per generation (Gyapay *et al.* 1994). Most observed mutations result in a change of a single repeat unit (Weber and Wong 1993; DiRienzo *et al.* 1994), but there are rare events with larger changes and a tendency toward increasing allelic length (Brinkmann *et al.* 1998). On the basis of these observations, a plausible range of values for microsatellite mutation parameters includes $10^{-4} \leq \mu \leq 10^{-3}$, $0.5 \leq \alpha < 1$, and $0.5 \leq P < 1$. Selecting $P = 0.999$ implies 99.9% of mutations result in a size change of 1 repeat unit, whereas $P = 0.500$ implies that >99% of mutations are expected to result in a change of 7 or fewer repeat units. In further simulations, alternative parameter values were also explored, by perturbing the D8S1179 values one at a time ($\mu = 1 \times 10^{-4}$, 3×10^{-4}, 5×10^{-4}, 9×10^{-4}, 1×10^{-3}; $\alpha = 0.50, 0.60, 0.72, 0.90, 0.99$; and $P = 0.500, 0.800, 0.900, 0.999$) and by examining values at the end points of the plausible range, for the New Zealand demographic model.

Allelic associations: Following the notation of Evett and Weir (1998), consider a microsatellite locus **A** with alleles A_i of length i repeat units in a randomly mating subpopulation. Let p_i and P_{ij} denote, respectively, the probability of drawing an allele A_i and the probability of drawing an individual with genotype $G = A_iA_j$ in a subpopulation at present. In both genetic models, p_i is assumed to be in steady state. We emphasize that p_i and P_{ij} are, respectively, the allelic and genotypic probabilities averaged over repeated replicates of a subpopulation, not fixed-population probabilities. Under the source-invariant mutation model, genotype probabilities may then be described by

$$P_{ii} = \theta p_i + (1 - \theta)p_i^2$$
$$P_{ij} = 2(1 - \theta)p_ip_j, \quad i \neq j,$$

where the coancestry coefficient θ is specific to the genetic model. In the genetic model of Weir and Cockerham (1984), θ is the probability that two alleles drawn from a subpopulation at present are identical by descent with respect to the base population. In the genetic model of Balding and Nichols (1994), θ is the probability that two alleles from the same subpopulation coalesce with no intervening mutation events on the lines of descent. However, for the high mutation rates typical of microsatellite markers, both measures are virtually identical given the New Zealand demographic parameters. Figure 1 shows the coancestry coefficient, measured first with respect to the base population at 5000 gbp, and then without reference to a base population. The coancestry coefficient, like the allelic and genotypic probabilities p_i and P_{ij}, is defined in terms of subpopulation replicates and is not a fixed-population parameter. Coancestry coefficients were determined empirically, on the basis of 10^7 coalescent replicates for a random pair of chromosomes from a subpopulation. At the mutation rate $\mu = 0.0005$ selected for D8S7911, both coancestry coefficients are $\sim$0.008, a reasonable value given numerical estimates from population surveys (Cavalli-Sforza *et al.* 1994).

To gain insight into the adequacy of the source-invariant mutation model for microsatellites, we compared association parameters describing P_{ij} under the stepwise mutation model to the analogous quantity under the source-invariant mutation model. Associations were determined empirically, based on 10^7 coalescent replicates for a random pair of chromosomes within a random sample of 1000 chromosomes. The within-subpopulation correlation for an allele of length i repeat units is

$$\theta_{ii} = \frac{P_{ii} - p_i^2}{p_i(1 - p_i)}.$$

Under the source-invariant mutation model, this correlation coincides with the coancestry coefficient θ; hence $\theta_{ii} \equiv \theta$. More generally, however, θ_{ii} can vary with allele length i. Another measure of association within a subpopulation, between two alleles of different lengths $i \neq j$, is

$$1 - \theta_{ij} = \frac{P_{ij}}{2p_ip_j}.$$

In a stepwise mutation model, θ_{ij} is expected to vary with the allelic states (Balding and Nichols 1994), with positive association between alleles of similar length. Such variation is not accommodated by the source-invariant mutation model, which constrains $\theta_{ij} \equiv \theta$. As a diagnostic for the fit of the source-invariant mutation model, we examined the departure of θ_{ii} and θ_{ij} from the coancestry coefficient θ.

Predicted match probabilities: Balding and Nichols (1994) showed that, within a randomly mating subpopulation of constant size, joint allele probabilities match the moments of a Dirichlet distribution, provided that the marginal probabilities of an allele are in steady state and that there are no length-dependent correlations among frequencies. They expressed joint allele probabilities in terms of the marginal probabilities p_i of an allele and the coancestry coefficient θ and used them to derive formulas for match probabilities. For a suspect ($\mathcal{S}$) with genotype G_S and perpetrator ($\mathcal{P}$) with genotype G_P, (1) gives the expressions for genotypes A_iA_i and A_iA_j, $i \neq j$, respectively (Evett and Weir 1998):

$$\Pr(G_P = A_iA_i|G_S = A_iA_i) = \frac{[p_i + \theta(2 - p_i)][p_i + \theta(3 - p_i)]}{(1 + \theta)(1 + 2\theta)}$$

$$\Pr(G_P = A_iA_j|G_S = A_iA_j) = \frac{2[p_i + \theta(1 - p_i)][p_j + \theta(1 - p_j)]}{(1 + \theta)(1 + 2\theta)}. \tag{1}$$

Empirical match probabilities were compared to those predicted by these equations, using empirically determined values of p_i and assigned values of $\theta = 0.010$, 0.050, 0.100, and 0.150. The values of $\theta = 0.100$ and 0.150 are particularly conservative for forensic calculations (Cavalli-Sforza *et al.* 1994). Empirical values were based on 10^7 coalescent replicates for a random sample of four chromosomes, within a random sample of 1000 chromosomes from a subpopulation. Empirical match probabilities were calculated by dividing the observed probability of drawing two members of a subpopulation with the given genotype by the observed probability of that genotype.

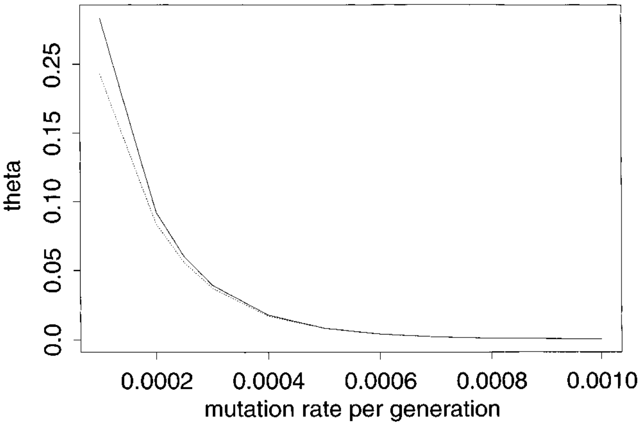

Figure 1.—Coancestry coefficient *vs.* mutation rate. Solid line, pairwise probability of identity by descent without reference to a base population; dotted line, with reference to the base population at 5000 gbp.

Estimated match probabilities: We also examined the bias, over subpopulation replicates, of the product-rule estimator and an estimator based on the Dirichlet equations (1). The Dirichlet-based estimator is formulated unconditionally, over repeated realizations of populations or sets of populations. In contrast, the product-rule estimator is formulated conditional on the observed population. However, bias of the product-rule estimator across subpopulation replicates should reflect a tendency toward bias at the fixed population level.

Typically, forensic databases are constructed using convenience samples from a limited number of subpopulations. To mimic such data, we simulated the ancestry of random samples of 1000 chromosomes from each of five subpopulations with demographic and mutation model parameter values reflecting D8S7911 in New Zealand. For each coalescent replicate, the samples were used to build a database of simulated microsatellite allele frequencies. The overall database frequency f_i of an allele of size i repeats was used to estimate the expected frequency p_i. The product-rule estimator of match probability is $2f_if_j$ for a suspect and perpetrator with genotype A_iA_j, $i \neq j$, and f_i^2 for a suspect and perpetrator with homozygous genotype A_iA_i.

Under a known coancestry coefficient and Dirichlet allele frequencies within subpopulations, a biased estimator that takes into account coancestry may be constructed by substituting database frequencies f_i for p_i into the Dirichlet equations (1). The Dirichlet match probability formulas are of the form $a_1 + b_1p_i + c_1p_i^2$ for A_iA_i homozygotes, and $a_2 + b_2(p_i + p_j) + c_2p_ip_j$ for A_iA_j heterozygotes, where a_1, a_2, b_1, b_2, c_1, and c_2 are constants with respect to p_i and p_j. Although the f_i are unbiased for p_i, substituting f_i^2 for p_i^2, or f_if_j for p_ip_j, into the formulas leads to bias because $E(f_i^2) = p_i^2 + k_1p_i(1 - p_i)\theta$, and

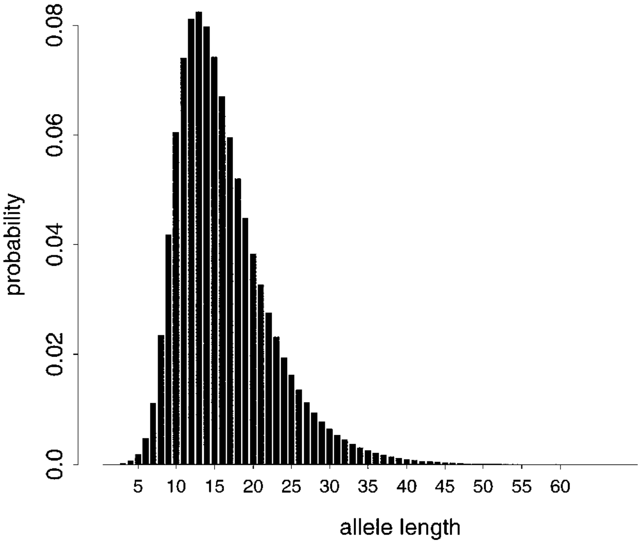

Figure 2.—Probability of sampling an allele of a given length.

$E(f_i f_j) = p_i p_j - k_2 p_i p_j \theta$, where k_1 and k_2 are constants with respect to θ, p_i, and p_j.

When the coancestry coefficient must be estimated, or when the Dirichlet approximation no longer holds, the properties of an estimator based on naive substitution are uncertain. We chose a moment estimator of θ, which is easy to calculate and combines coancestry information across subpopulations (Weir 1996). Lange (1995) used a maximum-likelihood approach to estimate the Dirichlet parameters with samples from several subpopulations. Balding and Nichols (1997) introduced a Bayesian approach to modeling variation in θ among subpopulations to address the possibility that subpopulations may have different degrees of coancestry, owing to differing demographic histories. However, in the current study, all subpopulations were simulated to have the same coancestry coefficient. Hence, modeling of variation in θ is unnecessary.

RESULTS

Figure 2 shows the probability of drawing an allele A_i of length i repeat units from a subpopulation at present under parameter values selected to reflect D8S7911 in New Zealand. The marginal distribution has a longer right tail, with a mode of 13 repeat units, and a mean of ~16 units. Over 90% of the time, an allele is between 9 and 28 repeat units in length. The mode and longer right tail of the distribution are consistent with the ancestral allele $A = 9$ and the parameter $\alpha = 0.720$ describing the probability of an increase in allelic length given a mutation event. Generally, over time, the mode of allele frequencies within a subpopulation tends to drift toward higher repeat numbers. Long ancestral trees tend to have more such drift and a larger spread of allele lengths than shorter trees. Shorter ancestral trees result in more tightly clustered lengths, closer to the ancestral allele. As predicted (Moran 1975), the spread of allele lengths within a subpopulation tends to be more stable than the mode, which can occasionally drift toward high repeat numbers. In fact, most variation in allelic length (~80% for the D8S7911 simulations) is observed across coalescent replicates rather than within a replicate.

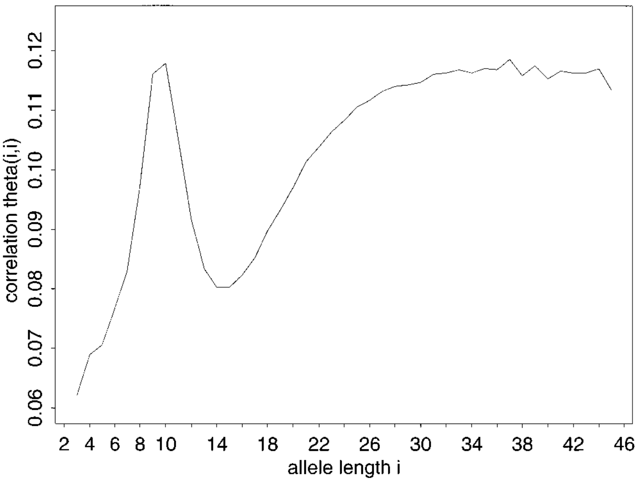

Figure 3.—Within-subpopulation correlations θ_{ii} for allele A_i.

Allelic associations: Allelic correlations θ_{ii} are plotted in Figure 3 for parameter values reflecting D8S7911 in New Zealand. The stepwise mutation model introduces excess correlation, above the correlation of $\theta = 0.008$ (the coancestry coefficient) that would hold under the source-invariant mutation model. The average correlation weighted by allele frequency is $\Sigma_i p_i \theta_{ii} \approx 0.095$. Correlation is high for alleles of length 9 and 10 repeat units, which are associated with shorter ancestral trees. Short ancestral trees have alleles that tend to be more tightly clustered in length. Correlation is lowest for alleles with very low repeat numbers, which tend to derive from long ancestral trees carrying alleles with a wider range of lengths. Further simulations indicate that, as expected, correlation is diminished at higher mutation rates and, when $\alpha = 0.5$, drops off symmetrically from the ancestral allele length of $A = 9$. Correlation is also reduced as the mutation model parameter P decreases, or the change in allelic length due to a mutation becomes more variable. The more variable the change in length, the wider the range of alleles within a subpopulation, and the lower the allelic correlation.

Figure 4 displays associations $1 - \theta_{ij}$ in the natural logarithmic scale for selected genotypes $A_i A_j$, $i < j$, under parameter values selected to reflect D8S7911 in New Zealand. The figure illustrates the general finding that the rarer the allele, the stronger the association with alleles of similar but unequal length. Further simulation

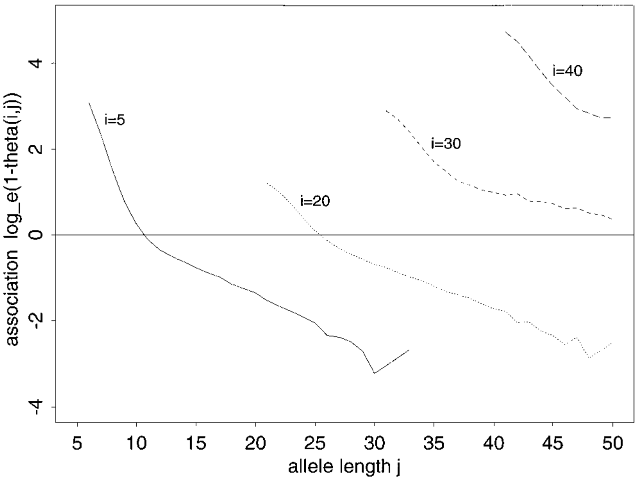

Figure 4.—Within-subpopulation associations $\log_e(1 - \theta_{ij})$ for alleles $j > i$; solid line, $i = 5$; dotted line, $i = 20$; short dashed line, $i = 30$; long dashed line, $i = 40$.

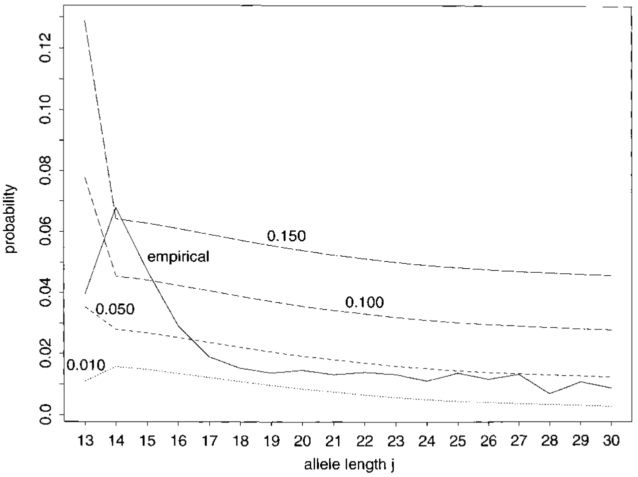

Figure 5.—Empirical and predicted match probabilities for selected genotypes A_iA_j, $i = 13$ and $j \geq i$. Solid line, empirical probabilities; dotted line, predicted probabilities $\theta = 0.010$; short dashed line, $\theta = 0.050$; medium dashed line, $\theta = 0.100$; long dashed line, $\theta = 0.150$.

results indicate that as the step-size parameter P is reduced, or the mutation rate μ is increased, the strength of association decreases. Smaller values of P imply more variable changes in allelic size (and larger mean step size), which, like larger mutation rates μ, lead to more variability in allelic size within a subpopulation. The positive association between distinct alleles of similar length is at odds with the negative Dirichlet association that is predicted by the source-invariant mutation model. Indeed, the overall weighted sum $\Sigma_{ij}P_{ij}\theta_{ij}$ for the D8S7911 simulations is $\sim$−3.6, quite far from the value of $\theta = 0.008$ predicted by the source-invariant mutation model.

These diagnostics indicate that the Dirichlet distribution does not fully capture the pairwise dependence of alleles under a stepwise mutation model. It is therefore reasonable to expect that the joint distribution of three and four alleles, and hence the predicted match probabilities, would also be misspecified. In the next section, we investigate the impact of the stepwise mutation model on Dirichlet match probabilities predicted by the source-invariant mutation model.

Predicted match probabilities: Figure 5 shows empirically determined match probabilities and those predicted under Dirichlet allele frequencies within a subpopulation for parameter values selected to reflect D8S7911 in New Zealand. Assumed values of the coancestry coefficient θ have been substituted in (1) for selected genotypes A_iA_j, $i = 13$ and $j \geq i$. This is consistent with current forensic practice of using assigned values for θ. For the common genotypes, predicted match probabilities systematically understate the empirical (true) match probabilities, except when the coancestry coefficient is taken to be very high. For example, the coancestry coefficient must be inflated to a value of 0.15, >18 times the true $\theta = 0.008$, to make the predicted match probability for the most common genotype $A_{13}A_{14}$ approximately correct. However, the resulting match probabilities for the $A_{13}A_{13}$ homozygote and the less common heterozygotes are then too conservative.

In further simulations, match probabilities increased with the mutation parameter P as the distribution of allelic lengths within a subpopulation became more concentrated. As the mutation rate increased, Dirichlet-based predictions of match probabilities became worse, particularly under small mean step-size ($P \rightarrow 1$) and asymmetric mutation ($\alpha \rightarrow 1$). For instance, under $\mu = 10^{-3}$, $\alpha = 0.990$, and $P = 0.999$, the true coancestry $\theta = 0.0005$ must be inflated by a factor of $\sim$260 to avoid understating match probabilities for more common genotypes. However, the resulting predictions for rarer genotypes were then as much as eight times too conservative.

Overall, the Dirichlet approximation performed better under low than under high mutation rates. For instance, at the low rate of $\mu = 10^{-4}$, Dirichlet match probability predictions based on the true coancestry coefficient were reasonably accurate for common genotypes, especially under symmetric mutation ($\alpha = 0.500$). However, predictions for rare genotypes were still conservative, with some more than twice the true match probability. The variability in length of allelic change due to mutation (controlled by P) had little effect on performance.

Under additional population subdivision early in human history, both match probabilities and the coancestry coefficient ($\theta = 0.009$) were slightly increased, as expected. The true coancestry coefficient required inflation by a factor of $\sim$10 to avoid understating match probabilities for more common genotypes. However,

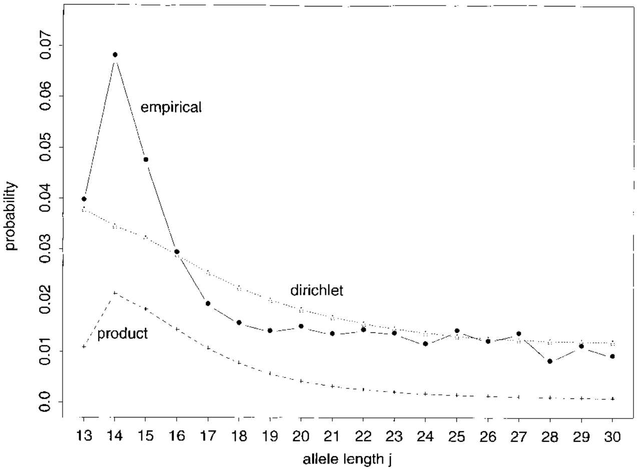

Figure 6.—Empirical match probabilities and expected values of match probability estimators for selected genotypes A_iA_j, $i = 13$ and $j \geq i$. Solid line, empirical probabilities; dotted line, expected value of Dirichlet-based estimator; dashed line, expected value of product-rule estimator.

the resulting predictions for rarer genotypes were then as much as three times too conservative.

Estimated match probabilities: Figure 6 shows empirically determined match probabilities and the expected values of match probability estimators under simulations reflecting D8S7911 in New Zealand for selected genotypes A_iA_j, $i = 13$ and $j \geq i$. The product-rule estimator is systematically biased, with a tendency to underestimation. The Dirichlet-based estimator is less biased, but still tends to understate match probabilities for common genotypes. For example, estimated match probabilities for a suspect with the more common genotype $A_{13}A_{14}$ are expected to be ∼51 and 31% of the true match probability for the Dirichlet-based and product-rule estimators, respectively. Under additional population subdivision early in human history, these match probability estimators were expected to be ∼40 and 15% of the true value, respectively.

The poorer performance of the product-rule estimator under increased subdivision is not surprising. Given that the product-rule estimator understates the true match probability, it is also unsurprising that for common genotypes so does the Dirichlet-based estimator. Predicted match probabilities for common genotypes A_iA_j involve larger marginal probabilities p_i and p_j in the numerator of (1). Larger p_i and p_j reduce the importance of the coancestry coefficient in the numerator and make predicted match probabilities more similar to those under the product rule.

To consider the implications of these results, suppose profile data from a subpopulation with demographic history similar to the hypothetical New Zealand population are available on 5 unlinked microsatellite loci, all with mutation parameters similar to those reflecting D8S7911. Then, in the case that the suspect carried the common genotype at all 5 loci, we would expect match probabilities to be underestimated by a factor of $0.51^5 = 3 \times 10^{-2}$ with the Dirichlet-based estimator and by a factor of $0.31^5 = 3 \times 10^{-3}$ with the product-rule estimator, assuming statistical independence of alleles at unlinked loci. For 10 loci, we would expect underestimation by factors of $\sim 1 \times 10^{-3}$ and 9×10^{-6}, respectively. This has implications for current FBI practice (reported in *Science* 278:1407, 1997) of not quoting match probabilities when these drop to some threshold value: it would seem to be important for these thresholds to be determined appropriately.

Further simulations explored the behavior of estimators under values at the end points of the plausible range for microsatellite mutation parameters. Estimated match probabilities for a suspect with the most common genotype were expected to be between 43 and 72% of the true match probability for the Dirichlet-based estimator and between 12 and 47% for the product-rule estimator.

DISCUSSION

Several aspects of population genetics require conditional and unconditional genotype probabilities. In forensic assessment of DNA profiles, conditional genotype probabilities are used to calculate match probabilities, which account for the effects of coancestry (Balding and Nichols 1995). The theory is based on a Dirichlet approximation to the joint distribution of allele frequencies. Assumptions include a source-invariant mutation model with steady-state distribution of allele frequencies (Wright 1951; Griffiths 1979) and populations of constant size. However, modern DNA profiles are often constructed using microsatellite markers, for which a stepwise mutation model would seem more realistic. Under stepwise mutation, there is no steady-state distribution of allele frequencies (Moran 1975).

We have used simulation to investigate the fit of Dirichlet-based match probabilities to those under a generalized stepwise model of mutation. Although a variety of demographic and stepwise mutation models may be applied, we have opted for simple versions as useful first approximations. Demographic parameters are chosen to reflect a modern population such as New Zealand caucasians, as well as historical human population size estimates from the literature. Mutation parameters are selected to be consistent with data for a microsatellite locus (D8S7911) used in forensic profiling of New Zealand caucasians. Further simulations explore the effect of additional population subdivision early in human history and different parameter values for the mutation model. Perturbations of the D8S7911 parameter values are explored, as well as more extreme values within the plausible range observed for human microsatellites.

Our results confirm that it is important to account for coancestry in assessments of DNA evidence. We find that the product-rule estimator is systematically biased, with a tendency to underestimate match probabilities. However, our results also illustrate potential problems with the growing use of the Dirichlet approximation for microsatellite profiles. As shown in Figure 6, the Dirichlet-based estimator is less biased, but still tends to underestimate match probabilities for more common genotypes. However, as shown in Figure 5, such underestimation may be avoided by setting the coancestry coefficient to be very high. The price for such corrections is overly conservative predictions for rarer genotypes. For example, in the simulations reflecting D8S7911 in New Zealand, some predicted match probabilities were more than three times the empirical value.

It is clear that allelic associations must be taken into consideration when estimating match probabilities for microsatellite profiles. However, as shown in Figures 3 and 4, these associations are inadequately characterized by the coancestry coefficient. Estimation procedures formulated under the source-invariant mutation model will therefore be ineffective. One alternative suggested by the current study is a coalescent-based estimator. For a given microsatellite locus, available data from well-characterized populations could be used to estimate the appropriate mutation parameters. Fu and Chakraborty (1998) describe one such analysis. Estimated parameters could then be used to evaluate match probabilities empirically, in conjunction with a variety of plausible demographic histories for the population of the suspect. However, for a given mutation model, such a procedure would only be as good as the parameter estimates. The statistical properties of estimators of mutation parameters are uncertain and require further investigation.

We thank Dr. John Buckleton and the Institute of Environmental Science and Research Limited of New Zealand for the data, and Dr. Ian Painter for helpful discussions. This work was supported in part by National Institutes of Health grant GM-45344 to North Carolina State University and by National Science Foundation grants DMS-9208758, DMS-9700867, and DMS-9711365 to the National Institute of Statistical Sciences, and in part by a postdoctoral fellowship from the New Zealand Foundation for Research in Science and Technology (FORST) to J.M.C.

LITERATURE CITED

Balding, D. J., and R. A. Nichols, 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 R. A. Nichols, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, pp. 3–12 in *Human Identification: The Use of DNA Markers*, edited by B. S. Weir. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Balding, D. J., and R. A. Nichols, 1997 Significant genetic correlations among Caucasians at forensic DNA loci. Heredity **78:** 583–589.

Brinkmann, B., M. Klintschar, F. Neuhuber, J. Hühne and B. Rolf, 1998 Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. **62:** 1408–1415.

Cavalli-Sforza, L., P. Menozzi and A. Piazza, 1994 *The History and Geography of Human Genes*. University Press, Princeton, NJ.

DiRienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin *et al.*, 1994 Mutational processes of simple sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

Evett, I., and B. Weir, 1998 *Interpreting DNA Evidence*. Sinauer Associates, Sunderland, MA.

Fu, Y.-X., and R. Chakraborty, 1998 Simultaneous estimation of all the parameters of a stepwise mutation model. Genetics **150:** 487–497.

Griffiths, R. C., 1979 A transition density expansion for a multi-allele diffusion model. Adv. Appl. Prob. **11:** 310–325.

Gyapay, J., J. Morisette, A. Vignal, C. Dib, C. Fizames *et al.*, 1994 The 1993–94 Généthon human genetic linkage map. Nat. Genet. **7:** 246–249.

Harpending, H. C., M. A. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers *et al.*, 1998 Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95:** 1961–1967.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

Lange, K., 1995 Applications of the Dirichlet distribution to forensic match probabilities, pp. 107–117 in *Human Identification: The Use of DNA Markers*, edited by B. S. Weir. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Levinson, G., and G. A. Gutman, 1987 Slipped strand mispriming: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

Li, W. H., 1976 Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. Theor. Popul. Biol. **10:** 303–308.

Li, Y.-J., 1996 *Characterizing the Structure of Genetic Populations*. Ph. D. thesis, North Carolina State University, Raleigh, NC.

Moran, P. A. P., 1975 Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. **8:** 318–330.

National Research council, 1992 *DNA Technology in Forensic Science*. National Academy Press, Washington, DC.

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

Weber, J., and C. Wong, 1993 Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123–1128.

Wehrhahn, C., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics **80:** 375–394.

Weir, B. S., 1994 The effects of inbreeding on forensic calculations. Annu. Rev. Genet. **28:** 597–621.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Weir, B. S., and C. Cockerham, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

Weir, B. S., A. H. D. Brown and D. R. Marshall, 1976 Testing for selective neutrality of electrophoretically detectable protein polymorphisms. Genetics **84:** 639–659.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Communicating editor: A. G. Clark