7-1-2002

A Bayesian approach to inferring population structure from dominant markers

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

Paul O. Lewis
University of Connecticut, paul.lewis@uconn.edu

Dipak Dey
University of Connecticut, DIPAK.DEY@uconn.edu

A Bayesian approach to inferring population structure from dominant markers

Kent E. Holsinger
Paul O. Lewis
Department of Ecology & Evolutionary Biology, U-3043
University of Connecticut
Storrs, CT 06269-3043

Dipak K. Dey
Department of Statistics, U-3120
University of Connecticut
Storrs, CT 06296-3120

Abstract

Molecular markers derived from PCR amplification of genomic DNA are an important part of the toolkit of evolutionary geneticists. RAPDs, AFLPs, and ISSR polymorphisms allow analysis of species for which prior DNA sequence information is lacking, but dominance makes it impossible to apply standard techniques to calculate F-statistics. We describe a Bayesian method that allows direct estimates of F_{st} from dominant markers. In contrast to existing alternatives, we do not assume prior knowledge of the degree of within-population inbreeding. In particular, we do not assume that genotypes within populations are in Hardy-Weinberg proportions. Our estimate of F_{st} incorporates uncertainty about the magnitude of within-population inbreeding. Simulations show that samples from even a relatively small number of loci and populations produce reliable estimates of F_{st}. Moreover, some information about the degree of within population inbreeding (F_{is}) is available from data sets with a large number of loci and populations. We illustrate

the method with a reanalysis of RAPD data from 14 populations of a North American orchid, *Platanthera leucophaea*.

Introduction

Since Lewontin and Hubby (1966) introduced evolutionary geneticists to allozyme electrophoresis almost 40 years ago, molecular methods have been a vital source of data for evolutionary analysis. Advances in molecular technology in the last decade, notably the introduction of PCR-based DNA amplification, have resulted in the development of many new methods for assessing genetic diversity. In particular, random amplified polymorphic DNA (RAPD: Williams *et al.* 1990), amplified fragment length polymorphism (AFLP: Vos *et al.* 1995), and inter-simple sequence repeat polymorphism (ISSR: Wolfe and Liston 1998) allow investigators to obtain large amounts of data on variation within and among populations without detailed prior knowledge of DNA sequences within the species being studied.

Because these markers are based on PCR amplification, however, a diploid individual need carry only one copy of the sequence necessary for an amplification product to be produced. A homozygote for the "null" allele at a particular locus will not produce a band, but both a heterozygote and a homozygote at that locus will. As a result, the fundamental data available from a population at any given locus is the number of individuals with a band corresponding to that locus and the number of individuals lacking that band.

Investigators collecting genetic data from multiple populations are usually interested in assessing the degree of genetic differentiation among those populations. When a genetic marker is co-dominant all genotypes are distinguishable from one another. Estimating allele frequencies within populations and the variance of allele frequencies among populations is straightforward. When a genetic marker is dominant, however, estimating allele frequencies requires prior knowledge of the inbreeding coefficient. As a result, current approaches to partitioning genetic diversity as assessed with dominant marker data involve either assuming that the inbreeding coefficient within populations is known (Lynch and Milligan 1994; Zhivotovsky 1999) or treating the multilocus phenotype as a haplotype and using a similarity index (e.g., Nei and Li 1979) or Euclidean distance (Schneider et al. 2000) to describe distances among haplotypes in an analysis of molecular variance (AMOVA:

Excoffier et al. 1992; Isabel et al. 1999)

In this paper we present a Bayesian hierarchical model appropriate for analysis of data derived from dominant markers. The method we present is constructed in terms of the classical F-statistics of Wright (1951) and Malécot (1948), and it represents a special case of a more general Bayesian approach to analysis of hierarchical data in genetics described by Holsinger (1999). Although our method does not provide precise estimates of inbreeding within populations, it allows us to incorporate the effect of our uncertainty about the magnitude of inbreeding into our estimates of F_{st}. We present simulation results showing that our estimator performs well, even with relatively small numbers of loci and populations. We also illustrate how the method can be applied using data derived from a recent survey of RAPD variation in a North American orchid, *Platanthera leucophaea* (Wallace 2000, submitted).

Materials and methods

F-statistics as defined by Wright (1951) and Malécot (1948) are the most widely used method for describing the hierarchical structure of genetic data derived from multiple populations. F_{is} is defined as "the average over all [populations] of the correlation between uniting gametes relative to those of their own population," and F_{st} is defined as "the correlation between random gametes within [populations], relative to gametes of the total [set of populations]" (Wright 1969, p. 294). When considering one locus with two alleles, F_{is} is equal to the average within-population inbreeding coefficient,

$$F_{is} = \sum_{k=1}^{K} \left(1 - \frac{H_k}{2p_k(1-p_k)} \right) \quad , \tag{1}$$

where H_k is the frequency of heterozygotes and p_k is the allele frequency in population k. F_{st} is equal to the variance in allele frequency among populations divided by the maximum possible variance given the mean allele frequency across all populations, $\bar{p}$,

$$F_{st} = \frac{\mathrm{Var}(p)}{\bar{p}(1-\bar{p})} \quad . \tag{2}$$

Weir and Cockerham (1984, see also Cockerham 1969) introduced an approach for estimating F_{is} and F_{st} derived from analysis of variance. In their

notation f corresponds to F_{is}, and θ corresponds to F_{st}. We follow the Weir and Cockerham notation because the method we propose is directly analagous to the random-effects model of population sampling underlying their analysis of variance approach (compare Holsinger 1999; Roeder et al. 1998; Weir 1996).

The statistical model

The sample consists of data on genetic variation in K populations at I loci, each locus having 2 alleles A_1 and A_2. We assume , without loss of generality, that A_1 is dominant to A_2 at every locus. Let $x_{A1,ik}$ be the frequency of the dominant phenotype at locus i in population k, and let $n_{A1,ik}$ be the number of dominant phenotypes and $n_{A2,ik}$ be the number of recessive phenotypes in the sample at locus i in population k. If we assume that phenotypes are sampled at random across loci, which corresponds to assuming that the magnitudes of gametic and identity disequilibrium within populations are negligible, then the likelihood of the sample is

$$P(\mathbf{n}|\mathbf{x}_{A1}) \propto \prod_{i=1}^{I} \prod_{k=1}^{K} x_{A1,ik}^{n_{A1,ik}} x_{A2,ik}^{n_{A2,ik}} \quad , \tag{3}$$

where

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

and $f = F_{is}$ (see equation (1)). To incorporate the hierachical structure implicit in the data, we assume that frequency distribution of p_{ik} among all populations of interest (including those not sampled) is given by a Beta distribution with parameters α_i and β_i, $i = 1, \ldots, I$. For neutral genetic markers, the stationary distribution of allele frequency for a single locus subject to drift, migration, and mutation is a Beta distribution (Crow and Kimura 1970; Ewens 1979).

Let $\alpha_i = ((1-\theta)/\theta)\pi_i$ and $\beta_i = ((1-\theta)/\theta)(1-\pi_i)$, be the parameters of this Beta distribution. Its mean is then

$$\begin{aligned} \frac{\alpha_i}{\alpha_i + \beta_i} &= \frac{((1-\theta)/\theta)\pi_i}{((1-\theta)/\theta)\pi_i + ((1-\theta)/\theta)(1-\pi_i)} \\ &= \pi_i \end{aligned}$$

and its variance is

$$\begin{aligned} \frac{\alpha_i \beta_i}{(\alpha_i + \beta_i)^2(\alpha_i + \beta_i + 1)} &= \frac{\pi_i(1-\pi_i)}{(\alpha_i + \beta_i + 1)} \\ &= \pi_i(1-\pi_i)\theta \quad . \end{aligned}$$

Since $F_{st} = \mathrm{Var}(p)/\left(\bar{p}(1-\bar{p})\right)$ (see equation (2)), it follows that $\theta = F_{st}$ in this formulation, if we assume that all loci show the same pattern of within and among population diversity (Holsinger 1999; Roeder et al. 1998).

Notice that p_{ik} and f enter the likelihood only through $x_{A1,ik}$ and $x_{A2,ik}$. As a result, p_{ik} and f are not identifiable in a likelihood context, and likelihood methods or other classical methods of analysis are possible only when either f or the p_{ik} are specified before the analysis. In contrast, Bayesian analysis is possible by specifying separate priors for π_i, f, and θ. Prior information may either be vague and non-informative, or it may use information from previous comparable studies to refine estimates provided from newly-collected data.

Specifically, if $P(\pi_i)$ is the prior distribution for π_i, $P(\theta)$ is the prior distribution for θ, and $P(f)$ is the prior distribution for f, the posterior probability distribution for p_{ik}, π, θ, and f is given by

$$P(\mathbf{p}, \pi, \theta, f | \mathbf{n}_{A1}, \mathbf{n}_{A2}) \propto \left\{ \prod_{i=1}^{I} \left\{ \prod_{k=1}^{K} x_{A1,ik}^{n_{A1,ik}} x_{A2,ik}^{n_{A2,ik}} P(\mathbf{x}_{ik} | \pi_i, \theta, f) \right\} P(\pi_i) \right\} P(\theta) P(f) \quad . \tag{4}$$

The posterior mean of θ provides a point estimate of F_{st}, and the posterior mean of f provides a point estimate of F_{is}.

Analytical expressions for the posterior distributions of θ and f derived from (4) are not available in standard form, but they can be numerically approximated through the use of Markov chain Monte Carlo (MCMC) simulation. Because the full conditional distributions for some of the parameters in (4) cannot be sampled directly, we use a single-component Metropolis-Hastings algorithm (Gilks et al. 1996) in our MCMC implementation.

Validation of the method

MCMC methods depend both on convergence of the Markov chain to its stationary distribution and on independence of samples taken from the chain once it has converged. We take samples from the chain only after discarding a fixed number of initial samples (the *burn-in*) to ensure that the chain has converged to its stationary distribution. Once the chain has converged, we

retain samples from it only at fixed intervals (the *thin*) to ensure that the samples are independent of one another.

We assessed convergence of the MCMC sampler in a series of preliminary simulations using a battery of standard diagnostic tests (Brooks and Gelman 1998; Gelman and Rubin 1992; Raftery and Lewis 1992a,b). We then simulated data derived from a broad range of realistic values for F_{is} and F_{st} and assessed the performance of f and θ as estimates of these parameters by computing three summary statistics for every simulated combination of parameters: (1) *bias*, which is the average difference between the known parameter value and estimated parameter value; (2) *root mean squared error*, which is the square root of the average squared difference between the known parameter value and the estimated parameter value; and (3) *realized coverage*, which is the fraction of 95% credible intervals that include the true parameter value.

Results

Convergence of the sampler

Reliable point estimates for F_{is} and F_{st} can be obtained with a burn-in of 5000 iterations and sampling run of 25,000 iterations from which only every fifth sample is retained for posterior calculations. Reliable estimates of the 95% credible intervals, however, requires a burn-in of 50,000 iterations and a sampling run of 250,000 iterations from which every fiftieth sample is retained for posterior calculations (data not shown). Figure 1 illustrates results derived from one run of the sampler (burn-in: 50,000; sample: 250,000; thin: 50) with simulated data from 10 loci and 5 populations produced when $F_{is} = 0.1$ and $F_{st} = 0.1$.

Inspection of the figure illustrates that even with small amounts of data accurate and precise estimates of F_{st} are possible. For this example, $\theta = 0.16$ and its 95% credible interval is [0.093,0.24]. Estimates of F_{is}, not surprisingly, are uninformative. For this example, $f = 0.41$ and its 95% credible interval is [0.027,0.93]. Panels (c) and (d) illustrate that the sampler thoroughly explores the relevant parts of parameter space, and panels (e) and (f) illustrate that thinning the sample has successfully produced nearly independent samples from the posterior distribution. Formal convergence diagnostics (Brooks and Gelman 1998; Gelman and Rubin 1992; Raftery and Lewis 1992a,b) con-

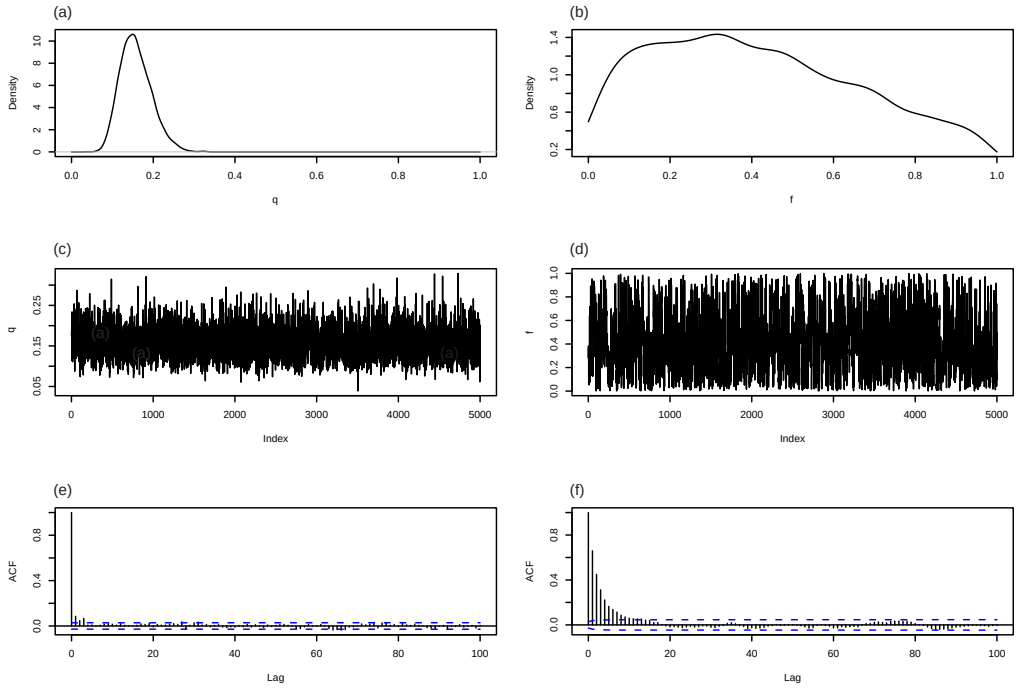

Figure 1: Convergence of the sampler and posterior densities for θ and f with a sample data set generated with $F_{st} = 0.1$ and $F_{is} = 0.1$. (a) Posterior density of θ. (b) Posterior density of f. (c) Sample history for θ. (d) Sample history for f. (e) Autocorrelation for θ. (f) Autocorrelation for f.

firm these subjective impressions.

Performance of the estimtes

To assess performance of θ and f we performed a series of simulations in which we generated data with known values of F_{is} and F_{st} (0.01, 0.05, 0.1, 0.25, 0.5, 0.9). The mean allele frequency at each locus, π_i, was chosen at random from a Beta(2,2). We chose this distribution because we expect investigators to focus their attention on loci with moderate to high amounts of variability. 95% of the allele frequencies chosen at random from this distribution will lie between 9.4% and 90.6%. For a population with genotypes in Hardy-Weinberg proportions that corresponds to dominant band frequencies between 18% and 99%. We chose a sample size of 25 individuals per

population as representative of a moderate degree of within-population sampling effort, and we investigated performance of the estimates both with 10 loci and 5 populations and, for a smaller number of parameter combinations, with 50 loci and 15 populations. We used non-informative (uniform) priors for θ, f, and π_i in all of our analyses.

The results from 100 independently generated data sets for each set of parameter conditions (shown in Table 1) illustrate that θ provides reliable estimates of F_{st} even when the amount of data collected is relatively small. Both the bias and root mean squared error of θ are consistently smaller than about 3%. Moreover, the 95% credible intervals for θ cover the true value of F_{st} at least 90% of the time. As expected, the bias and root mean squared error of the estimates are substantially smaller when samples are available from more loci and more populations, and the 95% credible intervals are also narrower.

Although these results illustrate that good estimates of F_{st} are possible, they also illustrate that very little information about F_{is} is available from dominant markers, especially in data sets with a small number of loci and populations. The posterior mean of f differs only a little from its prior mean (0.5), and the credible intervals cover most of the range in simulated data sets with 10 loci and 5 populations (data not shown). Notice especially that the realized coverage of the credible intervals for F_{is} is less than 40% with 10 loci and 5 populations when its true value is 0.01. With larger amounts of data, however, the bias and root mean squared error of f are smaller, indicating that some inference about F_{is} is possible even with dominant markers (see also the analysis of data from *Platanthera leucophaea* in the next section).

Information on f comes from constraints placed on it through the assumption of a common θ across loci. Although p_{ik} and f are not identifiable in a likelihood context, the posterior distribution for p_{ik} is constrained by the data. Values for p_{ik} falling outside the interval $[x_{A1,ik}, 1-\sqrt{x_{A2,ik}}]$ have a low likelihood and a small posterior probability, regardless of the value of f. The posterior distribution of θ depends on the variance in allele frequency among populations. Thus, it is constrained to values consistent with a high likelihood for the p_{ik}. For any single locus, the posterior distribution of f is little constrained by the data, because for any value of p_{ik} in $[x_{A1,ik}, 1-\sqrt{x_{A2,ik}}]$ a value of f can be found that provides a high likelihood and a large posterior probability. With many loci, however, the value of θ is tightly constrained, because of the constraints on p_{ik} imposed by the observed within-population

		θ			f		
F_{st}	F_{is}	Bias	Root MSE	Coverage	Bias	Root MSE	Coverage
0.01[1]	0.01	0.0105	0.0149	0.93	0.2969	0.3258	0.33
	0.05	0.0095	0.0133	0.94	0.3303	0.3259	0.83
0.05	0.01	0.0219	0.0320	0.87	0.3139	0.3387	0.37
	0.05	0.0182	0.0295	0.91	0.3252	0.3606	0.73
0.10	0.10	0.0268	0.0418	0.96	0.2997	0.3293	0.96
	0.25	0.0165	0.0354	0.91	0.1823	0.2394	0.99
	0.50	0.0064	0.0310	0.97	0.0474	0.1495	1.00
	0.90	0.0059	0.0269	0.96	-0.2476	0.2760	0.96
0.25	0.10	0.0465	0.0737	0.87	0.3344	0.3536	0.98
	0.25	0.0299	0.0619	0.93	0.2092	0.2426	1.00
	0.50	0.0030	0.0518	0.93	0.0439	0.1321	1.00
	0.90	0.0071	0.0530	0.96	-0.2975	0.3209	0.96
0.50	0.10	0.0631	0.0971	0.84	0.0348	0.4401	0.98
	0.25	0.0409	0.0814	0.88	0.0302	0.2881	1.00
	0.50	0.0146	0.0723	0.92	0.0551	0.0877	1.00
	0.90	0.0077	0.0717	0.92	-0.3081	0.3173	0.99
0.90	0.10	-0.0551	0.0631	0.91	0.4231	0.4239	1.00
	0.25	-0.0466	0.0531	0.90	0.2805	0.2820	1.00
	0.50	-0.0571	0.0670	0.87	0.0299	0.0340	1.00
	0.90	-0.0508	0.0583	0.86	-0.3690	0.3702	1.00
0.01[2]	0.01	0.0018	0.0031	0.93	0.1231	0.1410	0.61
0.10	0.10	0.0098	0.0145	0.85	0.1421	0.1785	0.87
0.50	0.50	0.0171	0.0598	0.95	0.0751	0.2698	0.80
0.90	0.90	-0.0063	0.0070	0.84	-0.2314	0.2456	1.00

[1] 10 loci, 5 populations

[2] 50 loci, 15 populations

Table 1: Peformance of θ and f. F_{st} and F_{is} are the known parameter values used to generate the simulated data sets with 25 individuals sampled from every population. All statistics are based on 100 independently generated data sets.

phenotype frequencies. As a result, estimates for f are also constrained, although the constraints on f are much looser than those on θ, as our results indicate (see O'Hagan 1994, p. 72 for an excellent discussion of nonidentifiability in a Bayesian context).

An example from *Platanthera leucophaea*

Wallace (2000, submitted) presents data on genetic diversity in a rare orchid, *Platanthera leucophaea*. Seven 10-mer primers were used to genotype between 2 and 40 individuals (mean: 14.1) from 14 populations (Maine: 1; Michigan: 5; Ohio: 8). 69 bands were scored, of which 66 were polymorphic in the sample. Shannon-Weaver diversity indices calculated from band presence-absence data give an estimate of $F_{st} = 0.426$, while AMOVA using squared Euclidean distances among individuals gives an estimate of $F_{st} = 0.252$. Both are substantially smaller than the estimate of $F_{st} = 0.752$ derived from a survey of 7 polymorphic allozyme loci in the 8 Ohio populations. Because the allozyme survey revealed substantial inbreeding within populations ($F_{is} = 0.747$), the question naturally arises whether neglecting within population inbreeding in analyses of RAPD data can account for the discrepancy.

Results of our re-analysis of the RAPD data are illustrated in Figure 2. The posterior mean of f (our estimate of F_{is}) is 0.889, and its 95% credible interval is [0.675, 0.996]. Thus, our results are consistent with the evidence provided by allozymes that there is substantial inbreeding within populations. Moreover, given the considerable uncertainty associated with our estimate of F_{is} the values estimated from allozymes and RAPDs do not appear to be inconsistent with one another.

The posterior mean of θ (our estimate of F_{st}) is 0.392, and its 95% credible interval is [0.343, 0.443]. While our estimate is substantially larger than the AMOVA estimate (0.392 *versus* 0.252), it is consistent with the estimate derived from Shannon-Weaver band diversities. More importantly, our estimate is substantially smaller than the one derived from allozymes (0.392 *versus* 0.752), suggesting that the evolutionary processes associated with diversification at allozyme loci in these populations occur at different rates from those associated with diversification at RAPD loci (compare Balloux et al. 2000). Because of the relatively small number of populations and loci included in the allozyme survey, however, the apparent difference between F_{st} in allozyme and RAPD data should be interpreted as suggestive rather than definitive.

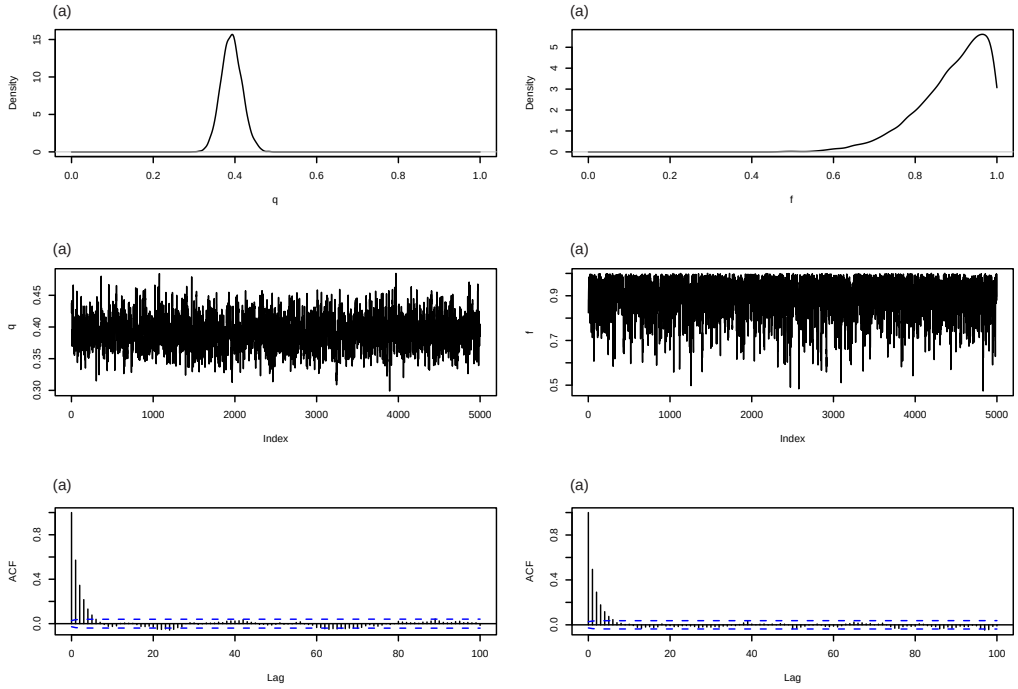

Figure 2: Convergence of the sampler and posterior densities for θ and f with data from *Platanthera leucophaea* (Wallace 2000, submitted). (a) Posterior density of θ. (b) Posterior density of f. (c) Sample history for θ. (d) Sample history for f. (e) Autocorrelation for θ. (f) Autocorrelation for f.

Discussion

Because of the relative ease with which they can be obtained, dominant markers (RAPDs, AFLPs, ISSRs) are widely used in conservation and evolutionary genetics. Unfortunately, existing methods for estimating F_{st} from dominant-marker data either assume that genotypes are found in Hardy-Weinberg proportions within populations (e.g., Lynch and Milligan 1994; Stewart and Excoffier 1996; Zhivotovsky 1999) or treat multi-locus RAPD phenotypes as haplotypes and use similarity or distance indices in an analysis of molecular variance (e.g., Isabel et al. 1999; Schneider et al. 2000). We propose an approach that takes full advantage of the data, allowing us to incorporate uncertainty about the magnitude of the within-population inbreeding coefficient into estimates of F_{st}.

Simulations demonstrate that estimates of F_{st} obtained using our approach are accurate and reliable. Estimates of within-population inbreeding are, not surprisingly, substantially less reliable. Nonetheless, both simulations and re-analysis of data from *Platanthera leucophaea* illustrate that plausible inferences about the magnitude of F_{is} are possible when data are available from enough loci and enough populations.

We developed our approach by assuming that each scorable band in a gel represents allelic variation at a single genetic locus. Rabouam et al. (1999) point out, however, that some RAPD fragments may not correspond to genomic DNA sequences. As a result, some of the variation scored in RAPD surveys, and possibly in surveys using other dominant markers, may be non-allelic. To the extent that differences among individuals and populations in band presence and absence are due to non-allelic differences, conclusions about the genetic structure of populations are necessarily suspect.

We also assumed that a Beta distribution provides an adequate description of the pattern of allele frequency variation among populations. While a Beta distribution can accomodate many patterns of allele frequency differentiation, it cannot accomodate all of them. If the actual pattern of allele frequency variation is multimodal, for example, a Beta distribution will fit the pattern poorly, and a mixture of Beta distributions may be more appropriate. Nonetheless, estimates of F_{st} depend only on the variance of allele frequencies among populations, not on the shape of the distribution. We conjecture that estimates of F_{st} will be robust to violations of this assumption, and we intend to test that conjecture in future work.

Acknowledgments

We are indebted to Lisa Wallace for sharing her data on *Platanthera leucophaea* with us and to the the statistical ecology discussion group at the University of Connecticut for comments and suggestions on early versions of this work.

References

Balloux F, Brunner F, Goudet J (2000) Microsatellites can be misleading: an empirical and simulation study. *Evolution*, **54**, 1414–1422.

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

Cockerham CC (1969) Variance of gene frequencies. *Evolution*, **23**, 72–84.

Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory*. Harper & Row, New York.

Ewens WJ (1979) *Mathematical Population Genetics* Springer-Verlag, Berlin.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

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

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Introducing Markov chain Monte Carlo. In: *Markov Chain Monte Carlo in Practice* (ed. Gilks WR, Richardson S, Spiegelhalter DJ), pp. 1–19, Chapman & Hall, London.

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

Isabel N, Beaulieu J, Thériault P, and Bousquet J (1999) Direct evidence for biased gene diversity estimates from dominant random amplified polymorphic CNA (RAPD) fingerprints. *Molecular Ecology*, **8**, 477–483.

Lewontin RC, Hubby JL (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, **54**, 595–609.

Lynch M, Milligan BG (1994) Analysis of population genetic structure with RAPD markers. *Molecular Ecology*, **3**, 91–99.

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

Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the USA*, **76**, 5269–5273.

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

Rabouam C, Comes AM, Bretagnolle V, *et al.* (1999) Features of DNA fragments obtained by random amplified polymorphic DNA (RAPD) assays. *Molecular Ecology* **8**, 493–503.

Raftery AL, Lewis S (1992a) Comment: One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, **7**, 493–497.

Raftery, AL, Lewis, S (1992b) How many iterations in the Gibbs sampler? In: *Bayesian Statistics 4* (ed 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.

Schneider S, Roessli D, Excoffier L (2000) *ARLEQUIN v2.0: A Software for Population Genetic Analysis*. Genetics and Biometry Laboratory, University of Geneva.

Stewart CN, Excoffier L (1996) Assessing population genetic structure with RAPD data: application to *Vaccinium macrocarpon* (American cranberry). *Journal of Evolutionary Biology* **9**, 153–171.

Vos P, Hogers R, Bleeker M *et al.* (1995) A new technique for DNA fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

Wallace LE (2000) The significance of population size on reproductive success and genetic variability in the Eastern prairie white fringed orchid, *Platanthera leucophaea*. *American Journal of Botany*, **87**, 165.

Wallace LE (submitted) Using RAPD's to assess suitable units of conservation in a threatened orchid, *Platanthera leucophaea*. *Plant Species Biology*.

Weir BS (1996) *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

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

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, **18**, 6531–6535.

Wolfe AD, Liston A (1998) Contributions of PCR-based methods to plant systematics and evolutionary biology. In: *Plant Molecular Systematics II* (ed Soltis DE, Soltis PS, Doyle JJ), pp. 43–86. Kluwer Academic Publishers, Boston.

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

Wright, S (1969) *Evolution and the Genetics of Populations. Volume 2, The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

Zhivotovsky LA (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology*, **8**, 907–913.