April 2004

Bayesian approaches for the analysis of population genetic structure: an example from Platanthera leucophae (Orchidaceae)

Kent E. Holsinger
University of Connecticut, kent.holsinger@uconn.edu

Lisa E. Wallace
University of South Dakota

Bayesian approaches for the analysis of population genetic structure: an example from *Platanthera leucophaea* (Orchidaceae)

Kent E. Holsinger

Department of Ecology & Evolutionary Biology, U-3043

University of Connecticut

Storrs, CT 06269-3043

USA

Lisa E. Wallace

Department of Biology

University of South Dakota

Vermillion, SD 57069

USA

Author for correspondence: Kent Holsinger (kent@darwin.eeb.uconn.edu)

Keywords: Bayesian statistics, *F*-statistics, population genetic structure

Abstract

We describe four extensions to existing Bayesian methods for analysis of genetic structure in populations: (1) use of beta distributions to approximate the posterior distribution of f and θ^B, (2) use of an entropy statistic to describe the amount of information about a parameter derived from the data, (3) use of the Deviance Information Criterion (DIC) as a model choice criterion for determining whether there is evidence for inbreeding within populations or genetic differentiation among populations, and (4) use of samples from the posterior distributions for f and θ^B derived from different data sets to determine whether the estimates are consistent with one another. We illustrate each of these extensions by applying them to data derived from previous alloyzme and RAPD surveys of an endangered orchid, *Platanthera leucophaea*, and we conclude that differences in θ^B from the two data sets may represent differences in the underlying mutational processes.

Introduction

For more than seventy years population and evolutionary geneticists have been interested in describing how genetic diversity is distributed within and among populations, and since Sewall Wright (1951) and Gustave Malécot (1948) introduced them, *F*-statistics have been the descriptor of choice. Unfortunately, neither Wright nor Malécot paid careful attention to the problem of estimating these statistics from sample data, but with the advent of modern computers and modern molecular methods interest in the estimation of *F*- statistics has shown a dramatic increase (see Excoffier 2001; Weir & Hill 2002 for recent, comprehensive reviews).

Coincident with the increasing interest in analysis of population genetic structure has been an explosion of interest in the use of Bayesian statistical techniques for analysis of many complex, hierarchical statistical problems. More recently, these and closely-related likelihood techniques have been applied to analysis of genetic data (e.g., Balding & Nichols 1995; Roeder et al. 1998; Holsinger 1999; Balding 2003).

In this paper we describe further extensions of the Bayesian approach for analysis of structure in population genetic data. We illustrate that a beta distribution (with appropriately chosen parameters) provides a good fit to the posterior distributions of estimators for Wright's F_{is} and F_{st} and that an entropy statistic provides a useful measure of the amount of information provided by the data about the parameters. We focus our attention, however, on two problems of more immediate importance: (1) Developing a model choice criterion that can be used to determine when the data provide strong evidence for inbreeding within populations or for genetic differentiation among them. (2) Developing a method for comparing non-zero estimates of F_{is} and F_{st} derived from different data sets.

We illustrate these extensions by showing that there is substantial inbreeding within populations of *Platanthera leucophaea*, an endangered orchid in the United States and Canada.

The data also show that there is substantially more genetic differentiation among the sampled populations at allozyme loci than at loci coding for RAPD markers. Because both data sets include exactly the same set of populations and virtually the same set of individuals, the differences in genetic structure are unlikely to reflect differences in migration rates or demographic history of the populations. One likely explanation is a higher rate of mutation at loci encoding variation in RAPD markers.

Materials and Methods

Data

Platanthera leucophaea (Nuttall) Lindley is a showy orchid species that was once abundant in prairies and sedge meadows of the midwestern United States and Canada, primarily east of the Mississippi River. Presently, it is listed as a threatened species under the U.S. Endangered Species Act as only 59 extant populations are known from Illinois, Iowa, Maine, Michigan, Ohio, and Wisconsin (USFWS 1999). In addition to having an extremely fragmented distribution, many populations are also very small (i.e., fewer than 50 aboveground individuals). Population size is necessarily tied to the lifestyle of a species, which is quite complex in *Platanthera leucophaea*. This species is adapted to graminoid habitats that are routinely disturbed, especially as a result of fire and drought, and populations exhibit periods of dormancy or mass flowering. Populations are maintained only through sexual reproduction by seed, which requires pollination by hawkmoths and the formation of mycorrhizae for seedling establishment.

The populations included in this survey represent seven of the ten known populations of *P. leucophaea* in Ohio (USFWS 1999). These populations occur in prairie situated in the plains of Lake Erie (Metzger, Pickerel, and Yandota) or in wet sedge meadow in the central part of the state (Meadow, Medway, Conservation, Cemetery). Of the three unsampled populations, no

individuals were found at two sites in 1998 when tissue was collected, and the other unsampled population occurs on private land and could not be visited in 1998. In the three smallest populations (Meadow, Conservation, and Cemetery), all individuals were sampled, while in the four larger populations (Medway, Metzger, Pickerel, and Yandota), a representative sample of individuals was chosen randomly (Table 1). With only a few exceptions, all individuals scored for RAPD markers were also genotyped at polymorphic allozyme loci. The populations included in these analyses are a subset of those considered in Holsinger et al. (2002).

Wallace (2002) provides details on buffer systems and genetic interpretations of allozymes and on scoring of RAPD markers. Briefly, the allozyme data set includes 7 polymorphic loci (TPI-1, TPI-2, EC 5.3.1.1; MDH-2, EC 1.1.1.37; PGM-1, EC 2.7.5.1; GOT-1, GOT-2, EC 2.6.1.1; CAT-1, EC 1.11.1.6), each with two alleles per locus except for CAT-1, which had three alleles. The RAPD data set includes 63 polymorphic loci.

Statistical analysis

We analyzed the data using `Hickory` v0.8 (Holsinger & Lewis 2003). Following the notation in Holsinger (1999) and Holsinger et al. (2002), which was inspired by the close analogy to Weir and Cockherham's (1983) formulation of *F*-statistics,we use f to refer to F_{is}, and θ^B to refer to F_{st}. For dominant marker data, the full conditional distribution of the parameters is given by:

$$P(f,\theta,p,\pi \mid N_{A1}N_{A2}) \propto \left(\prod_{i=1}^{I}\left(\prod_{k=1}^{K}\gamma_{A1,ik}^{N_{A1,ik}}\gamma_{A2,ik}^{N_{A2,ik}}P(\gamma_{ik}\mid\pi_i,\theta,f)\right)P(\pi_i)\right)P(\theta)P(f)\ ,$$

where $P(f)$ and $P(\theta)$ are the prior distributions on f and θ and where $N_{A1,ik}$ and N_{A2},ik refer to the number of dominant and recessive phenotypes at locus i in population k. $P(\pi_i)$ is the prior distribution on the mean allele frequency at each locus, and $P(\gamma_{ik}|\pi_i,\theta,f)$ is the induced prior on

phenotype frequencies in each population. To calculate $P(\gamma_{ik}|\pi_i,\theta,f)$ we assume that the prior distribution of allele frequcncies at locus i in the k^{th} population, p_{ik}, is given by a beta distribution with parameters $((1-\theta)/\theta)\pi_i$ and $((1-\theta)/\theta)(1-\pi_i)$ (compare Roeder et al. 1998; Holsinger et al. 2002; Balding 2003). The corresponding phenotype frequencies are calculated as:

$$\begin{aligned}\gamma_{A1,ik} &= p_{ik}^2(1-f)+p_{ik}f+2p_{ik}(1-p_{ik})(1-f)\\ \gamma_{A2,ik} &= (1-p_{ik})^2(1-f)+(1-p_{ik})f\end{aligned},$$

Notice that the first-stage likelihood consists of a binomial sample of phenotypes.

For co-dominant marker data, the full conditional distribution follows the same form with two differences: (1) Loci may have more than two alleles and (2) All genotypes are distinguishable. To accomodate loci with more than two alleles we use a Dirichlet distribution, the multivariate generalization of a beta distribution, to describe the among-population distribution of allele frequencies (compare Roeder et al. 1998; Holsinger 1999; Balding 2003). Similarly, the first-stage likelihood with co-dominant marker data is constructed as a multinomial sample of genotypes rather than a binomial sample of phenotypes (see Holsinger 1999 for details).

Notice that these formulations assume independent sampling across both loci and populations. Independent sampling across loci ignores any statistical dependence associated with gametic or identity disequilibrium, but we expect the effect to be small. Independent sampling across populations is more problematic because high correlations among populations can occur in realistic genetic models, especially when the number of populations and the mutation rates are small (Fu et al. 2003). Models that account for among-population correlation are considerably more complex and are currently under development. As often happens in relatively complex hierarchical Bayesian models, closed form expressions for the posterior distributions of f and θ^B

are not available. `Hickory` uses standard Monte Carlo Markov Chain (MCMC) methods to approximate the posterior distributions of f and θ^B from either type of genetic data.

We used non-informative, uniform priors on all parameters in the analyses. As described in Holsinger et al. (2002), using informative beta priors has relatively little effect on posterior estimates of f and θ^B for either data set and relatively little effect on posterior estimates of f (details provided below). After a burn-in of 50,000 iterations, each sample chain consisted of 250,000 iterations, and we retained values at every 50th iteration for an MCMC sample size of 5000.

The output of a MCMC run is a large number of individual values for each of the parameters. While these points could be summarized with histograms, the true posteriors are continuous. Thus, a non-parametric continuous kernel density estimate is more appropriate. We use a Gaussian kernel density with parameters chosen to match the default parameters in the widely used statistical package R (Venables and Ripley 2002, p. 127). Plots of our kernel density estimates for f and θ^B suggested that a beta distribution with appropriate parameters might fit the posterior distributions well. To assess that possibility for f and θ^B in the full model, we calculated the Hellinger distance between a Gaussian kernel density estimate of each parameter and the corresponding beta density, where parameters of the beta density were estimated by matching its mean and variance with the mean and variance of the posterior distribution. Specifically, if $\bar{x}$ is the posterior mean for a parameter and s^2 its posterior variance, we chose the parameters of the Beta distribution, $Be(\nu,\omega)$ as

$$\begin{aligned} \nu &= \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right) \\ \omega &= (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right) \end{aligned} \qquad (1)$$

The Hellinger distance between two densities *f(x)* and *g(x)* is defined as (LeCam 1986, pp. 46-47):

$$H(f,g) = \int \left(f(x)^{1/2} - g(x)^{1/2} \right)^2 dx.$$

It ranges between 0 and 1, and can be interpreted as the fractional difference between the two densities. We approximate *H(f,g)* by a discrete sum evaluated at the 1024 points included in our kernel density estimate.

Lindley (1956) suggested that the entropy of a distribution is a natural measure of the "information" it provides about a random variable. If $f(\phi)$ is the probability density for random variable, then the entropy of f, $H(\phi)$, is given by

$$H(\phi) = -\int f(\phi) \log f(\phi) d\phi \quad .$$

In Bayesian inference, the difference between the entropy of the posterior and prior distributions of a parameter, $H(\phi|x) - H(\phi)$, is a widely used measure of the information provided about that parameter by the data, $I_E(\phi)$ (compare O'Hagan 1994, p. 87). In particular, if the posterior distribution for ϕ is Beta(α,β) and its prior distribution is Beta(ν,ω), then

$$\begin{aligned} I_E(\phi) = \ln \frac{Be(\nu,\omega)}{Be(\alpha,\beta)} &+ (\alpha - 1)\left(\Psi(\alpha) - \Psi(\alpha+\beta)\right) + (\beta - 1)\left(\Psi(\beta) - \Psi(\alpha+\beta)\right) \\ &- (\nu - 1)\left(\Psi(\nu) - \Psi(\nu+\omega)\right) - (\omega - 1)\left(\Psi(\omega) - \Psi(\nu+\omega)\right) \end{aligned} \qquad (2)$$

(compare Lindley 1957), where $Be(\nu,\omega)$ is the Beta function and $\psi(x)$ is the digamma function (Abramowitz & Stegun 1965). A uniform distribution corresponds to a beta distribution with both parameters equal to one. Thus, our reported values for $I_E(\phi)$ correspond to the difference in entropy between a Beta distribution with $\alpha = \beta = 1$ and a Beta distribution with parameters ν and ω chosen according to (1).

Results

We present point and interval estimates of f and θ^B for all of the models we consider in Table 2, and we present posterior plots and sample traces in Figure 1. We report results assuming a uniform prior, but both the point and interval estimates are relatively unaffected by the choice of prior (compare Holsinger et al 2002). The point estimates of f and θ^B from the RAPD data set, for example, are changed little (0.8774 *versus* 0.8362 for f, 0.2198 *versus* 0.2490 for θ^B) when we use beta priors with parameters as extreme as (48,11) for f and (33,55) for θ^B – values which were chosen to mimic the posterior distribution of the parameters in the allozyme data (see below). Because of the small number of loci in the allozyme data set, the estimates of θ^B are somewhat more sensitive to the choice of priors. Choosing beta priors to mimic the posterior RAPD data sets gives a point estimate of θ^B of 0.2505 (*versus* 0.3773 with a uniform prior), but the parameters of the prior for θ^B in this case (82,290) make the prior highly informative. A less informative prior with the same mean (8.2, 29) gives a point estimate for θ^B of 0.3336.

The estimate of θ^B presented here for RAPD data under the full model differs from the value originally reported in Holsinger et al. (2002). The data sets used are slightly different, but numerical error in our analytical routines also affected our original estimate. Further simulations (not shown) demonstrated that this error did not have a noticeable effect on our original reports of bias or root mean squared error. Standard convergence diagnostics (Brooks & Gelman 1998; Gelman & Rubin 1992; Raftery & Lewis 1992a; 1992b) indicate that sample chains for all models we consider had reached stationarity before we began constructing our posterior sample. Similarly, autocorrelation analyses showed that each parameter vector in our posterior sample can be regarded as an independent sample from the posterior distribution. Although point estimates of parameters are not greatly affected by high levels of autocorrelation, ensuring

independence of elements in the posterior sample is important for accurate estimates of credible intervals (Raftery and Lewis 1992b) and for our comparisons of f and θ^B between models. Although estimates of f from dominant marker data may be unreliable in certain data sets (Holsinger and Lewis 2003), analyses of alternative models in which f is fixed only by the prior specification either to a uniform (0,1) or a Beta with parameters matching the posterior distribution of f in the allozyme data produces estimates of θ^B that are barely distinguishable from those presented here.

Beta distributions with parameters matched to the mean and variance of the posterior distributions for f and θ^B provide good approximations to the exact posterior distribution for these data sets. Specifically, the Hellinger distance between a Gaussian kernel density estimate and the corresponding beta distribution was less than 0.007 (i.e., the overlap between the fitted beta distribution and the Gaussian kernel density estimate was greater than 99.3%) for all combinations of models, parameters, and data sets. Thus, we can use the analytical expression in (1) to calculate $I_E(x)$ for f and θ^B, which is provided for the full models in Table 3. Not surprisingly, the results show allozyme data provide substantially more information on f than the RAPD data, even with a much smaller number of loci. But precisely because the RAPD data set includes a much larger number of polymorphic loci than the allozyme data set, the RAPD data provide substantially more information on θ^B.

Model choice.

Spiegelhalter et al. (2002) introduced the Deviance Information Criterion (DIC) as a model choice criterion in Bayesian contexts. It is analogous to the more familiar Akaike Information Criterion (AIC; Akaike 1973) in that it combines a measure of model fit, $\overline{D}$ (-2 times the mean posterior log likelihood), with a measure of model complexity, pD. The preferred model is the

one that minimizes $\overline{D}+pD$, i.e., the one that represents the best compromise between model fit and number of parameters. *pD* is the "effective number of parameters", and it is calculated as $\overline{D}-\hat{D}$, where $\hat{D}$ is –2 times the log likelihood at the posterior mean. (The distribution of log likelihoods around the posterior mean is approximately χ^2, with degrees of freedom equal to the number of parameters and mean equal to the number of parameters; see Spiegelhalter et al. 2002 for details.) We present DIC calculations for the models we consider in Table 4. Because DIC is derived from the fit of a model to the data, it is only appropriate for comparisons of models as applied to the same data set. In our context, therefore, DIC comparisons are appropriate within the RAPD data set and within the allozyme data set, but they are not appropriate for comparisons between the data sets.

In classical applications, twice the difference between log likelihood of a nested model and that of a more complex model is distributed approximately as a χ^2 random variable with degrees of freedom equal to the difference in number of parameters (Mood et al. 1974). For models that differ in only one parameter, a significant difference corresponds, approximately, to a difference of 2 log likelihood units. Because DIC involves an average log likelihood, it is not surprising that Spiegelhalter et al. (2002) suggest that difference models that differ by only 1-2 DIC units deserve consideration, while those that differ by 3-7 DIC units have considerably less support.

In both data sets we analyze, the full model is preferred to alternatives with $f = 0$ or $\theta^B = 0$ – conclusively so in the case of the allozyme data. Both data sets therefore provide evidence of inbreeding within populations ($f > 0$) and of genetic differentiation among populations ($\theta^B > 0$). Nonetheless, a closer look at the DIC calculations reveals an important difference. With the allozyme data, $\overline{D}$ is substantially smaller for the full model than for the $f = 0$ model, i.e., the model with $f > 0$ fits the data substantially better than the model with $f = 0$. The difference in fit

to the data is almost entirely responsible for the difference in DIC, giving us considerable confidence that the full model should be preferred. With the RAPD data, however, $\overline{D}$ is about the same in the two models. The difference in DIC arises almost entirely as a result of differences in model dimension, indicating that the full model should be only weakly preferred.

A weak preference for the full model in the RAPD data seems paradoxical in light of a 95% credible interval for *f* whose lower limit is greater than 0.5. Nonetheless, it is consistent with prior intuition suggesting that it should be difficult to recover reliable information about inbreeding from a dominant marker. It is also consistent with our calculations of $I_E(x)$, which show that the RAPD data provide substantially less information about *f* than the allozyme data (Table 3). It appears that the weak identifiability of *f* with dominant marker data (Holsinger et al. 2002) is responsible for the apparent inconsistencies between conclusions about *f* based on DIC comparisons and those based on credible intervals. In light of these consistencies, we recommend that estimates of *f* derived from dominant marker data be regarded with considerable caution (see also Holsinger and Lewis 2003).

Comparing estimates

While exact permutation and approximate bootstrapping methods for testing for the presence of inbreeding within populations or for the presence of genetic differentiation among populations are well known (e.g., Rousset & Raymond 1995; Goudet et al 1996; Rousset 1997) methods for comparing different, non-zero estimates have not been developed. Fortunately, making such comparisons within a Bayesian framework is straightforward. Let $P(f_A|x_A)$ be the posterior distribution of *f* as determined from data set A, and let $P(f_B|x_B)$ be the posterior distributiion of *f* as determined from data set B. Different sets of data are used to estimate f_A and f_B, but if f_A and f_B reflect the same evolutionary processes the differences in $f_A - f_B$ should not be distinguishable

from zero. Let f_{Ai} be the i^{th} MCMC sample from $P(f_A|x_A)$, and let f_{Bi} be the i^{th} MCMC sample from $P(f_B|x_B)$. We can approximate the posterior distribution of $f_A - f_B$ to an arbitrary degree of accuracy from a large sample of $f_{Ai} - f_{Bi}$. If the 100(1-α)% credible interval is strictly positive, we have evidence (at the α% level) that f_A is larger than f_B. If it is strictly negative, we have evidence that f_A is smaller than f_B. If the 100(1-α)% credible interval on the difference includes zero, then we have no evidence that one inbreeding coefficient is larger than the other. Clearly, we can use the same approach to compare estimates of θ^B derived from different data sets. Ayres and Balding (1998) and Shoemaker et al. (1998) employ similar strategies in developing Bayesian methods for assessing departures from Hardy-Weinberg.

Notice that the simpler approach of requiring non-overlapping credible intervals would be overly conservative. Suppose, for example, that our estimate of f_A is less than our estimate of f_B. For the 100(1-α)% credible intervals not to overlap then $P(f_A > p|x_a) < \alpha/2$ and $P(f_b < p\ |x_b) < \alpha/2$ must hold for some p. The probability that both hold, assuming the data sets are independent, is $\alpha^2/4$. Thus, the credibility level associated with $f_A < f_B$, would be $1-\alpha^2/4$ if non-overlap of credibility intervals were required.

While the point estimate of f derived from both data sets in *Platanthera leucophaea* is quite similar and the 95% credible intervals are broadly overlapping, the point estimate of θ^B derived from the RAPD data set is substantially smaller than the estimate derived from the allozyme data set**,** and the 95% credible intervals are non-overlapping (Table 2, Figure 2). Thus, both sets of loci appear to indicate similar levels of within-population inbreeding, but also to indicate different levels of among-population differentiation. Posterior comparisons of the difference in parameters between the data sets bear these impressions out. The posterior mean of $f_{RAPD} - f_{Allozyme}$ is 0.0640, but its 95% credible interval broadly overlaps zero: (-0.2745, 0.2494). The posterior

mean of $\theta^B_{RAPD} - \theta^B_{Allozyme}$, on the other hand, is –0.1569, and its 95% credible interval does not include zero: (-0.2662, -0.0535).

Discussion

The analyses presented in this paper extend previous work on the application of Bayesian methods to analysis of population genetic structure in four ways. First, they show that for these data the posterior distribution of f and θ^B are well approximated by a beta distribution in which the parameters of the beta distribution are chosen to match the mean and variance of the posterior distribution (see Equation 1), and our limited experience with other data sets suggests that this is likely to be a general result. We encourage investigators to report the mean and variance of the posterior distributions so that later investigators can take advantage of existing data to use informative priors when appropriate (Holsinger et al. 2002). Second, the difference between the entropy of the prior distribution of a parameter and the entropy of its posterior distribution, $I_E(x)$, is a useful summary of the amount of information about a paramter provided by the data. Although we do not explore use of this measure in sampling design here, it is apparent that preliminary studies using simulated data sets could help investigators to choose an allocation of samples within and among populations that provide the most information possible about the parameters. Third, we illustrate the use of DIC as a model choice criterion in the context of population structure analysis, and we point out that investigators should pay attention not only to DIC but also to its components. While DIC itself suggested strong support for a model including both inbreeding and among population differentiation with the RAPD data (a difference of approximately 15 units), a closer examination showed that it was preferred primarily because of its reduced complexity (pD), not because a noticeable improvement in average fit. Under such

circumstances, the model with lower DIC should be preferred only weakly, if they are preferred at all. Finally, we illustrate how to determine whether two, non-zero estimates of θ (or f) can be compared to determine whether differences in point estimates derived from different data sets reflect the operation of different evolutioanry processes or of sampling error.

Our analysis of allozyme and RAPD data sets from *Platanthera leucophaea* confirm earlier results suggesting that there is substantial inbreeding within populations and substantial genetic differentiation among populations. We find, however, that the degree of population differentiation depends on the data set used for analysis. Because exactly the same populations (and nearly all of the same individuals) are included in both the RAPD and the allozyme data sets, the observed differences in θ^B are not likely to reflect differences in migration rates or demographic history. Instead, they are likely to reflect differences either in the mutational processes by which variation arises at these loci or in the patterns of natural selection to which they are subject. In models incorporating drift, migration, and mutation (see Fu et al. 2003 for a recent review), higher rates of mutation are associated with lower amounts of among-population differentiation. Thus, the smaller value of θ^B associated with RAPD loci may reflect a higher rate of mutation at RAPD loci than at allozyme loci (see Flint et al. 1999 for a similar example involving human minisatellite data). Additional analyses of the RAPD data set (Fu, Dey, and Holsinger unpublished) have also shown that there are detectable differences in θ^B among loci *within* the RAPD data set. Both sets of results illustrate, yet again, that migration-rate estimates derived from *F*-statistics are highly problematic (compare Whitlock & McCauley 1999). Migration-rate estimates from genetic data must take into account mutational processes if they are to be reliable (Beerli & Felsenstein 1999).

Acknowledgments

We are grateful to Dipak Dey and Paul Lewis for discussion, advice, and for helpful comments on earlier versions of this paper.

Literature cited

Abramowitz M, Stegun IA (eds.) (1965) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications Inc., New York.

Ayres, KL, Balding DJ (1998) Measuring departures from Hardy-Weinberg: a Markov Chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity* **80**, 769-777.

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: *2nd International Symposium on Information Theory* (eds Petrov BN, Csáki F), pp. 267-281. Akadémiai Kiadó, Budapest.

Balding DJ (2003) Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology* **63**, 221-230.

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

Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152,** 763-773.

Brooks S, Gelman A (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434-455.

Excoffier L (2001) Analysis of population subdivision. In: *Statistical Genetics* (eds. Balding DJ, Bishop M, Cannings C), pp. 271-307. John Wiley & Sons, Chichester.

Flint, J, Bond J, Rees DC, Boyce AJ, Roberts-Thomson JM, Excoffier L, Clegg JB, Beaumont MA, Nichols RA, Harding RM (1999) Minisatellite mutational processes reduce F_{st} estimates. *Human Genetics* **105**, 567-576.

Fu R, Gelfand AE, Holsinger KE (2003) Exact moment calculations for genetic models with migration, mutation, and drift. *Theoretical Population Biology* **63***,* 231-243.

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-511.

Goudet JM, Raymond M, de Meeus T, Rousset F (1996) Testing differentiation in diploid populations. *Genetics* **144**, 1933-1940.

Holsinger KE (1999) Analysis of genetic diversity in geographically structured populations: a Bayesian perspective. *Hereditas*, **130**, 245-255.

Holsinger KE, Lewis PO (2003) Hickory v0.8. http://darwin.eeb.uconn.edu/hickory/hickory.html

Holsinger KE, Lewis PO, Dey DK (2002) A Bayesian approach to inferring population structure from dominant markers. *Molecular Ecology,* **11**, 1157-1164.

LeCam L (1986) *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, Berlin.

Lindley DV (1956) On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, **35**, 1622-1643.

Lindley DV (1957) Binomial sampling schemes and the concept of information. *Biometrika*, **44**, 179-186.

Malécot G (1948) *Les Mathématiques D l'Hérédité*. Masson et Cie, Paris.

Mood AM, Graybill FA, Boes DC (1974) *Introduction to the Theory of Statistics*, 3rd ed. McGraw-Hill, New York, NY.

O'Hagan A (1994) *Kendall's Advanced Theory of Statistics*, vol 2B. *Bayesian Inference*. Edward Arnold, London.

Raftery AL, Lewis S (1992a) One long run with diagnostics: implementation strategies for Markov Chain Monte Carlo [Comment]. *Statistical Science*, **7**, 493-497.

Raftery AL, Lewis S (1992b) How many iterations in the Gibbs sampler? In: *Bayesian Statistics 4* (eds Bernardo JM, Berger JO, Dawid AP, Smith AFM), pp. 763-774. Oxford University Press, Oxford.

Roeder K, Escobar M, Kadane J, Balazs I (1998) Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika*, **85**, 269-287.

Rousset F (1997) Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics* **145,** 1219-28.

Rousset F, Raymond, M (1995) Testing heterozygote excess and deficiency. *Genetics* **140**, 1413-1419.

Shoemaker, JI, Painter I, Weir, BS (1998) A Bayesian characterization of Hardy-Weinberg disequilibrium. *Genetics* **149**, 2079-2088.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, **64**, 583-639.

United States Fish & Wildlife Service (USFWS). 1999. Eastern prairie fringed orchid, *Platanthera leucophaea* (Nuttall) Lindley Recovery Plan. Region 3, Fort Snelling, Minnesota, USA.

Wallace LE (2002) Examing the effects of fragmentation on genetic variation in *Platanthera leucophaea* (Orchidaceae): inferences from allozyme and random amplified polymorphic DNA markers. *Plant Species Biology,* **17,** 37-49.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*, 4th ed. Springer-Verlag, Ne York, NY

Weir, BS and Cockerham, CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38**,1358-1370.

Weir BS, Hill WG (2002) Estimating *F*-statistics. *Annual Reviews of Genetics* **36**,721-750.

Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: F_{ST} is not equal to 1/(4*Nm* + 1). *Heredity*, **82**, 117.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323-354.

Author information box

Kent Holsinger is Professor of Biology and Adjunct Professor of Statistics at the University of Connecticut. His research concerns the evolution of plant reproductive systems and the development of statistical methods in population genetics and population ecology. Lisa Wallace is a postdoctoral fellow at the University of South Dakota. She is interested in the evolution of populations and mechanisms of speciation in plants as well as conservation genetics of rare plants. Software (source code in C++, Windows, and Linux executables) implementing the analyses described here is distributed under terms of the GNU General Public License at http://darwin.eeb.uconn.edu/hickory/hickory.html.

Figure Legends

Figure 1. Posterior distributions and sample traces for f and θ^B estimated from the RAPD data (**a**) and from the allozyme data (**b**). The dashed line on the density plots is a Gaussian kernel density based on 1024 points. The solid line is a beta density with parameters chosen to match the posterior mean and variance of each parameter. The dashed line is most easily visible in the posterior density for f estimated from the RAPD data. The scale on the y-axis (labeled "Probability density") is chosen so that the area underneath it integrates to one. **a.** For f, $(\nu,\omega) = (5.606, 0.7833)$. For θ^B, $(\nu,\omega) = (81.84, 290.4)$. **b.** For f, $(\nu,\omega) = (48.67, 11.05)$. For θ^B, $(\nu,\omega) = (33.35, 55.06)$.

Figure 2. Box plots of the posterior densities for f and θ^B. The line in the box is at the position of the median, the lower boundary is at the lower quartile, the upper boundary of the box is at the upper quartile, and the whiskers extend to the most extreme data points no more than 1.5× the interquartile range beyond the box. Individual points outside the interquartile range are plotted.

Table 1. Sample sizes for the allozyme and RAPD data sets from *Platanthera leucophaea* (see Wallace 2002).

Population	Census Size	Sample Size	
		Allozyme	RAPD
Meadow	13	13	13
Medway	33	12-15	13
Conservation	24	24	24
Cemetery	17	10-13	17
Metzger	104	25	25
Pickerel	1065	36-39	40
Yandota	118	17-23	23

Table 2. Posterior means, standard deviation, and 95% credible intervals of f and θ^B for RAPD and allozyme data in *Platanthera leucophaea* under three alternative models.

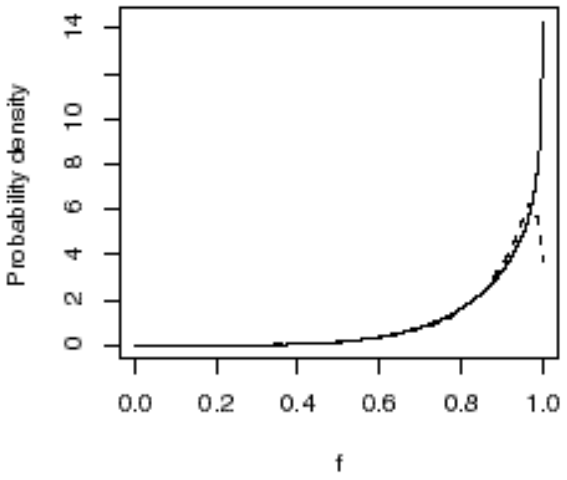
		f			θ^B		
Model		Mean	s.d.	95% c.i.	Mean	s.d.	95% c.i.
RAPD	Full	0.8774	0.1207	(0.5540, 0.9971)	0.2198	0.0214	(0.1797, 0.2646)
	$f = 0$				0.1644	0.0176	(0.1327, 0.2008)
	$\theta = 0$	0.9078	0.0924	(0.6661, 0.9977)			
Allozyme	Full	0.8149	0.0498	(0.7061, 0.8986)	0.3772	0.0513	(0.2830, 0.4789)
	$f = 0$				0.4135	0.0504	(0.3187, 0.5169)
	$\theta = 0$	0.9268	0.0192	(0.8855, 0.9605)			

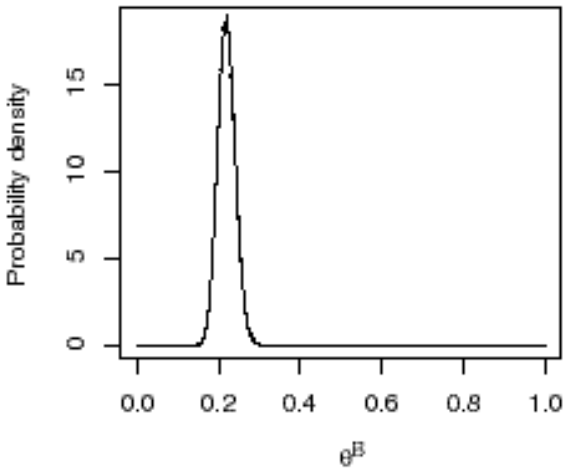
Table 3. Information provided by the data ($I_E(x)$) for each parameter in the full model when applied to RAPD and allozyme data from *Platanthera leucophaea*. The greater the value of $I_E(\phi)$, the more information about the parameter provided by the data.

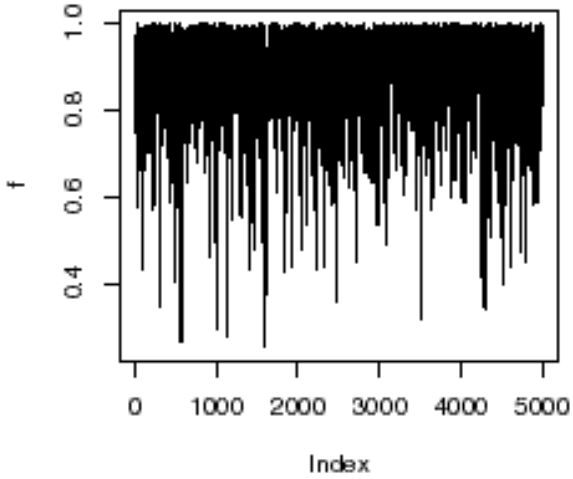
Dataset	Parameter	$I_E(\phi)$
RAPD	f	1.1135
	θ^B	2.4254
Allozyme	f	1.5954
	θ^B	1.5530

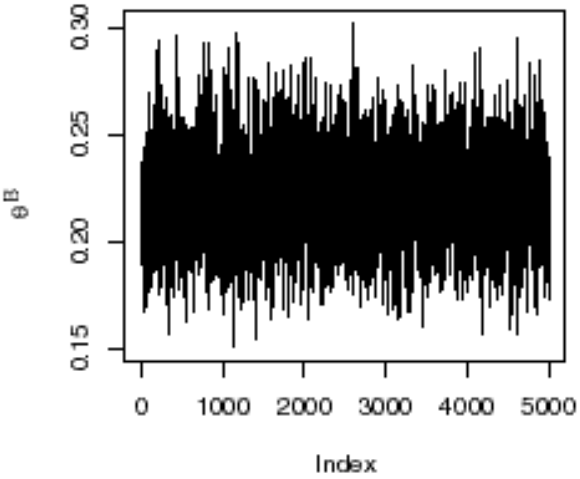
Table 4. DIC calculations for RAPD and allozyme data in *Platanthera leucophaea* under three alternative models. The smallest DIC for each data set is highlighted in bold.

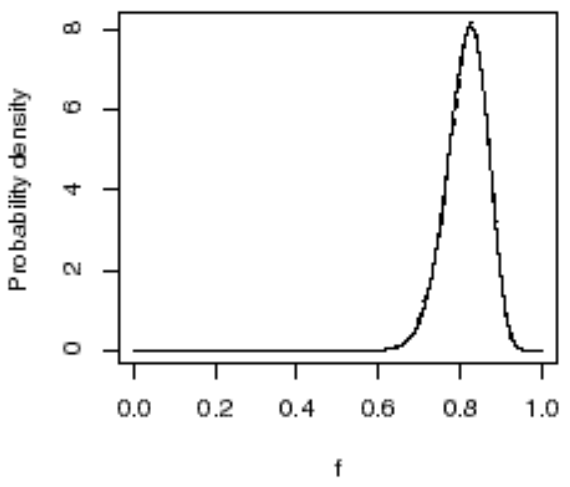
Model		$\overline{D}$	$\hat{D}$	*pD*	DIC
RAPD	Full	1033.8745	826.1775	207.6970	**1241.5715**
	$f = 0$	1032.7101	809.0914	223.6187	1256.3288
	$\theta = 0$	2174.1822	2114.1996	59.9826	2234.1647
Allozyme	Full	155.2735	140.9082	14.3653	**169.6387**
	$f = 0$	240.3963	226.4958	13.9005	254.2968
	$\theta = 0$	560.8112	553.7171	7.0941	567.9053

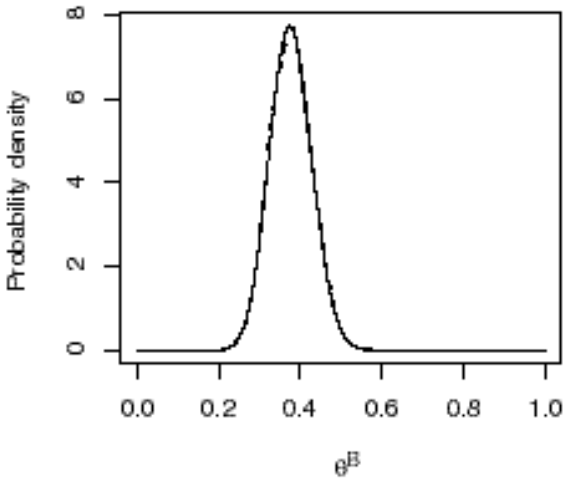

Probability density
0
2
4
6
8
10
14
0.0
0.2
0.4
0.6
0.8
1.0
f

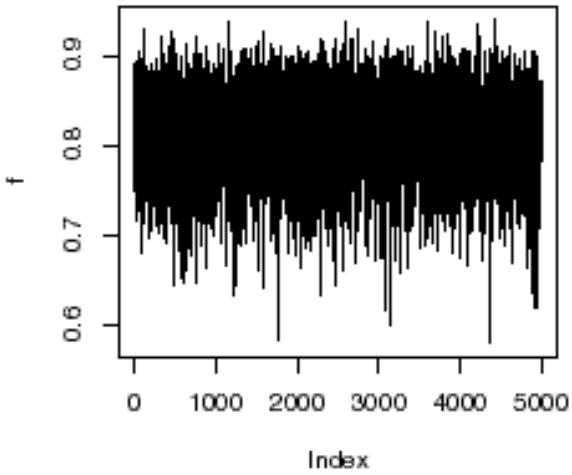

Probability density
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
θB

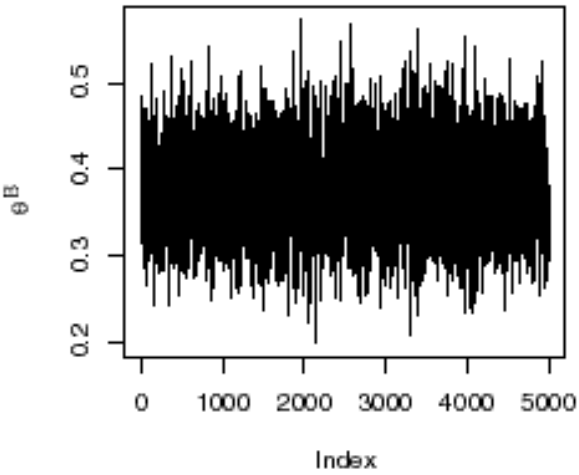

f
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
Index

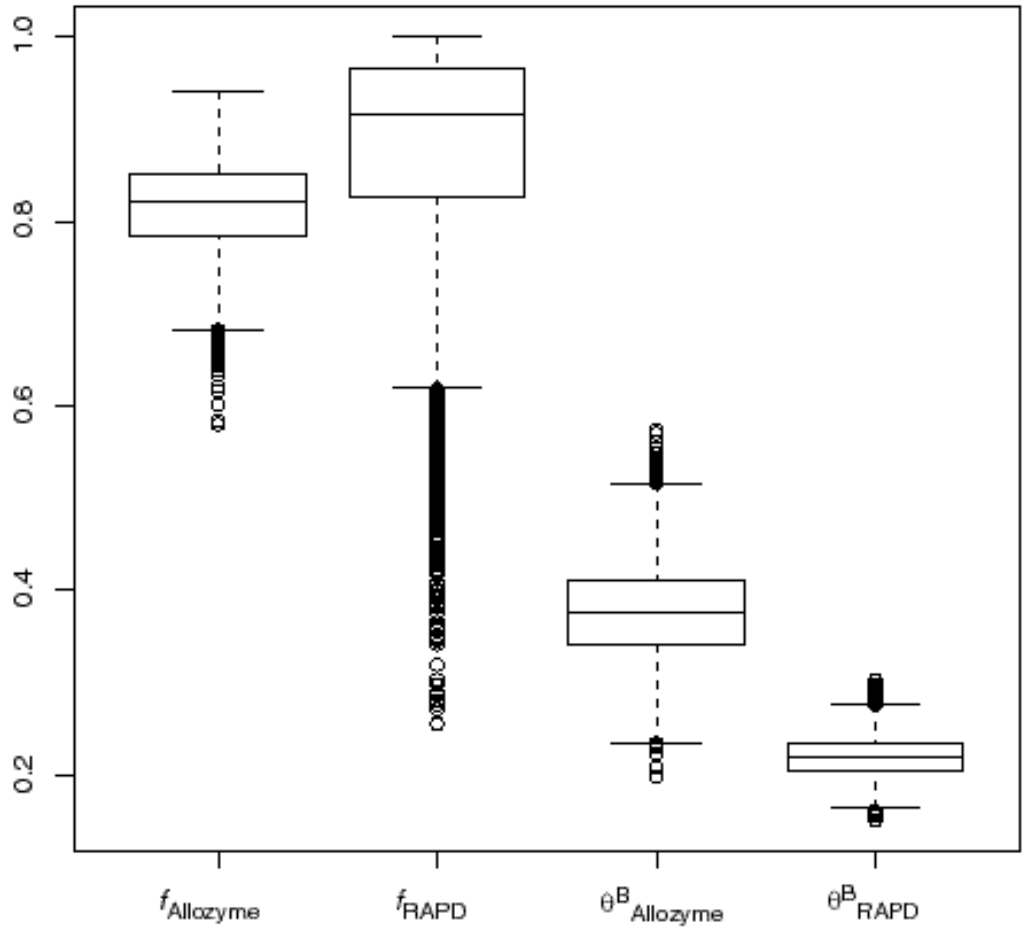

θB
0.15
0.20
0.25
0.30
0
1000
2000
3000
4000
5000
Index


Probability density
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
f


Probability density
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
θ^B


f
0.6
0.7
0.8
0.9
0
1000
2000
3000
4000
5000
Index


θ^B
0.2
0.3
0.4
0.5
0
1000
2000
3000
4000
5000
Index


1.0
0.8
0.6
0.4
0.2
$f_{Allozyme}$
f_{RAPD}
$\theta^{B}_{Allozyme}$
θ^{B}_{RAPD}