# Measuring Gametic Disequilibrium From Multilocus Data

**Karen L. Ayres and David J. Balding**

*Department of Applied Statistics, University of Reading, Reading RG6 6FN, United Kingdom*

Manuscript received June 20, 1999
Accepted for publication September 19, 2000

ABSTRACT

We describe a Bayesian approach to analyzing multilocus genotype or haplotype data to assess departures from gametic (linkage) equilibrium. Our approach employs a Markov chain Monte Carlo (MCMC) algorithm to approximate the posterior probability distributions of disequilibrium parameters. The distributions are computed exactly in some simple settings. Among other advantages, posterior distributions can be presented visually, which allows the uncertainties in parameter estimates to be readily assessed. In addition, background knowledge can be incorporated, where available, to improve the precision of inferences. The method is illustrated by application to previously published datasets; implications for multilocus forensic match probabilities and for simple association-based gene mapping are also discussed.

DEPARTURES from gametic (or linkage) and Hardy-Weinberg (HW) equilibria can provide clues about aspects of population histories and mating behavior (see, *e.g.*, Lewontin 1974) and can be useful in locating disease genes (Jorde 1995; Feder *et al.* 1996; Nielsen *et al.* 1998). They also play an important role in the forensic use of DNA profile evidence. Match probability calculations either rely on assumptions of equilibrium (National Research Council 1996) or else allow for patterns of departures that hold in simplified population models (Weir 1994; Balding and Nichols 1995; Ayres and Overall 1999). It is important that the validity of such assumptions in actual populations is verified empirically, as far as is feasible.

Traditional statistical treatments usually focus on testing hypotheses of equilibrium, with recent developments involving randomization tests (*e.g.*, Zaykin *et al.* 1995; Slatkin and Excoffier 1996). Although they may form a useful first step, such hypothesis tests represent a limited form of statistical inference, since the tests concern only whether or not the data are consistent with equilibrium, rather than directly assessing how large are the departures from precise equilibrium that inevitably exist in real populations (*e.g.*, Smith 1970). In forensic applications, for example, a hypothesis of equilibrium may be rejected with a sufficiently large sample, whereas a forensic scientist may nevertheless believe that the magnitude of the departure is sufficiently small that the hypothesis of equilibrium, though strictly false, is adequate for the application at hand.

Point estimation methods for disequilibrium parameters have been developed (see, *e.g.*, Weir 1979). Here, we propose a Markov chain Monte Carlo (MCMC) method to investigate probability distributions for gametic disequilibrium measures given the data. This extends previous work (Ayres and Balding 1998; Shoemaker *et al.* 1998) on assessing departures from HW.

Perhaps the most important advantage of our approach is *interpretability*: the questions of interest are answered directly in terms of probabilities that can conveniently be presented graphically via probability density curves, providing an immediate yet detailed assessment of the variability associated with an estimate. A further advantage is that, since the approach is likelihood based, it is statistically powerful and can incorporate a wide range of modeling assumptions. Previous treatments assume random union of gametes (RUG) to infer population haplotype proportions from genotype data (*e.g.*, Excoffier and Slatkin 1995). Although we also implement the RUG model in the following analyses, we note that other models can be readily applied, such as those that incorporate inbreeding measures.

The choice of prior distribution is sometimes seen as a barrier to the implementation of direct probability, or Bayesian, methods. We introduce a class of hierarchical prior distributions for the haplotype proportions, which allows the scientist some flexibility either to incorporate relevant background information, if desired, or to adopt a relatively "vague" prior.

We illustrate our method by analyzing samples of genotypes at two unlinked loci and at three linked loci. We also briefly discuss its application to forensic identification, and to haplotype data and simple disequilibrium gene mapping. Computer programs (C code) for the MCMC algorithms are available from the authors on request.

*Corresponding author:* David J. Balding, Department of Applied Statistics, University of Reading, P.O. Box 240, Earley Gate, Reading RG6 6FN, United Kingdom. E-mail: d.j.balding@rdg.ac.uk

## METHODS

**Measures of gametic disequilibrium:** Genetic equilibrium corresponds to statistical independence, and many authors (see, *e.g.*, Weir 1979) measure gametic disequilibrium in terms of the differences between population haplotype proportions and the values that would be expected under equilibrium, given the allele proportions. Following this approach, for a two-locus haplotype consisting of alleles $A_i$ and $B_j$, we introduce the notation

$$D_{ij} = h_{ij} - p_i q_j, \tag{1}$$

where $h_{ij}$ denotes the population proportion of haplotype $A_iB_j$, while $p_i = \Sigma_j h_{ij}$ and $q_j = \Sigma_i h_{ij}$, the proportions of, respectively, alleles $A_i$ and $B_j$.

The range of $D_{ij}$ depends on $p_i$ and $q_j$, which makes cross-locus and cross-population comparisons difficult. To alleviate this problem, Lewontin (1964) defined the normalized difference, with range $[-1, 1]$, by

$$D'_{ij} = D_{ij}/D_{\max},$$

where $D_{\max}$ is

$$\begin{array}{ll} \min(p_iq_j,\ (1-p_i)(1-q_j)) & \text{if } D_{ij} < 0 \\ \min(p_i(1-q_j),\ (1-p_i)q_j) & \text{if } D_{ij} > 0. \end{array}$$

When there are only two alleles at each locus, there is a unique value of $|D'_{ij}|$. Otherwise, it is usually of interest to have a summary measure of the gametic disequilibrium between the two loci; Hedrick (1987) proposed

$$D' = \sum_i\sum_j p_i q_j |D'_{ij}|. \tag{2}$$

The range of $D'$ is $[0, 1]$, independent of the $p_i$ and $q_j$. However, there remain difficulties in interpreting the value of $D'$. Lewontin (1988) noted that values of $D'$ at different loci and in different populations tend to vary with the values of the $p_i$ and $q_j$, so that the problem of cross-locus and cross-population comparisons is not fully overcome by use of $D'$.

Moreover, in practice the range of values of $D'$ consistent with gametic equilibrium is not readily apparent and can vary from locus to locus. Under equilibrium, each $D_{ij}$, and hence $D'$, takes value zero. However, just as a $\chi^2$ goodness-of-fit statistic is unlikely to be very close to zero even when the model is valid, so estimates of $D'$ based on data from equilibrium populations are unlikely to be very close to zero (furthermore, variances of $D'$ are difficult to calculate; see Zapata *et al.* 1997). Insight into whether or not the data are consistent with gametic equilibrium can be gained by reanalyzing them with the alleles at each locus randomly permuted, thus simulating samples from a population in equlibrium with the same $p_i$ and $q_j$ as the population under study (*e.g.*, Slatkin and Excoffier 1996). However, if the primary aim of an analysis is to test the hypothesis of equilibrium, then the likelihood-ratio (LR) statistic or a Bayes factor would be more appropriate than $D'$.

Measures of gametic disequilibrium not based on $D_{ij}$ have also been proposed. Smouse (1974) specifies a log-linear model for the $h_{ij}$, with allele-specific parameters $a_i$ and $b_j$, and an interaction term $c_{ij}$ that can be employed as an alternative to $D_{ij}$. Weir (1996, pp. 127–133) details a closely related multiplicative model and extends the analysis to genotype data.

Here, we focus on $D'$ as a summary measure of gametic disequilibrium (together with an extension $D''$, introduced below). This measure is widely used and, although it suffers from the interpretability drawbacks described above, there seems to be no univariate measure that avoids such difficulties. When interest focuses on gametic disequilibrium due to linkage, such as in "simple" genetic mapping, then a natural criterion for choosing between disequilibrium measures is correlation with physical distance and Devlin and Risch (1995) find that $D'$ has good properties in that setting.

**Random union of gametes model:** When only genotype counts are available, a model is required to relate the $h_{ij}$ to genotype proportions, which then implies a model for the $D_{ij}$. For two loci at which the population proportion of genotype $A_iA_{i'}B_jB_{j'}$ is denoted $p_{ii'jj'}$ (with $i \le i'$ and $j \le j'$), perhaps the simplest plausible model assumes RUG:

$$p_{ii'jj'} = \begin{cases} h_{ij}^2 & \text{if } i = i', j = j' \\ 2h_{ij}h_{ij'} & \text{if } i = i', j < j' \\ 2h_{ij}h_{i'j} & \text{if } i < i', j = j' \\ 2h_{ij}h_{i'j'} + 2h_{i'j}h_{ij'} & \text{if } i < i', j < j'. \end{cases} \tag{3}$$

Inbreeding and selection, for example, will invalidate this model: haplotype proportions will be incorrectly estimated because no allowance is made for the dependence of haplotypes within multi locus genotypes. However, for human populations and approximately neutral loci, the effect on inference should be negligible, and so the RUG assumption may be reasonable in such cases.

The log-likelihood for a random sample of genotypes is obtained by substituting (3) into the multinomial log-likelihood function,

$$\log L = \sum n_{ii'jj'} \log(p_{ii'jj'}), \tag{4}$$

where the $n_{ii'jj'}$ are the observed genotype counts. The maximum-likelihood (ML) estimates $\hat{h}_{ij}$ can then be obtained by maximizing log $L$ using any suitable method, such as the expectation-maximization (EM) algorithm of Excoffier and Slatkin (1995). Substitution of the $\hat{h}_{ij}$ into (1) and (2) then leads to point estimates $\hat{D}'_{ij}$ and $\hat{D}'$.

**Modeling background information:** Here, we are primarily concerned not with point estimates but with the full joint distribution of the $h_{ij}$, and hence of the $d_{ij}$ and $D'$, given the genotype data. This requires a probability model for the $h_{ij}$ prior to observing the data. Perhaps the simplest such model is given by the (multivariate) uniform distribution, which may be interpreted as corresponding to no background information about haplotype proportions. However, a uniform prior for the $h_{ij}$ does not correspond to an uninformative prior for $D'$, and the level of informativeness is fixed and cannot be controlled. Moreover, the uniform-on-haplotypes prior does not encapsulate the fact that haplotypes are composed of alleles and hence, for example, the $h_{1j}$, $j \neq 1$ and the $h_{i1}$, $i \neq 1$ are informative about $p_1$ and $q_1$ and thus may well be informative about $h_{11}$.

Suppose that information was available in advance, perhaps from surveys in other populations, which indicated that $p_i$ and $q_j$ were likely to be close to, say, $\alpha_i$ and $\beta_j$, respectively. (Conceptually, $\alpha_i$ and $\beta_j$ might be thought of as metapopulation allele proportions.) A tractable family of prior distributions for the $h_{ij}$ would then be the Dirichlet family with parameters $\lambda\alpha_i\beta_j$, where $\lambda$ is a constant, so that each $h_{ij}$ has prior expectation and variance given by

$$E[h_{ij}] = \alpha_i\beta_j, \quad \mathrm{Var}[h_{ij}] = \frac{\alpha_i\beta_j(1 - \alpha_i\beta_j)}{1 + \lambda}. \qquad (5)$$

Under this assumption, the $p_i$ and the $q_j$ are also Dirichlet, with parameters $\lambda\alpha_i$ and $\lambda\beta_j$, respectively. If $\lambda$ is large then $h_{ij}$, $p_i$, and $q_j$ will be close to, respectively, $\alpha_i\beta_j$, $\alpha_i$,

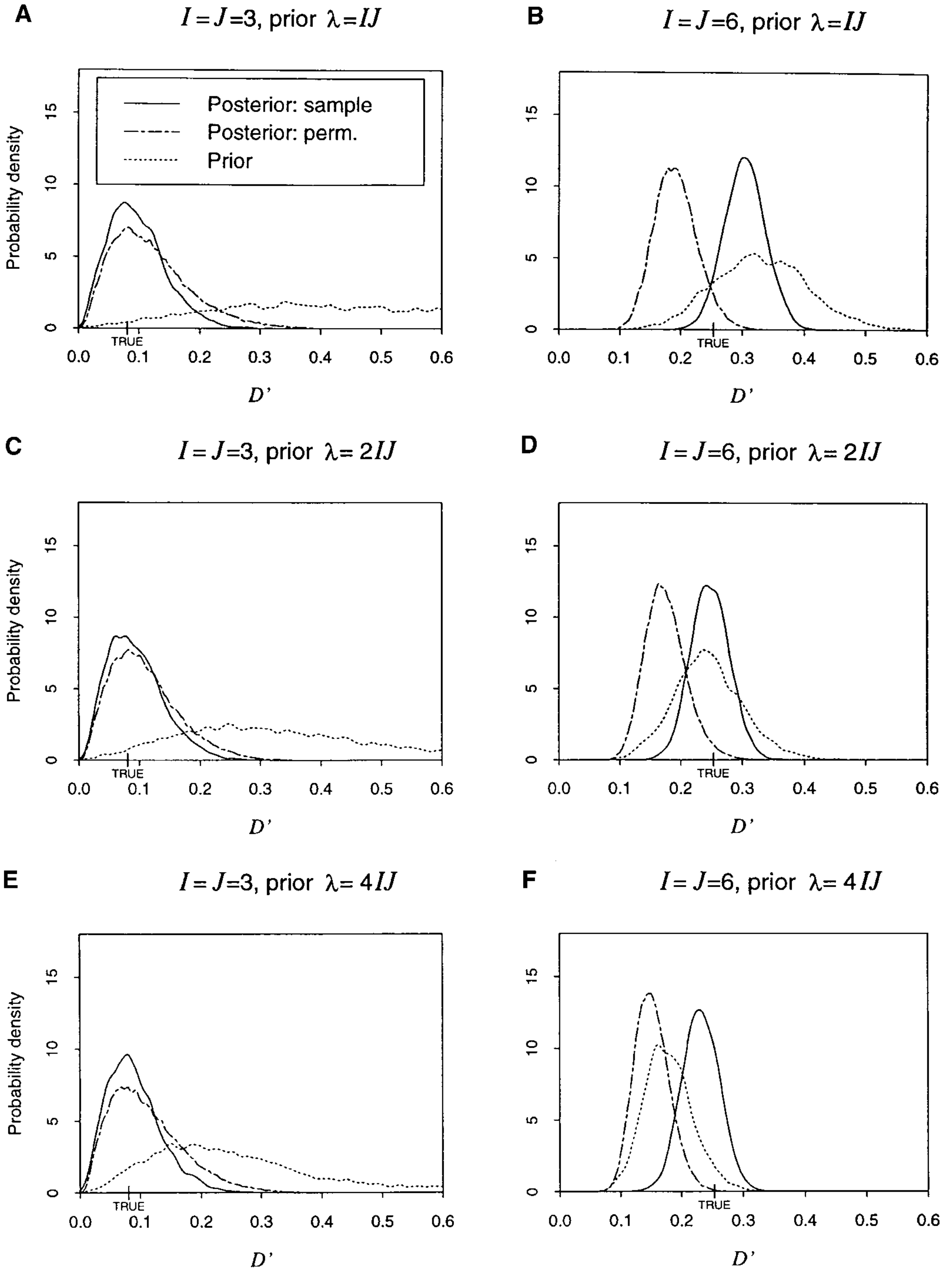

FIGURE 1.—Prior (dotted curves) and posterior (solid curves) densities for the overall disequilibrium measure $D'$, based on two data sets simulated under the RUG model, each with $n$ = 200 genotypes and with three and six alleles at each locus. The "true" values of $D'$ underlying the simulated data are 0.081 and 0.253. The posterior curves are obtained by Gaussian kernel density estimation using 5000 MCMC algorithm outputs. The three prior densities for the $h_{ij}$ are Dirichlet with parameters $IJ\alpha_i\beta_j$ (A and B), $2IJ\alpha_i\beta_j$ (C and D), and $4IJ\alpha_i\beta_j$ (E and F), conditional on the $\alpha_i$ and the $\beta_j$, which are each (multivariate) uniform. Dot-dashed curves show average posterior densities from 50 random permutations of the genotypes at each locus, mimicking 50 samples from populations in gametic equilibrium.

**TABLE 1**

**Medians and equal-tailed 90% intervals of the prior and posterior distribution for $D'$ shown in Figure 1**

| No. alleles | Prior | Prior | | Posterior | |
|---|---|---|---|---|---|
| true $D'$ | parameters | Median | 90% interval | Median | 90% interval |
| $I = J = 3$ | $\lambda = IJ = 9$ | 0.453 | (0.132–0.952) | 0.090 | (0.030–0.182) |
| | $\lambda = 2IJ = 18$ | 0.327 | (0.102–0.825) | 0.088 | (0.031–0.180) |
| $D' = 0.081$ | $\lambda = 4IJ = 36$ | 0.236 | (0.077–0.622) | 0.084 | (0.028–0.172) |
| $I = J = 6$ | $\lambda = IJ = 36$ | 0.329 | (0.211–0.463) | 0.305 | (0.250–0.359) |
| | $\lambda = 2IJ = 72$ | 0.242 | (0.155–0.343) | 0.247 | (0.198–0.299) |
| $D' = 0.253$ | $\lambda = 4IJ = 144$ | 0.176 | (0.119–0.248) | 0.231 | (0.182–0.281) |

and $\beta_j$, and hence the implied prior for $D_{ij}$ will be peaked at zero, implying little gametic disequilibrium. Decreasing the value of $\lambda$ makes strong disequilibrium more probable (the tails of the implied prior distribution for $D_{ij}$ are longer).

The sum of the Dirichlet parameters provides a measure of the information conveyed by the distribution. Choosing $\lambda$ so that the average of the $\lambda\alpha_i\beta_j$ is one would give a distribution that has the same information content as the uniform (for which all the parameters equal one) and may provide a reasonable vague prior for the $h_{ij}$.

This framework for specifying a prior distribution for $h_{ij}$ does not require that $\alpha_i$ and $\beta_j$ be specified precisely. Instead, they can be assigned probability distributions, leading to a hierarchical prior model. Below, we adopt independent uniform distributions for the $\alpha_i$ and $\beta_j$, although background information could in practice be incorporated into more informative distributions.

**MCMC algorithm:** We implement an MCMC stochastic simulation algorithm for genotype data to approximate the joint distribution of the $h_{ij}$, and hence of the gametic disequilibrium measures, under the RUG model and the hierarchical prior distribution described above. The MCMC algorithm adopted is of the Metropolis-Hastings type (Metropolis *et al.* 1953; Hastings 1970). At each iteration of the algorithm, a decision is made whether to keep the current vector of parameter values or reject it in favor of a new vector. The accept/reject decision is made in such a way that the proportion of iterations at which the current vector lies in any region of the parameter space approximates the probability that the true parameter vector lies in that region, with the approximation becoming more accurate as the number of iterations increases. Further details of the MCMC algorithm are given in the appendix.

Figure 1 shows the posterior density curves for $D'$, approximated via the MCMC algorithm, given two samples of two-locus genotypes simulated under the RUG model with $D' = 0.081$ (three alleles) and $D' = 0.253$ (six alleles). Three prior distributions were employed, shown as dotted curves. Key quantiles of the prior and posterior distributions are given in Table 1.

Even with a reasonably large sample size (200 individuals), $D'$ is a difficult parameter to estimate. This is because the data bear directly on the population genotype proportions, whereas differences between allele and haplotype proportions are the quantities of interest. This difficulty is reflected by the posterior curves of

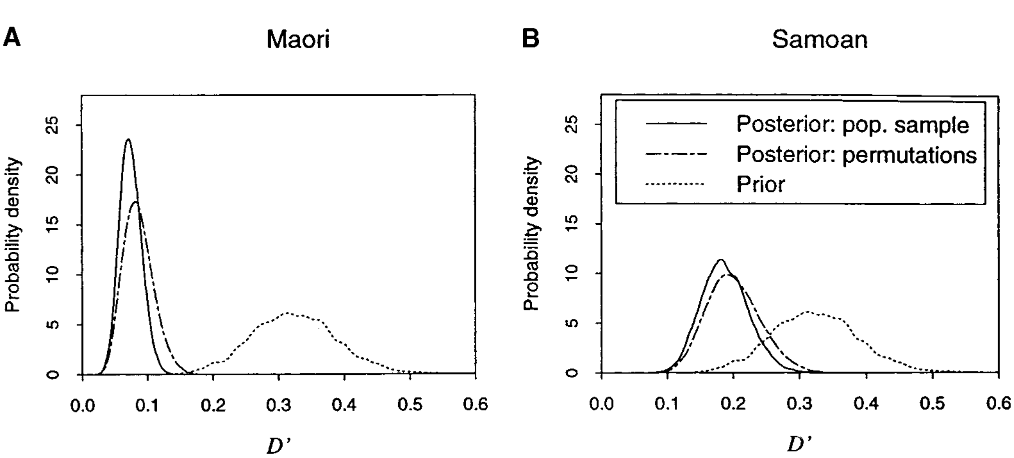

Figure 2.—Prior (dotted curves, corresponding to the prior distribution in Figure 1 with $\lambda = IJ$, where $I = 8$ and $J = 6$) and posterior (solid curves) densities for the overall disequilibrium measure $D'$ for samples of Maoris (1091 individuals) and Samoans (139 individuals) at STR loci THO1 and TPOX. Allele labels are numbers of repeats. The dot-dashed curves show posterior densities averaged over 50 random permutations of the genotypes. Data provided by John Buckleton (Institute of Environmental Science and Research, New Zealand).

Figure 1, which support a rather broad range of values for $D'$ and display some sensitivity to the choice of prior. However, in each case the posterior median is close to the true value and usually closer than the corresponding ML-based estimates (0.058 and 0.315), for which the sampling variance is difficult to calculate. Moreover, since $D'$ is univariate it is relatively easy to plot both prior and posterior density curves and hence assess visually the effect of the prior from the plots. Background information, when available, can be incorporated via the prior and may be invaluable in situations of little data and/or many alleles.

Also shown in Figure 1 are density curves averaged over 50 random permutations of the alleles, mimicking 50 samples from populations in gametic equilibrium with the same allele proportions. The data with three alleles at each locus are clearly consistent with equilibrium, but those with six alleles are not. These results are in accord with the $P$ values 0.56 and 0.00 obtained from an LR-based permutation test for gametic disequilibrium (Slatkin and Excoffier 1996).

Figure 1 corresponds to a single simulated dataset. We also applied the MCMC method (for the prior with $\lambda = IJ$) to 100 datasets of size $n = 1000$, simulated with $I = J = 3$. The underlying $h_{ij}$ were such that $D' = 0.404$. For each dataset we calculated the posterior median: the 100 estimated posterior medians had mean 0.403 and standard deviation 0.039. These values compare favorably with the MLE-based estimates $\hat{D}'$, for which the mean and standard deviation over these 100 simulated datasets were 0.407 and 0.040.

## RESULTS

**Two unlinked loci used for forensic identification:** The MCMC method was applied to the genotypes at two unlinked forensic short tandem repeat (STR) loci, THO1 and TPOX, for samples of Maoris ($n = 1091$) and Samoans ($n = 139$) resident in New Zealand. Eight alleles were observed for locus THO1 and six for TPOX (additional alleles observed in other populations are ignored here, although they could be incorporated into the analysis if desired).

Figure 2 shows prior ($\lambda = IJ$) and posterior curves for the overall measure $D'$ together with a curve obtained from 50 random permutations of the data (mimicking equilibrium). There is a substantial overlap of these curves, suggesting that both samples are consistent with gametic equilibrium in the underlying populations; these conclusions are in agreement with $P$ values obtained from the LR-based permutation test (0.42 and 0.13).

A full multilocus match probability involves correlations of genes both within and between individuals (in

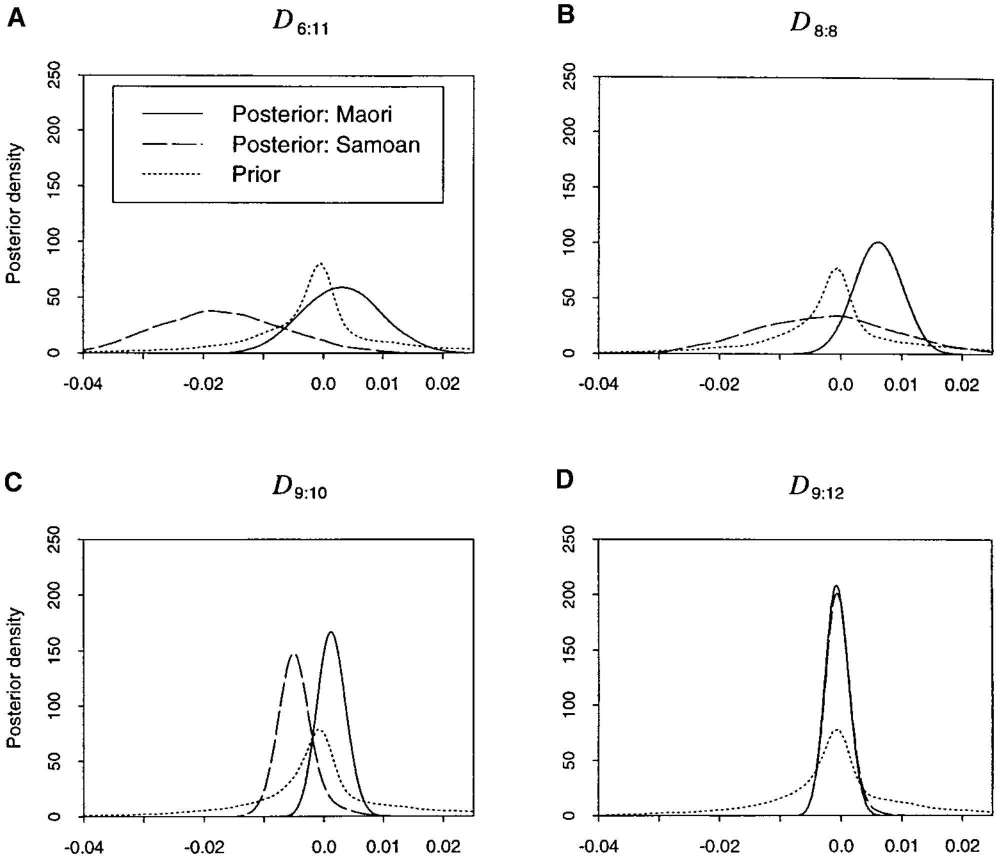

Figure 3.—Posterior densities for disequilibrium coefficients $D_{6:11}$, $D_{8:8}$, $D_{9:10}$, and $D_{9:12}$ for the Maori and Samoan samples of Figure 2. The prior density (dotted curve) is the same as that underlying Figure 2.

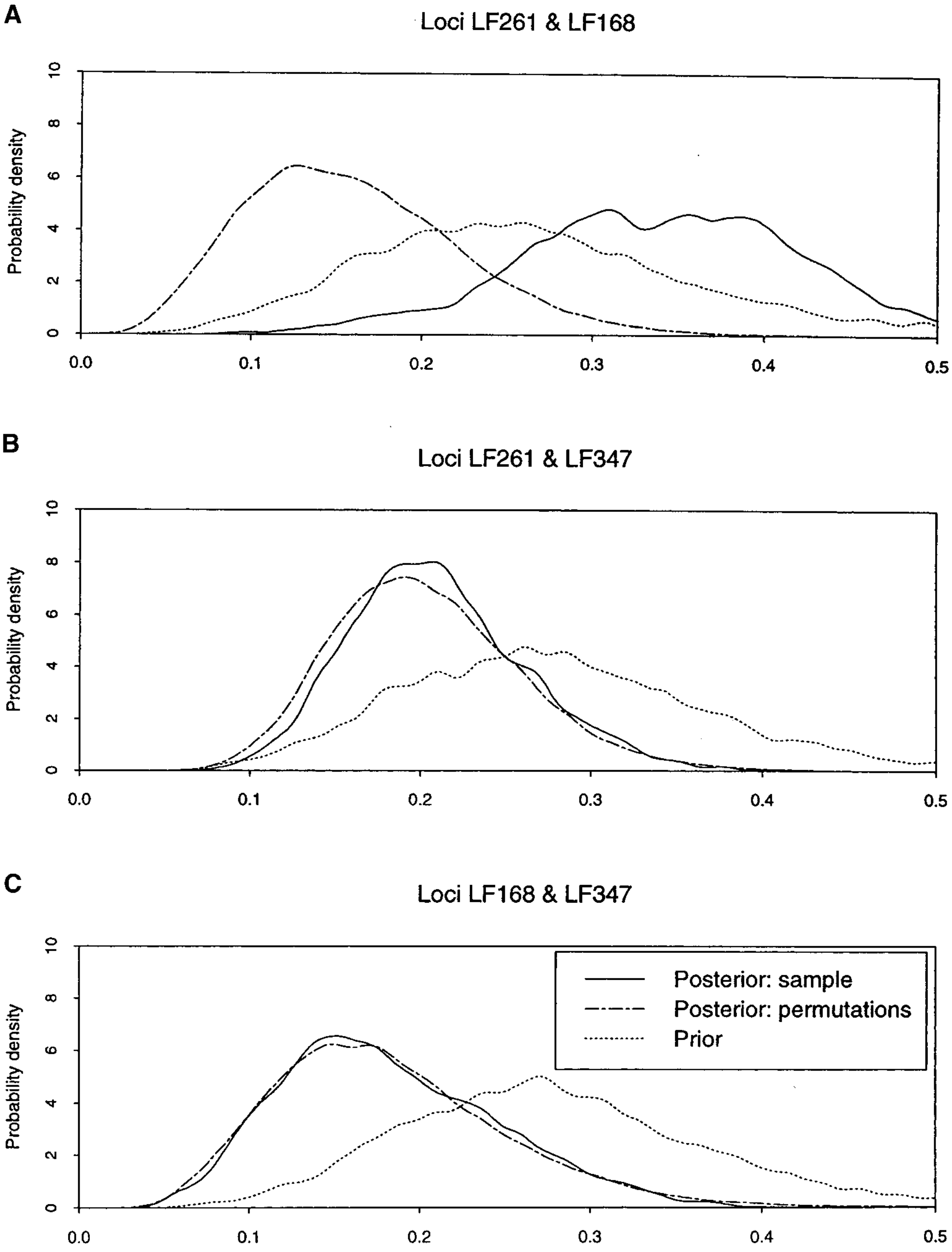

FIGURE 4.—Posterior densities (solid curves) for pairwise disequilibrium coefficients $D'$ from a sample of $n = 96$ genotypes, at loci LF261 (four alleles), LF168 (four alleles), and LF347 (five alleles) of the MOYO strains of the *A. aegypti* mosquito data of YAN *et al.* (1997). Point estimates (from ML estimates) are 0.492, 0.220, and 0.242. Dot-dashed curves are posterior densities averaged over 50 random permutations of the genotypes. The prior densities (dotted curves) are based on a Dirichlet prior for the $h_{ij}$, with parameters $40\alpha_i\beta_j\gamma_k$, conditional on the $\alpha_i$, $\beta_j$, and $\gamma_k$, which are each (multivariate) uniform.

the latter case, between the defendant and an alternative possible source of the crime scene DNA). In current practice (see, *e.g.*, EVETT and WEIR 1998), adjustment is often made for between-individual correlations at a single locus. However, multilocus forensic match probabilities are usually obtained by taking the product of the single-locus probabilities, thereby assuming independence between loci. Strong gametic disequilibrium may invalidate this assumption.

Although THO1 and TPOX are unlinked, gametic disequilibrium may nevertheless arise (due to founder effects, selection, or drift) and affect multilocus forensic match probability calculations involving these loci. It is therefore important to investigate levels of gametic disequilibrium, and a selection of marginal posterior density curves for the $D_{ij}$ is shown in Figure 3. All the posterior distributions support values close to zero, encouraging optimism that the effect of gametic disequilibrium on two-locus forensic match probabilities involving these loci may indeed be negligible.

Although these results tend to support current practice, note that we have not simultaneously taken all relevant correlations into account. In particular, other forms of assocation may invalidate the independence of genes assumption in the match probability (see AYRES 2000 for the case of between-locus dependence due to

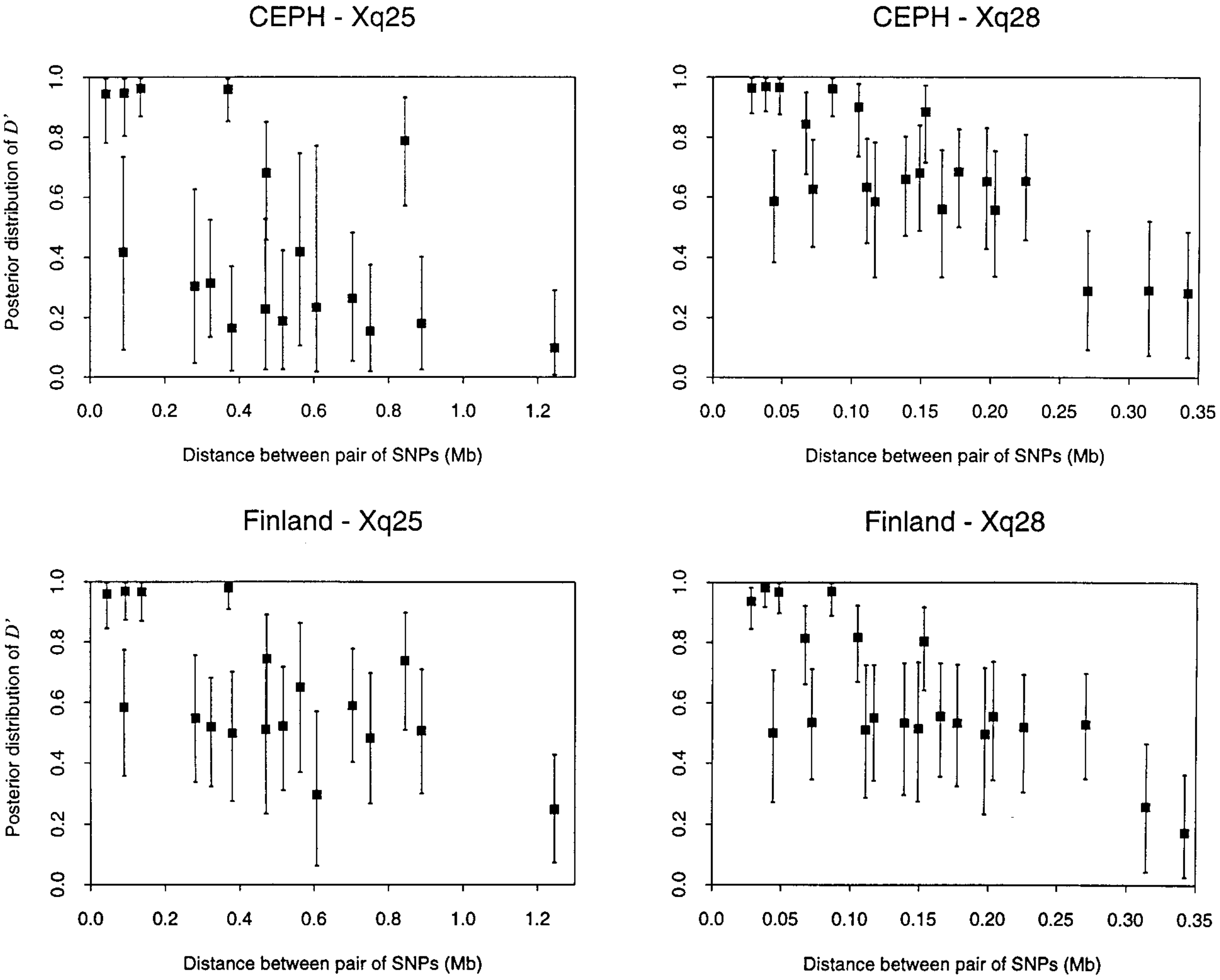

Figure 5.—Posterior median (■) and central 90% posterior intervals (—) for $D'$ in two populations, for the Xq25–Xq28 SNP data of Taillon-Miller *et al.* (2000); marker pairs analyzed here correspond to those presented in Table 2 of Taillon-Miller *et al.* (2000). A multivariate uniform prior was used for the $h_{ij}$.

inbreeding). Also, forensic identification involves many loci, often >10, whereas here we have considered only the two-locus case.

**Three linked loci:** The MCMC algorithm for approximating the joint posterior of the haplotype proportions is readily extended to three loci. For forensic applications, we may be interested in investigating the difference $h_{ijk} - p_i q_j r_k$, which can be readily obtained from the MCMC output. For other problems, simultaneous estimation of the pairwise disequilibrium measures may be of more interest. However, multilocus systems impose additional constraints on the $D_{ij}$. For three diallelic loci, Robinson *et al.* (1991) describe a new pairwise normalized measure $D''_{ij}$ based on these adjusted bounds and with range [−1, 1]. These authors also note that additional loci beyond three add no further constraints on the $D_{ij}$. The multiallelic analogue of the formulas of Robinson *et al.* (1991) is given in the appendix.

Yan *et al.* (1997) analyzed multilocus genotype data from chromosome 3 of the MOYO strain of the mosquito *Aedes aegypti*. For the three RFLP loci LF261, LF168, and LF347, they reported that the values of $D'$ for all three pairs were significant at the 1% level. We have calculated posterior density curves for $D''$ on the basis of their original data, as well as curves based on random permutations (Figure 4).

Our results suggest strong disequilibrium between LF 261 and LF168, since the curve based on the randomly permuted data has little overlap with that based on the observed data. In contrast, Figure 4 suggests little or no disequilibrium between the other two pairs of loci. The latter conclusion differs from that of Yan *et al.* (1997). It is, however, consistent with the relative map positions of the loci—LF168 is situated between LF261 and LF347, much closer to LF261 than to LF347. Although the MCMC results correctly identify LF168 and LF261 as the closest pair, disequilibrium between the other marker pairs is so weak that the ordering of all three

**TABLE 2**

**Posterior summaries of $D'$ for the analyses of Figure 5**

| Markers | Distance (Mb) | $D'_{\text{CEPH}}$ median | $D'_{\text{Fin}}$ median | Posterior Pr$(D'_{\text{CEPH}} < D'_{\text{Fin}})$ |
|---|---|---|---|---|
| Xq4007-1, Xq3774-2 | 0.044 | 0.94 | 0.96 | 0.59 |
| Xq3804-1, Xq3812-1 | 0.090 | 0.42 | 0.58 | 0.75 |
| Xq3774-2, Xq3773-1 | 0.093 | 0.95 | 0.97 | 0.59 |
| Xq4007-1, Xq3773-1 | 0.137 | 0.96 | 0.97 | 0.50 |
| Xq3812-1, Xq3917-1 | 0.281 | 0.30 | 0.55 | 0.82 |
| Xq3862-1, Xq3773-1 | 0.323 | 0.31 | 0.52 | 0.89 |
| Xq3804-1, Xq3917-1 | 0.371 | 0.96 | 0.98 | 0.68 |
| Xq3773-1, Xq3804-1 | 0.380 | 0.16 | 0.50 | 0.97 |
| Xq3773-1, Xq3812-1 | 0.470 | 0.23 | 0.51 | 0.88 |
| Xq3774-2, Xq3804-1 | 0.473 | 0.68 | 0.74 | 0.66 |
| Xq4007-1, Xq3804-1 | 0.517 | 0.19 | 0.52 | 0.97 |
| Xq3774-2, Xq3812-1 | 0.563 | 0.42 | 0.65 | 0.79 |
| Xq4007-1, Xq3812-1 | 0.607 | 0.23 | 0.30 | 0.58 |
| Xq3862-1, Xq3804-1 | 0.703 | 0.26 | 0.59 | 0.96 |
| Xq3773-1, Xq3917-1 | 0.751 | 0.15 | 0.48 | 0.95 |
| Xq3774-2, Xq3917-1 | 0.844 | 0.79 | 0.74 | 0.34 |
| Xq4007-1, Xq3917-1 | 0.888 | 0.18 | 0.51 | 0.96 |
| Xq3774-2, Xq3698-1 | 1.245 | 0.10 | 0.25 | 0.83 |
| Xq2816-1, Xq3274-1 | 0.028 | 0.96 | 0.94 | 0.28 |
| Xq3471-1, Xq4001-1 | 0.038 | 0.97 | 0.98 | 0.65 |
| Xq3413-1, Xq2816-1 | 0.044 | 0.59 | 0.50 | 0.33 |
| Xq3449-1, Xq3471-1 | 0.048 | 0.97 | 0.97 | 0.53 |
| Xq4001-1, Xq3413-1 | 0.067 | 0.84 | 0.81 | 0.41 |
| Xq3413-1, Xq3274-1 | 0.072 | 0.63 | 0.53 | 0.29 |
| Xq3449-1, Xq4001-1 | 0.086 | 0.96 | 0.97 | 0.56 |
| Xq3471-1, Xq3413-1 | 0.105 | 0.90 | 0.82 | 0.24 |
| Xq4001-1, Xq2816-1 | 0.111 | 0.63 | 0.51 | 0.24 |
| Xq3476-1, Xq3449-1 | 0.117 | 0.59 | 0.55 | 0.45 |
| Xq4001-1, Xq3274-1 | 0.139 | 0.66 | 0.53 | 0.25 |
| Xq3471-1, Xq2816-1 | 0.149 | 0.68 | 0.52 | 0.19 |
| Xq3449-1, Xq3413-1 | 0.153 | 0.89 | 0.80 | 0.24 |
| Xq3476-1, Xq3471-1 | 0.165 | 0.56 | 0.56 | 0.50 |
| Xq3471-1, Xq3274-1 | 0.177 | 0.69 | 0.53 | 0.21 |
| Xq3449-1, Xq2816-1 | 0.197 | 0.65 | 0.50 | 0.18 |
| Xq3476-1, Xq4001-1 | 0.203 | 0.56 | 0.56 | 0.51 |
| Xq3449-1, Xq3274-1 | 0.225 | 0.65 | 0.52 | 0.22 |
| Xq3476-1, Xq3413-1 | 0.270 | 0.29 | 0.53 | 0.94 |
| Xq3476-1, Xq2816-1 | 0.314 | 0.29 | 0.26 | 0.43 |
| Xq3476-1, Xq3274-1 | 0.342 | 0.28 | 0.17 | 0.28 |

Posterior medians (based on 1000 sampled values) for $D'$ in two populations (CEPH, $n = 92$; Finland (Fin), $n = 100$), for the Xq25–Xq28 SNP data of Taillon-Miller *et al.* (2000); also shown is the probability that $D'$ in the CEPH population is less than $D'$ in the Finnish population. Marker pairs analyzed here correspond to those presented in Table 2 of Taillon-Miller *et al.* (2000) and are ordered by distance between the markers. A multivariate uniform prior was used for the $h_{ij}$. Each entry is based on 1000 values sampled from the posterior distribution of $D'$.

markers on the basis of the joint posterior distribution of the $D''$ cannot be achieved with confidence, the correct order being assigned a probability of 42%.

**Haplotype data:** In some cases haplotype counts may be available, simplifying the direct probability approach. For example, for three-locus haplotypes $h_{ijk}$ and a hierarchical prior, we have

$$p(\{h_{ijk}\}, \alpha, \beta, \gamma|\mathbf{N}) = p(\alpha, \beta, \gamma|\mathbf{N})p(\{h_{ijk}\}|\alpha, \beta, \gamma, \mathbf{N}), \qquad (6)$$

where $\mathbf{N} \equiv \{n_{ijk}\}$ denotes the sample haplotype counts and $\alpha$, $\beta$, and $\gamma$ are vectors of hyperparameters specifying prior distributions for the population allele proportions at the three loci.

A method for sampling from this distribution is given in the appendix, together with a summary of implications for disequilibrium mapping. However, we focus below on the simpler case when, for two loci, a straightforward (nonhierarchical) Dirichlet distribution with parameters $k_{ij}$ is implemented for the $h_{ij}$. When the likelihood is multinomial, the posterior distribution for the $h_{ij}$ will again be Dirichlet with parameters $n_{ij} + k_{ij}$, and a sample from this distribution can be obtained by standard random number generation (see, *e.g.*, Appendix A of Gelman *et al.* 1995).

Taillon-Miller *et al.* (2000) analyzed several pairs of single nucleotide polymorphism (SNP) markers in the human Xq25–Xq28 region for three populations [general European (CEPH), Finnish, and Sardinian]. They found significant $P$ values ($P < 0.001$, from a $\chi^2$ test for gametic equilibrium) for markers separated by up to ~900 kb. They also found that, in general, point estimates of disequilibrium measures (such as $D'$) did not differ greatly between the large outbred population (represented by the CEPH sample) and the genetically isolated populations of Finland and Sardinia. These results were consistent with an STR analysis of similar populations (Eaves *et al.* 2000), though both conflict with the suggestion that genetically isolated populations tend to exhibit higher levels of disequilibrium and are therefore more useful for disease gene mapping (see, *e.g.*, Wright *et al.* 1999). In summarizing the conclusions of Taillon-Miller *et al.* (2000) and Eaves *et al.* (2000), Boehnke (2000) argues that the levels of disequilibrium observed appeared slightly stronger in the isolates than in the general mixed populations.

We have reanalyzed the CEPH and Finnish SNP data of Taillon-Miller *et al.* (2000), implementing a multivariate uniform prior for $h_{ij}$ for each pair of markers analyzed. This distribution imposes a prior belief that none of the alleles is very rare [the implied prior for the $p_i$ and $q_j$ is Beta(2,2)], which is reasonable for these markers as they have been selected on the basis of polymorphism. Results are shown in Figure 5: posterior medians and 90% intervals for the two populations are plotted against physical distance.

The measure of variability provided by our MCMC approach allows more careful comparison of the levels of disequilibrium across the populations analyzed. For almost all of the marker pairs given in Table 2, the posterior 90% intervals from the two populations overlap substantially, indicating that there is little evidence of any difference across the populations. This is quantified in Table 2, which gives the posterior probability for each marker pair that $D'$ is larger in the Finnish population than in the CEPH population: these probabilities exceed 90% for only a handful of markers, and in no case exceed 97%. (Note the values of $D'$ across closely linked markers are not independent.)

Our results therefore quantify the observation made by Taillon-Miller *et al.* (2000) that disequilibrium levels were similar across the populations. The data provide little evidence that gametic disequilibrium is higher in the Finnish population than in the general European population.

## DISCUSSION

The direct probability, or Bayesian, approach developed here permits interpretable visual answers to the question of interest about disequilibrium parameters. Moreover, it can readily incorporate complex models and background knowledge about a population, when available. For a discussion of the advantages of Bayesian approaches to problems in genetics, see Shoemaker *et al.* (1999). We have also developed a family of hierarchical prior distributions that allow the scientist some flexibility in specifying background knowledge.

Zapata *et al.* (1997) note that point estimates of $D'_{ij}$ are frequently reported without a corresponding measure of variability (such as the standard error), which can complicate comparisons over loci and populations. However, the calculation of $\mathrm{Var}(D'_{ij})$ is complicated by the different rescaling of positive and negative values in the definition of $D'_{ij}$. Zapata *et al.* derived an approximation to $\mathrm{Var}(D'_{ij})$ for biallelic loci only. Our direct probability approach provides an approximation not just for the variance of (multiallelic) $D'_{ij}$ and $D'$ but for their entire posterior distributions. A particular advantage is that the posterior intervals obtained can be directly interpreted in terms of probabilities, unlike standard confidence intervals that are routinely provided for some point estimates, which do not have such a direct interpretation.

There are no theoretical limits to the number of loci that can be analyzed simultaneously. However, for a fixed sample size, the information contained in the data decreases as the number of loci increases, and, as for hypothesis testing, useful inferences are usually not feasible for more than about three loci.

We thank the following for kindly providing the data analyzed in this study: John Buckleton (human STR data), Guiyun Yan (mosquito data), Patty Taillon-Miller, and Pui-Yan Kwok (SNP data). We thank Laurent Excoffier for helpful comments on an earlier draft, as well as two anonymous referees. Work was supported in part by the UK Biotechnology and Biological Sciences Research Council, under grant 45/G09617.

## LITERATURE CITED

Ayres, K. L., 1998 Measuring genetic correlations within and between loci, with implications for disequilibrium mapping and forensic identification. Ph.D. Thesis, The University of Reading, Reading, UK.

Ayres, K. L., 2000 A two-locus forensic match probability for subdivided populations. Genetica **108:** 137–143.

Ayres, K. L., and D. J. Balding, 1998 Measuring departures from Hardy-Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. Heredity **80:** 769–777.

Ayres, K. L., and A. D. J. Overall, 1999 Allowing for within-subpopulation inbreeding in forensic match probabilities. Forensic Sci. Int. **103:** 207–216.

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. Genetica **96:** 3–12.

Best, N. G., M. K. Cowles and S. K. Vines, 1995 *CODA Manual Version 0.30.* MRC Biostatistics Unit, Cambridge, UK.

Boehnke, M., 2000 A look at linkage disequilibrium. Nat. Genet. **25:** 246–247.

Brooks, S. P., 1998 Markov chain Monte Carlo method and its application. Statistician **47:** 69–100.

Devlin, B., and N. Risch, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics **29:** 311–322.

Eaves, I. A., T. R. Merriman, R. A. Barber, S. Nutland, E. Tuomilehto-Wolf *et al.*, 2000 The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nat. Genet. **25:** 320–323.

Evett, I. W., and B. S. Weir, 1998 *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists.* Sinauer, Sunderland, MA.

Excoffier, L., and M. Slatkin, 1995 Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. **12:** 921–927.

Feder, J. N., A. Gnirke, W. Thomas, Z. Tsuchihashi, D. A. Ruddy *et al.*, 1996 A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat. Genet. **13:** 399–408.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 1995 *Bayesian Data Analysis.* Chapman and Hall, London.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

Hedrick, P. W., 1987 Gametic disequilibrium measures: proceed with caution. Genetics **117:** 331–341.

Jorde, L. B., 1995 Linkage disequilibrium as a gene-mapping tool. Am. J. Hum. Genet. **56:** 11–14.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York.

Lewontin, R. C., 1988 On measures of gametic disequilibrium. Genetics **120:** 849–852.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equation of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1092.

National Research Council, 1996 *The Evaluation of Forensic DNA Evidence*, NRC2. National Academy Press, Washington, DC.

Nielsen, D. M., M. G. Ehm and B. S. Weir, 1998 Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am. J. Hum. Genet. **63:** 1531–1540.

Robinson, W. P., M. A. Asmussen and G. Thomson, 1991 Three-locus systems impose additional constraints on pairwise disequilibria. Genetics **129:** 925–930.

Shoemaker, J., I. Painter and B. S. Weir, 1998 A Bayesian characterization of Hardy-Weinberg disequilibrium. Genetics **149:** 2079–2088.

Shoemaker, J., I. Painter and B. S. Weir, 1999 Bayesian statistics in genetics: a guide for the uninitiated. Trends Genet. **15:** 354–358.

Slatkin, M., and L. Excoffier, 1996 Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity **76:** 377–383.

Smith, A. F. M., and J. M. Bernardo, 1994 *Bayesian Theory.* Wiley, Chichester, UK.

Smith, C. A. B., 1970 A note on testing the Hardy-Weinberg law. Ann. Hum. Genet. **33:** 377–383.

Smouse, P. E., 1974 Likelihood analysis of recombinational disequilibrium in multiple-locus gametic frequencies. Genetics **76:** 557–565.

Taillon-Miller, P., I. Bauer-Sardiña, N. L. Saccone, J. Putzel, T. Laitinen *et al.*, 2000 Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat. Genet. **25:** 324–328.

Weir, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35:** 235–254.

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

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

Wright, A. F., A. D. Carothers and M. Pirastu, 1999 Population choice in mapping genes for complex diseases. Nat. Genet. **23:** 397–404.

Yan, G., B. M. Christensen and D. W. Severson, 1997 Comparisons of genetic variability and genome structure among mosquito strains selected for refractoriness to a malaria parasite. J. Hered. **88:** 187–194.

Zapata, C., G. Alvarez and C. Carollo, 1997 Approximate variance of the standardized measure of gametic disequilibrium $D'$. Am. J. Hum. Genet. **61:** 771–774.

Zaykin, D., L. A. Zhivotovsky and B. S. Weir, 1995 Exact tests for association between alleles at arbitrary numbers of loci. Genetica **96:** 169–178.

Communicating editor: G. A. Churchill

## APPENDIX

**MCMC algorithm for genotype data:** Metropolis-Hastings algorithms are methods for generating a sample from an arbitrary probability distribution $\Pi$ (with probability density function $\pi$) by constructing a Markov chain whose stationary distribution is $\Pi$. If the current state of the chain is $x$, a candidate new state $x'$ is chosen with probability density $q(x'|x)$. The candidate is accepted with probability

$$\min\left(\frac{\pi(x')\,q(x|x')}{\pi(x)\,q(x'|x)},\ 1\right),$$

otherwise the current state $x$ is retained. A key feature of these algorithms is that $\pi$ need only be specified up to a normalizing constant, and so high-dimensional probability distributions can often be successfully handled. Although the states of the chain are correlated, selecting every $k$th iteration, after an initial "burn-in" period of length $b$, can lead to approximate random samples from $\Pi$ when suitable choices are made for $k$ and $b$. See Brooks (1998) for an introduction to MCMC algorithms.

For the algorithm implemented here, $\pi$ is the joint posterior density function of the $h_{ij}$. For two loci, each candidate $x'$ differs from $x$ at a randomly chosen pair of the $h_{ij}$, say $h_{rs}$ and $h_{wz}$. A proposal value $h'_{rs}$ is chosen uniformly in the interval

$$(\max(0,\ h_{rs} - \varepsilon),\ \min(h_{rs} + \varepsilon,\ h_{rs} + h_{wz})),$$

and $h'_{wz} = h_{wz} + h_{rs} - h'_{rs}$. The (positive) value of $\varepsilon$ is chosen to prevent proposed values from being rejected too often, which would result in slow movement of the chain around the sample space.

Slow convergence and poor mixing can arise in the presence of many alleles and/or loci. No difficulties were experienced with the examples discussed here that could not reasonably be overcome by choosing suitably large values for $k$ and $b$. A burn-in of $b = 30{,}000$ iterations was found to be adequate for the two-locus algorithm (50,000 for three loci), with every $k = 200$th (300) iteration output (these values having been determined by the inspection of sequential and autocorrelation plots of the output for initial runs). The output of each run underlying the figures and tables was analyzed with the MCMC diagnostic computer package CODA (Best *et al.* 1995), which indicated satisfactory convergence and mixing properties. For each application of the MCMC algorithm to the data, 5000 values were output (1000 for the permuted data).

**MCMC algorithm for haplotype data:** For three-locus haplotype data, assuming a multinomial log likelihood for the $h_{ijk}$ (given hyperparameters α, β, and γ), together with a Dirichlet prior distribution, after observing the $n_{ijk}$ the $h_{ijk}$ have a Dirichlet distribution with parameters $n_{ijk} + \lambda\alpha_i\beta_j\gamma_k$. A posterior sample from $p(\{h_{ijk}\}|\alpha, \beta, \gamma, \mathbf{N})$ can therefore be readily obtained using standard methods for the Dirichlet distribution (see, *e.g.*, Appendix A of GELMAN *et al.* 1995). To obtain a distribution for the $h_{ijk}$ that does not involve α, β, and γ, we can employ (6) together with a method of simulating from $p(\alpha, \beta, \gamma|\mathbf{N})$. A number of approaches are available, and details of an MCMC algorithm are given here.

The target distribution for the MCMC algorithm is $p(\alpha, \beta, \gamma|\mathbf{N})$, the probability density function of the hyperparameters α, β, and γ given the data **N**. The likelihood $p(\mathbf{N}|\alpha, \beta, \gamma)$ is of a standard form known as the multinomial-Dirichlet (SMITH and BERNARDO 1994, p. 135), and by Bayes theorem we can write

$$p(\alpha, \beta, \gamma|\mathbf{N}) = cp(\alpha, \beta, \gamma)\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\Gamma(n_{ijk} + \lambda\alpha_i\beta_j\gamma_k)}{\Gamma(\lambda\alpha_i\beta_j\gamma_k)}, \tag{7}$$

in which $c$ is a constant (and hence does not need to be known here) and $p(\alpha, \beta, \gamma)$ denotes the prior distribution for the hyperparameters, assumed here to be the product of multivariate uniforms so that $p(\alpha, \beta, \gamma)$ is also a constant.

A suitable Metropolis-Hastings algorithm can proceed as follows: first select a locus $l$, chosen uniformly at random. Suppose for notational convenience that α is the hyperparameter vector corresponding to the chosen locus $l$, then choose two elements of α, say $\alpha_v$ and $\alpha_w$. The proposal $\alpha'_v$ is chosen uniformly at random in the interval

$$(\max(0, \alpha_v - \varepsilon), \min(\alpha_v + \varepsilon, \alpha_v + \alpha_w)), \tag{8}$$

where ε is again a tuning parameter chosen to ensure that proposal values are not rejected either too frequently or too rarely. Finally, $\alpha'_w$ is assigned value $\alpha_w + \alpha_v - \alpha'_v$.

This algorithm, and modifications of it, can be useful in the location of disease loci via simple disequilibrium mapping. Briefly, under the assumption of a single disease mutation that arose sometime in the past, loci closest to the disease locus should exhibit higher levels of disequilibrium than those that are far away (*e.g.*, JORDE 1995). DEVLIN and RISCH (1995) detailed the use of point estimates of disequilibrium for inferring the closest of a number of biallelic markers (identifying the marker with the highest observed disequilibrium value as the closest). The direct probability approach, implementing the methods outlined above, adds a degree of interpretability to simple disequilibrium mapping, assigning probabilities to the event that a marker is closest to the disease locus—see AYRES (1998) for further details.

**Multiallelic three-locus normalized measures:** The following bounds apply to the $D_{ij}$ for two loci in a three-locus system (disequilibrium measures for the other locus pairs are denoted $D_{ik}$ and $D_{jk}$),

$$D_{ij_{\min}} = \max\{-p_iq_j, -(1 - p_i)(1 - q_j), -m_1, -m_2\}$$
$$D_{ij_{\max}} = \min\{p_i(1 - q_j), (1 - p_i)q_j, M_1, M_2\},$$

where

$$m_1 = p_iq_jr_k + (1 - p_i)(1 - q_j)(1 - r_k) + D_{ik} + D_{jk}$$
$$m_2 = p_iq_j(1 - r_k) + (1 - p_i)(1 - q_j)r_k - D_{ik} - D_{jk}$$
$$M_1 = p_i(1 - q_j)r_k + (1 - p_i)q_j(1 - r_k) + D_{ik} - D_{jk}$$
$$M_2 = p_i(1 - q_j)(1 - r_k) + (1 - p_i)q_jr_k - D_{ik} + D_{jk},$$

which are analogous to equations (12, a and b) and (13, a–d) of ROBINSON *et al.* (1991). The normalized parameters $D''_{ij}$ defined by these authors are then given by

$$D''_{ij} = \begin{cases} \dfrac{D_{ij}}{D_{ij_{\max}}} & \text{if } D_{ij} > 0 \text{ and } D_{ij_{\min}} \leq 0 \\ \dfrac{D_{ij} - D_{ij_{\min}}}{D_{ij_{\max}} - D_{ij_{\min}}} & \text{if } D_{ij} > 0 \text{ and } D_{ij_{\min}} > 0 \\ \dfrac{D_{ij}}{-D_{ij_{\min}}} & \text{if } D_{ij} < 0 \text{ and } D_{ij_{\max}} \geq 0 \\ \dfrac{D_{ij} - D_{ij_{\max}}}{D_{ij_{\max}} - D_{ij_{\min}}} & \text{if } D_{ij} < 0 \text{ and } D_{ij_{\max}} < 0. \end{cases}$$

The $D''_{ij}$ can be interpreted as the amount by which $|D_{ij}|$ exceeds its minimum value (given its sign), divided by its range. The overall pairwise measure $D''$ is calculated from the $D''_{ij}$ in the same way as $D'$ is defined in (2).