Received 16 July 1997

Measuring departures from Hardy–Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient

KAREN L. AYRES & DAVID J. BALDING*

Department of Applied Statistics, University of Reading, PO Box 240, Earley Gate, Reading RG6 6FN, U.K.

Many well-established statistical methods in genetics were developed in a climate of severe constraints on computational power. Recent advances in simulation methodology now bring modern, flexible statistical methods within the reach of scientists having access to a desktop workstation. We illustrate the potential advantages now available by considering the problem of assessing departures from Hardy–Weinberg (HW) equilibrium. Several hypothesis tests of HW have been established, as well as a variety of point estimation methods for the parameter f, which measures departures from HW under the inbreeding model. We propose a computational, Bayesian method for assessing departures from HW, which has a number of important advantages over existing approaches. The method incorporates the effects of uncertainty about the nuisance parameters — the allele frequencies — as well as the boundary constraints on f (which are functions of the nuisance parameters). Results are naturally presented visually, exploiting the graphics capabilities of modern computer environments to allow straightforward interpretation. Perhaps most importantly, the method is founded on a flexible, likelihood-based modelling framework, which can incorporate the inbreeding model if appropriate, but also allows the assumptions of the model to be investigated and, if necessary, relaxed. Under appropriate conditions, information can be shared across loci and, possibly, across populations, leading to more precise estimation. The advantages of the method are illustrated by application both to simulated data and to data analysed by alternative methods in the recent literature.

Keywords: Bayesian statistics, Hardy–Weinberg, inbreeding coefficient, inbreeding model, Markov chain Monte Carlo, Metropolis–Hastings algorithm.

Introduction

Statistical hypothesis testing for Hardy–Weinberg (HW) equilibrium has long had an important role in genetic studies (Haldane, 1954; Rousset & Raymond, 1995). It has received renewed attention in recent years (e.g. Guo & Thompson, 1992; Zaykin *et al.*, 1995), resulting in part from the debate over population genetics issues in the forensic use of DNA profiling. Although hypothesis testing provides some insight into the questions of scientific interest, it forms only the most basic level of statistical inference. Tests do not directly measure the size of the effect: for example, a deviation from HW may be statistically significant, and yet insignificant in the everyday sense for the application at hand.

This weakness has long been recognized, and a number of methods have emerged for obtaining point estimates of f, a parameter measuring departure from HW caused by inbreeding. However, such point estimates are also unsatisfactory for a number of reasons. First, an investigator will most often be interested in the distribution of plausible values, rather than just a (typically imprecise) point estimate. Secondly, standard errors can be attached to point estimates, but these are of limited value in connection with estimators having sampling distributions that may be highly skew and are bounded below by the requirement that all genotype frequencies be non-negative. Indeed, some point estimation methods can even produce estimates outside this bound. Finally, a satisfactory approach to estimation should allow the assumptions of the inbreeding model to be assessed and, if necessary, weakened.

*Correspondence. E-mail: d.j.balding@reading.ac.uk

The first attempt to evaluate a probability distribution for a parameter measuring departure from HW seems to have been that of Lindley (1988). This paper did not exploit modern computational methodology and dealt only with diallelic loci, for which a one-parameter model for deviations from HW is fully general. For highly polymorphic molecular genetic markers, a fully general model involves a distinct parameter for each heterozygous genotype, such as the fixation indices of Weir (1970) or the additive disequilibrium coefficients of Hernández & Weir (1989). Although they can be readily implemented, the data will provide little information for such highly parameterized models, and the resulting estimates may be very imprecise.

We do discuss a more general model below, but focus first on the one-parameter inbreeding model. We outline some of the current point estimation methods available for the parameter f, the inbreeding coefficient, and then describe a Markov chain Monte Carlo (MCMC) method for approximating the probability density of f based on a sample of genotypes from the population. We illustrate the method by applying it to simulated data and to data analysed by alternative methods in the recent literature. Next, we describe a method for investigating the validity of the inbreeding model. Finally, we discuss combining information over several loci, illustrating this with simulated data and with data from Samoan individuals at three short tandem repeat (STR) loci used in forensic work.

Computer programs (in C) for the MCMC algorithms are freely available from http://www.reading.ac.uk/~snsbalng/.

The inbreeding model

If inbreeding is expected to be the main cause of any deviation from HW, the inbreeding model may be appropriate (for example see Malécot, 1969). This model is completely general for diallelic loci but, in the multiallelic case, it cannot account for assortative mating or for some forms of selection. Under the inbreeding model, p_{ij}, the relative frequency of the genotype A_iA_j, is

$$p_{ii} = p_i(f + (1-f)p_i)$$
$$p_{ij} = 2p_ip_j(1-f), \qquad (1)$$

where p_i denotes the frequency of allele A_i, and f is the inbreeding coefficient. When $f = 0$, eqn (1) gives the HW proportions. When $f = 1$, the maximum value, heterozygotes never arise. The value of f can be negative. It is bounded below by the requirement that the population frequencies of each homozygote be non-negative, which leads to:

$$f \geq \frac{-p_{\min}}{1-p_{\min}}, \qquad (2)$$

where $p_{\min}$ is the smallest allele frequency.

The value of f can be interpreted as the correlation between an individual's two genes at a locus (for example see Crow & Kimura, 1970). It measures the deficit (or excess) of heterozygosity that results from inbreeding (or outbreeding). In some models for population subdivision, f can be interpreted as the probability that an individual's two genes are identical by descent, in which case it is constrained to be non-negative.

Point estimation methods for f

Nei & Chesser (1983) discuss an estimator for the inbreeding coefficient:

$$\hat{f}_{\mathrm{nc}} = \frac{\Sigma_i(X_{ii} - x_i^2) + (1 - \Sigma_i X_{ii})/2n}{(1 - \Sigma_i x_i^2) - (1 - \Sigma_i X_{ii})/2n}, \qquad (3)$$

where X_{ii} and x_i are the sample frequencies of A_iA_i genotypes and A_i alleles, and n is the number of individuals in the sample. This estimator is developed in terms of a subdivided population. In this setting, the parameter f in a single subpopulation is often denoted F_{IS}.

Robertson & Hill (1984) give an alternative estimator:

$$\hat{f}_{\mathrm{rh}} = \frac{1}{k-1}\sum_i \left(\frac{2(2n-1)n_{ii} - n_i(n_i-1)}{2n_i(n-1)}\right), \qquad (4)$$

where n_{ii} and n_i denote the sample counts of A_iA_i genotypes and A_i alleles, and k is the number of alleles at the locus.

These estimators do not explicitly take account of the inbreeding model and may, in the multiallelic case, give estimates that conflict with the bound (2). The maximum likelihood estimator under the inbreeding model does respect this bound. Assuming random sampling of genotypes, the likelihood is:

$$\mathrm{P}(\{n_{ij}\} \mid f, p_1, \ldots, p_k)$$
$$= c \prod_{i=1}^{k} (p_i(f + (1-f)p_i))^{n_{ii}} \prod_{j=i+1}^{k} (2p_ip_j(1-f))^{n_{ij}}, \quad (5)$$

where c is a constant. For $k = 2$, eqn (5) is readily maximized (for example, see Weir, 1996, p. 65) to obtain

$$\hat{f}_{\mathrm{mle}} = 1 - \frac{2n_{12}n}{(2n_{11}+n_{12})(n_{12}+2n_{22})}. \qquad (6)$$

For $k > 2$, the likelihood cannot be maximized analytically, but numerical methods such as that given in Robertson & Hill (1984) can be employed, although problems may arise with iterates going out of bounds. For the case when the maximum likelihood estimate (MLE) is non-negative, the EM algorithm given in Hill *et al.* (1995) may be used. More general mode-finding algorithms are described in chapter 9 of Gelman *et al.* (1995).

MCMC method

Although standard errors can be attached to the point estimators described above, these are of limited value for estimators with bounded, and possibly highly skewed, sampling distributions. The profile likelihood for f (the likelihood function obtained by setting each nuisance parameter p_i equal to its MLE $\hat{p}_i$) does provide a measure of the support given by the data to different possible values for f, but it ignores uncertainty in the p_i.

The nuisance parameter problem can be overcome by integration over the joint distribution of the p_i, leading to a marginal likelihood for f, which is also its posterior distribution when the joint prior distribution for f and the allele frequencies is (multivariate) uniform over the range of possible values. Informative prior distributions, reflecting, for example, knowledge that f is unlikely to be large or to be negative, or information about the allele frequencies from previous studies, can also be incorporated.

When exact integration is not feasible, approximate integration can be achieved via one of a range of stochastic simulation methods known as MCMC algorithms. These algorithms generate a sequence of realizations from a specified probability distribution, which can then be used to approximate properties of the distribution to any required accuracy. We implement an algorithm of the Metropolis–Hastings type (Metropolis *et al.*, 1953; Hastings, 1970; Smith & Roberts, 1993). Details are given in the Appendix.

Application to simulated data

The solid curves in Fig. 1 show the posterior density for f from samples of size $n = 200$ and $n = 1000$, at loci with $k = 2$, $k = 6$ and $k = 15$ alleles, simulated from the inbreeding model with $f = 5\%$ and assuming a uniform prior density. Corresponding point estimates of f are given in Table 1. These are given for comparison with Fig. 1 only and not to assess the properties of the point estimators, for which see, for example, Curie-Cohen (1982).

Note that for $k = 2$ and $n = 200$ (Fig. 1a), the density curve is relatively flat and visibly non-zero over a large interval, reflecting the fact that there is little information about f in the data. Fortuitously in this simulation, the point estimates are around 6%, close to the true value of f (5%). The posterior density curve, however, clarifies the level of uncertainty: values for f as small as -8% and as large as 20% remain plausible based on the data.

When either the number of alleles k or the sample size n is increased, additional information is available from the data, which is reflected by sharper peaks in the density curves (Fig. 1b–f, solid curves). For these simulations, only in the case $k = 15$ and $n = 1000$ is the hypothesis $f = 0$ unequivocally excluded: in the other cases, the density curve is visibly above zero at $f = 0$.

If additional data are unobtainable or expensive, a careful choice of prior distribution for f can be helpful in narrowing the range of plausible values. For example, the dashed curve in Fig. 1(a) shows the posterior probability density for f corresponding to an informative prior density (shown in Fig. 2), which reflects a belief that f is likely to be close to zero. The effect of incorporating this prior information is almost the same as the effect of increasing the sample size from 200 to 1000 (Fig. 1b, solid curve). In some cases, it may be reasonable to assume *a priori* that f is non-negative, in which case a simpler and more efficient algorithm can be implemented, because the bound (2) does not need to be recalculated at each iteration.

Each panel in Fig. 1 corresponds to a single data set. Although there will be some variation among replicate data sets with the same parameters, the patterns of increasing precision with greater k and n, and the effect of an informative prior density, are broadly unaltered (simulations not shown).

Comparison with analyses from recent literature

Table 2 gives point and interval estimates of f obtained by Hill *et al.* (1995) from a sample of size 60 of the human malaria parasite *Plasmodium falciparum*. Corresponding posterior density curves obtained via the MCMC method with a uniform prior density are shown in Fig. 3(a). As both approaches are likelihood based, the two sets of results do not conflict. However, the MCMC method is visual, and so immediately interpretable, reveals the support given to negative values, incorporates exactly the effect of uncertainty about the nuisance parameters and allows the inclusion of background information to narrow the range of plausible values.

Figure 3(b) shows the result of applying the MCMC method to a sample of $n = 8297$ Rhesus genotypes given in fig. 3 of Guo & Thompson (1992), who used hypothesis testing methods. The Rhesus locus is a highly polymorphic blood group genetic marker. Assortative mating with respect to blood groups is unlikely to occur, and any selection is thought to be negligible, so that the inbreeding model may be reasonable. Point estimates for f are extremely small:

$$\hat{f}_{\mathrm{nc}} = 0.00102, \quad \hat{f}_{\mathrm{rh}} = 0.00063, \quad \hat{f}_{\mathrm{mle}} = 0.00141,$$

and the exact-test P-values reported by Guo & Thompson (1992) are large, exceeding 0.69. Similarly, the posterior density for f indicates that values close to zero are highly plausible. However, the MCMC analysis also reveals that values for f in excess of 1% are consistent with the data, even though the locus is multiallelic and the sample is very large.

Note that our analysis assumes that alleles not present in the sample do not exist in the population under investigation. If there did exist a very rare allele not represented in the sample, the bound (2)

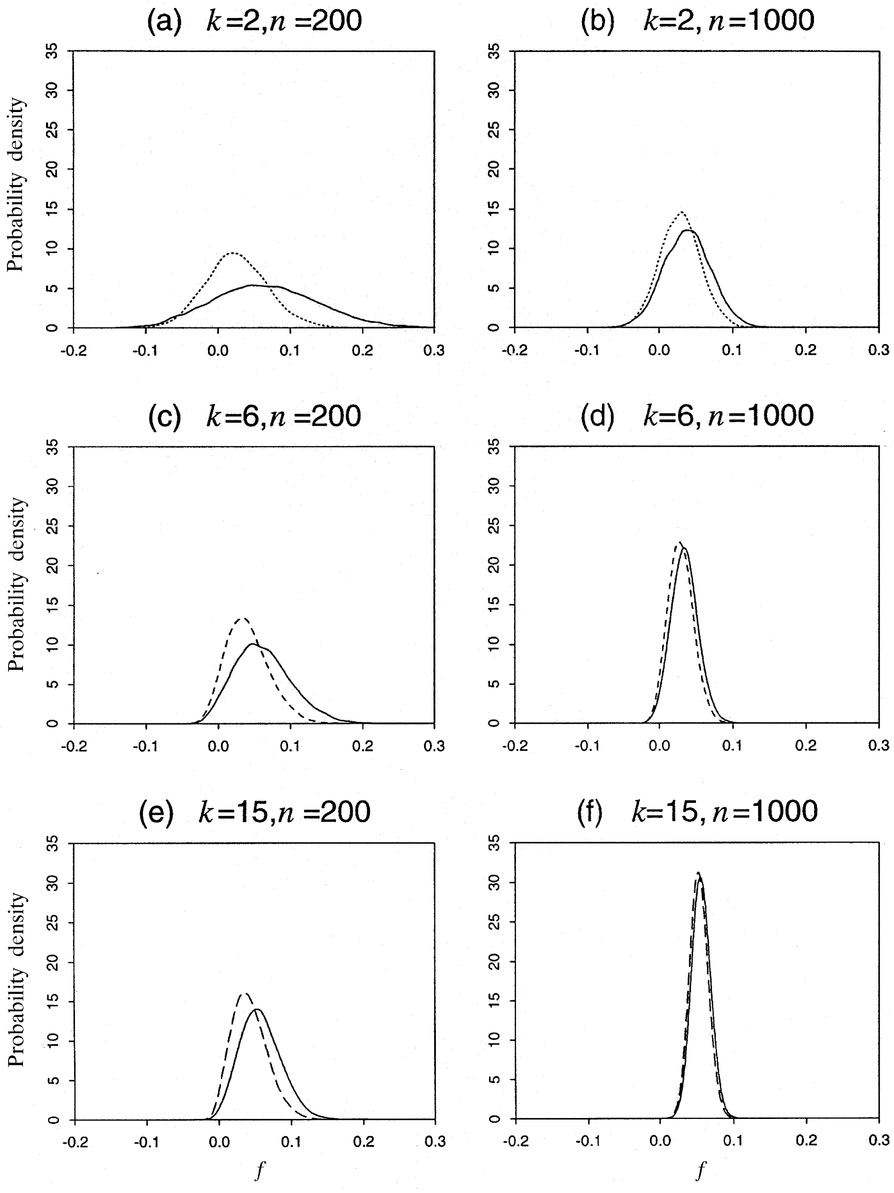

Fig. 1 Posterior density curves for the inbreeding coefficient f, for samples of size n = 200 and n = 1000, at loci with k = 2, k = 6 and k = 15 alleles, simulated from the inbreeding model with f = 5%. Solid and dashed curves correspond, respectively, to a uniform prior and an informative prior for f (shown in Fig. 2). A uniform prior was used for the allele frequencies in each case. Point estimates based on the same data are given in Table 1. The density curves are obtained by applying the density command of the statistical package S-PLUS to 10 000 values generated by an MCMC algorithm (burn-in length 5000; then every 50th value retained). See Appendix for further details.

would, in effect, restrict f to be non-negative because $p_{\min}$ would be very small.

Investigating the validity of the inbreeding model

The method thus far developed has concentrated on estimating the inbreeding coefficient f, the parameter of the inbreeding model (eqn 1). For more than two alleles, a thorough investigation into departures from HW should also examine the validity of this model. This can be performed by specifying a more general model, of which the inbreeding model is a special case, and examining whether or not the data support aspects of the general model not consistent with this special case. A natural extension of the one-parameter inbreeding model is the fixation indices model (Weir, 1970) for which genotype frequencies are

$$p_{ii} = p_i^2 + p_i \sum_{j \neq i} p_j f_{ij}$$

$$p_{ij} = 2p_i p_j (1 - f_{ij}), \qquad (7)$$

where f_{ij} is the fixation coefficient for the heterozygous genotype A_iA_j.

Table 1 Nei & Chesser (1983), Robertson & Hill (1984) and maximum likelihood estimates of the inbreeding coefficient f, for data sets of size $n = 200$ and $n = 1000$, with $k = 2$, $k = 6$ and $k = 15$ alleles, simulated from the inbreeding model with $f = 5\%$

k	n	$\hat{f}_{\mathrm{nc}}$	$\hat{f}_{\mathrm{rh}}$	$\hat{f}_{\mathrm{mle}}$
2	200	0.0613	0.0615	0.0588
2	1000	0.0371	0.0372	0.0366
6	200	0.0804	0.0566	0.0488
6	1000	0.0508	0.0263	0.0312
15	200	0.0558	0.0372	0.0489
15	1000	0.0515	0.0492	0.0539

The (population) allele frequency vectors (in %) were (25,75), (2,6,7.5,8.5,21,55) and (1,1,2,3,3,4,4,4,5,6,9,10,12,17,19).

For $k = 2$ alleles, eqn (7) reduces immediately to the inbreeding model (eqn 1). For $k > 2$, the inbreeding model is recovered when all the f_{ij} are equal. Inspecting the posterior densities of the coefficients f_{ij} will therefore provide an insight into the validity of the inbreeding model — if the posteriors do not overlap to any noticeable degree, the model may be invalid. For small sample sizes, the posteriors of the f_{ij} may each support a wide range of plausible values, so that the fact that they all overlap may reflect insufficient data to distinguish competing models rather than strong support for the inbreeding model. If the model does appear unsuitable, the posteriors of each f_{ij} that have been obtained may be used to infer the nature of departures from HW. Note that the lower bound of each f_{ij} is a function of the corresponding allele frequencies, which can complicate within-sample comparisons in the event that negative values are highly supported. An MCMC method for obtaining posterior densities for the fixation indices is detailed in the Appendix.

Combining information over loci

We noted from Fig. 1(a) that estimation of f when $n = 200$ and $k = 2$ is very imprecise. If deviations from HW are caused predominantly by inbreeding, then f should be approximately constant over loci. In this case, sharper estimation can be obtained by

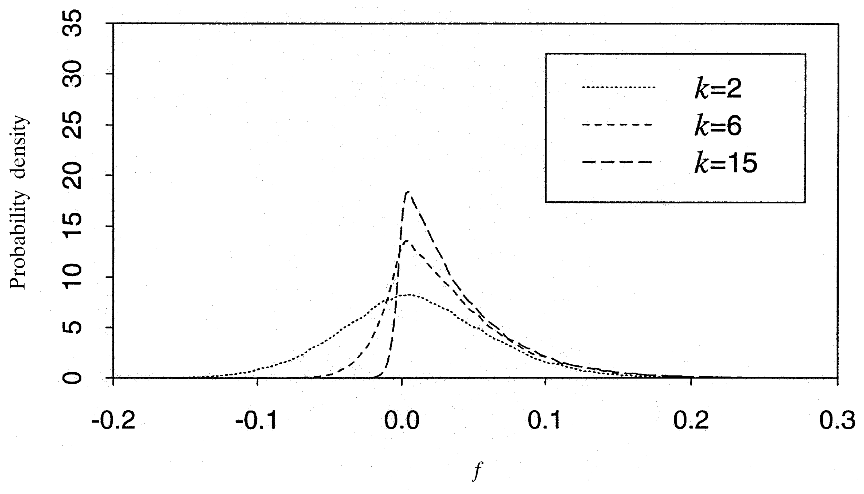

Fig. 2 Prior density curves corresponding to the dashed curves of Fig. 1.

Table 2 MLE of f, and intervals in which the log-likelihood is within two of the maximum value, based on a sample of size 60 of the human malaria parasite *Plasmodium falciparum* (taken from table 3 of Hill *et al.*, 1995)

Locus	k	MLE	Interval
MSP-1	3	15%	0–35%
MSP-2	2	39%	14–61%

pooling together information from different loci. There may also be situations in which it is reasonable to assume that f has the same value in different populations, in which case information could also be shared across populations.

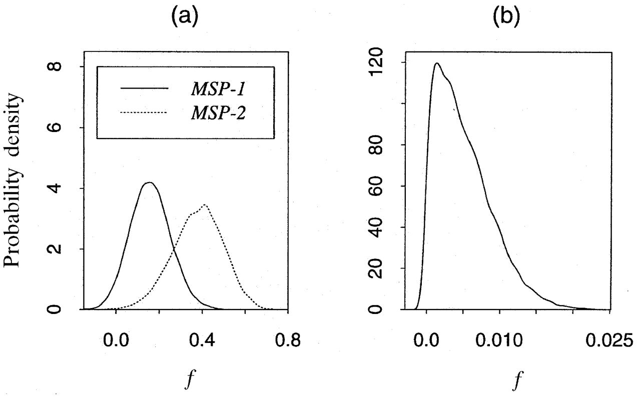

Fig. 3 Posterior density curves for f for (a) the *Plasmodium falciparum* data of Hill *et al.* (1995) and (b) the Rhesus data given in Guo & Thompson (1992). Details of the MCMC algorithm used to obtain these curves are the same as for the solid curves of Fig. 1.

The five dashed curves in Fig. 4 show posterior densities for f evaluated from samples of size $n = 200$ simulated with $f = 5\%$ at each of five loci. The solid curve shows the posterior density evaluated for all five loci assuming a common value of f. The prior density is uniform for each curve, and the likelihood for the combined data is given by the product of the likelihoods for each locus. This assumes independence of genotypes at different loci, given the value of f, which would be inappropriate in the presence of gametic disequilibrium or genotypic association.

Notice that the five dashed curves in Fig. 4 all overlap substantially: if this did not occur, it would suggest that the assumption of constant f is invalid. As expected, the solid curve has a higher peak than any of the dashed curves, indicating that more precise estimation is obtained by pooling information from the five data sets.

Hill *et al.* (1995) combine the information from the *MSP-1* and *MSP-2* loci to obtain an overall point estimate of 24%. Their corresponding interval is 9–39%, narrower than either of the single-locus intervals given in Table 2. Similarly, the posterior density curve for the combined data based on a uniform prior density (not shown) is more sharply peaked than either of the single-locus curves shown in Fig. 3(a).

Application to Samoan data

The dashed curves of Fig. 5 show the result of applying the MCMC methods outlined above to data from three STR loci (*THO1*, *TPOX* and *CSF1PO*) for a sample, collected in forensic work, of 143 Samoans resident in New Zealand. Six alleles were

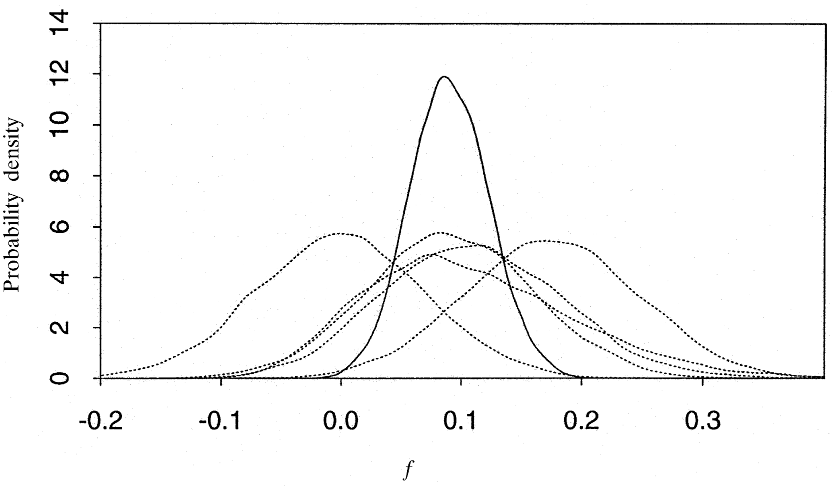

Fig. 4 Posterior density curves for f from each of five simulated single-locus data sets with $k = 2$, $n = 200$ and $f = 5\%$ (dashed curves) and from the combined data assuming f constant over loci (solid curve). Each prior density is uniform.

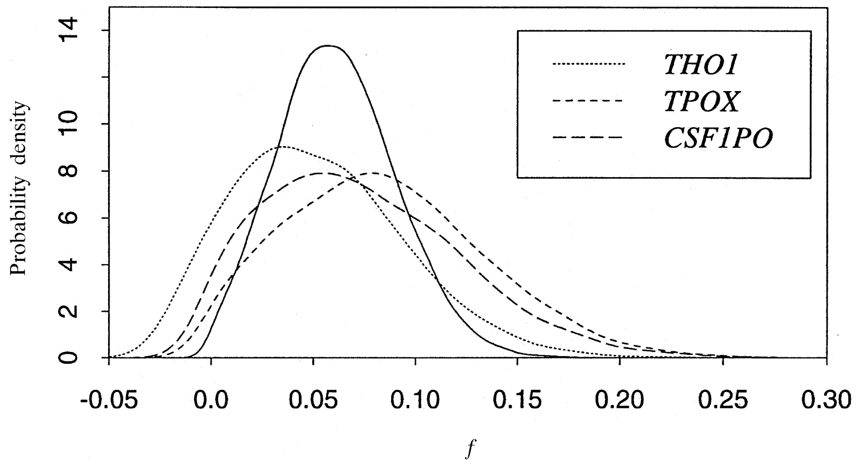

Fig. 5 The broken curves are the posterior density curves for f from three single-locus STR Samoan samples each with $k = 6$. The sample sizes are 142 (*THO1*), 140 (*TPOX*) and 141 (*CSF1PO*) individuals. The solid curve represents the posterior density for f obtained from the combined data assuming f constant over loci. Each prior density is uniform. Data provided by John Buckleton of ESR, New Zealand.

observed at each locus in the data set, although additional alleles are known to exist at these loci in other populations. As for the Rhesus locus considered earlier, our analysis assumes that alleles not represented in the sample do not exist in the Samoan population.

The fixation coefficient model yielded posteriors for each f_{ij} that were diffuse and overlapped substantially. (Uniform prior densities were used for the f_{ij}.) The inbreeding model may, therefore, be reasonable for these data. Moreover, the dashed curves of Fig. 5 overlap substantially, supporting a common value for f at the three loci. The solid curve represents the posterior density for f obtained by combining the data over the three loci, again assuming a uniform prior density for f.

The data support a large range of plausible values over the three loci, from about -3% to more than 20%. Combining information over the three loci results in some improvement in estimation, with the plausible range being narrowed to, say, 0–14%, and with values between 3% and 8% now being highly supported. Further improvement could be obtained by implementing an informative prior for f.

Discussion

The most obvious advantage of the MCMC method outlined here, compared with traditional approaches, is that results are represented visually in terms of posterior density curves and are thus readily interpretable. The method also has the advantage of allowing the scientist to incorporate background information if desired, thus reducing the amount of direct data required. The method is flexible and readily implemented. As well as being useful in practice, the method is also well supported in statistical theory: there are compelling reasons to support the view that uncertainty about an unknown parameter should, if possible, be described by its probability distribution (Smith & Bernardo, 1994).

Figure 1 shows that large ranges of plausible values often arise, particularly when few alleles can be distinguished and/or the sample size is small. The MCMC method highlights the resulting uncertainty in a direct and visual manner, making it preferable to point estimation methods. Figure 1 also shows that sharper estimation can be achieved by incorporating prior knowledge about the parameter, if desired. Alternatively, or in addition, estimation may be improved by combining information across loci (Fig. 4).

The likelihood basis of the method makes it very flexible, allowing investigation of the validity of the inbreeding model. If the model does not appear reasonable, the posteriors of the fixation indices f_{ij} may be used to quantify the departure from HW indicated by the sample.

Acknowledgements

We thank John Buckleton of the Institute of Environmental Science and Research Limited, Auckland, New Zealand, for kindly providing the Samoan data. K.L.A. was supported by a Ph.D. studentship from the UK EPSRC. D.J.B. was supported in part by the EPSRC under grant GR/K72599.

References

CROW, J. F. AND KIMURA, M. 1970. *An Introduction to Population Genetics Theory.* Harper & Row, New York.

CURIE-COHEN, M. 1982. Estimates of inbreeding in a natural population: a comparison of sampling properties. *Genetics*, **100**, 339–358.

GELMAN, A., CARLIN, J. B., STERN, H. S. AND RUBIN, D. B. 1995. *Bayesian Data Analysis.* Chapman & Hall, London.

GUO, S. W. AND THOMPSON, E. A. 1992. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361–372.

HALDANE, J. B. S. 1954. An exact test for randomness of mating. *J. Genet.*, **52**, 631–635.

HASTINGS, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **5**7, 97–109.

HERNÁNDEZ, J. L. AND WEIR, B. S. 1989. A disequilibrium coefficient approach to Hardy–Weinberg testing. *Biometrics*, **45**, 53–70.

HILL, W. G., BABIKER, H. A., RANFORD-CARTWRIGHT, L. C. AND WALLIKER, D. 1995. Estimation of inbreeding coefficients from genotypic data on multiple alleles, and application to estimation of clonality in malaria parasites. *Genet. Res.*, **65**, 53–61.

LINDLEY, D. V. 1988. Statistical inference concerning Hardy–Weinberg equilibrium. In: Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M. (eds) *Bayesian Statistics*, vol. 3, pp. 307–326. Oxford University Press, Oxford.

MALÉCOT, G. 1969. *The Mathematics of Heredity*. W. H. Freeman, San Francisco.

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

NEI, M. AND CHESSER, R. K. 1983. Estimation of fixation indices and gene diversities. *Ann. Hum. Genet.*, **47**, 253–259.

ROBERTSON, A. AND HILL, W. G. 1984. Deviations from Hardy–Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics*, **107**, 703–718.

ROUSSET, F. AND RAYMOND, M. 1995. Testing heterozygote excess and deficiency. *Genetics*, **140**, 1413–1419.

SMITH, A. F. M. AND BERNARDO, J. M. 1994. *Bayesian Theory.* Wiley, Chichester.

SMITH, A. F. M. AND ROBERTS, G. O. 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc. B*, **5**5, 3–23.

WEIR, B. S. 1970. Equilibria under inbreeding and selection. *Genetics*, **65**, 371–378.

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

ZAYKIN, D., ZHIVOTOVSKY, L. A. AND WEIR, B. S. 1995. Exact tests for association between alleles at arbitrary numbers of loci. *Genetica*, **96**, 169–178.

Appendix: Details of MCMC algorithms

The general Metropolis–Hastings algorithm generates a sample from a probability distribution $g(x)$, by constructing a Markov chain $X^1,\ldots, X^t,\ldots$ whose stationary distribution is $g(x)$. At step t, a proposed value x' for X^t is chosen from a distribution with probability density denoted by $q(x' \mid x)$, where $x = X^{t-1}$, the current state. The proposed value is accepted with probability

$$\min\left(\frac{g(x')q(x \mid x')}{g(x)q(x' \mid x)}, 1\right). \tag{8}$$

The key feature of the algorithm is that $g(x)$ need not be fully specified — the normalizing constant is not required. For Bayesian analyses, $g(x)$ can be expressed as the product of the prior and the likelihood, without evaluating the denominator of Bayes's rule, and even the prior distribution need only be specified up to a constant of proportionality. See Smith & Roberts (1993) for further details.

For the application to the inbreeding model, each X is a vector consisting of values for f and the p_i, and $g(x)$ is the joint posterior distribution of this vector given the sample genotype counts $(n_{11},\ldots, n_{kk})$. The specification of the proposal distribution q is complicated by the fact that the allele frequencies must sum to one. We update the allele frequencies in pairs, say p_u and p_v, with u and v chosen randomly at each iteration. A proposal value p_u' is chosen uniformly and randomly between $\max(0,p_u-\varepsilon_p)$ and $\min(p_u+\varepsilon_p,p_u+p_v)$, and then p_v' is set equal to $p_u+p_v-p_u'$. The (positive) value of ε_p is chosen to obtain reasonable acceptance rates: if ε_p is too large, the value of eqn (8) will usually be very small, leading to a chain that 'sticks' too much in one place and, hence, converges slowly; if too small, the chain will make frequent but very small moves and again converge slowly (for example, see Hastings, 1970).

Next, a proposal value f' is chosen uniformly and randomly in the interval

$$\left(\max\left(\frac{-p'_{\min}}{(1-p'_{\min})}, f-\varepsilon_f\right), \min(f+\varepsilon_f, 1)\right),$$

where f denotes the current value. Because the lower bound in selecting f' is a function of $p'_{\min}$ rather than $p_{\min}$, it may occur that the left limit of this interval exceeds the right limit. Setting $\varepsilon_f > k^2\varepsilon_p/[(k-1)(k-1-k\varepsilon_p)]$ avoids such problems and ensures irreducibility of the Markov chain, which is required for guaranteed convergence of the algorithm, for example see Smith & Roberts (1993).

The starting position of the chain can be chosen arbitrarily. For the simulations discussed here, we found it acceptable to generate starting values from the prior distributions. We found that a 'burn-in' of 5000 iterations adequately reduced the dependence of the output on the starting point. Output was recorded at every 50 iterations, which reduced serial correlation to a satisfactory level.

For the application to the fixation indices model, complications arise because of the dependence between allele frequencies and the f_{ij}. For example,

$$1-\frac{1}{2p_ip_j} \le f_{ij} \le 1 \tag{9}$$

must be satisfied for all i and j. Problems may be encountered when a proposed allele frequency invalidates some current f_{ij} values. If the sample size is not too small, the error arising by setting the allele frequencies equal to the sample values is negligible. Because small sample sizes may in any case fail to provide sufficient information for model discrimination, we make this assumption here, thus simplifying the algorithm by sidestepping the difficulties described above. The probability distribution $g(x)$ is therefore the joint posterior distribution of the f_{ij} given $(n_{11},\ldots, n_{kk})$ and the sample allele frequencies.

The updating of the f_{ij} is further complicated by the following constraints

$$-p_i \le \sum_{j\ne i} p_jf_{ij} \leqslant \frac{1}{p_i} - p_i, \tag{10}$$

for all i. We, therefore, update a single (randomly selected) parameter, f_{uv} say, with the proposed value f'_{uv} being chosen uniformly and randomly between the bounds

$$\max\left(\frac{-p_u+p_vf_{uv}-\sum\limits_{j\ne u} p_jf_{uj}}{p_v}, \frac{-p_v+p_uf_{uv}-\sum\limits_{i\ne v} p_if_{iv}}{p_u}, f_{uv}-\varepsilon, 1-\frac{1}{2p_up_v}\right)$$

and

$$\min\left(\frac{p_u^{-1}-p_u+p_vf_{uv}-\sum\limits_{j\ne u} p_jf_{uj}}{p_v}, \frac{p_v^{-1}-p_v+p_uf_{uv}-\sum\limits_{i\ne v} p_if_{iv}}{p_u}, f_{uv}+\varepsilon, 1\right),$$

where f_{ij} denotes the current value and ε can be chosen to ensure reasonable acceptance rates.

The starting position of the chain can again be chosen arbitrarily, provided that the constraints (9) and (10) are satisfied. Randomly generating starting values for each f_{ij} uniformly and independently between 0 and $1/k$ is acceptable. A 'burn-in' of 20 000 iterations was used, and output was recorded every 100 iterations.