Bayesian Analysis of Genetic Differentiation Between Populations

Jukka Corander,* Patrik Waldmann† and Mikko J. Sillanpää*,[1]

**Rolf Nevanlinna Institute, FIN-00014, University of Helsinki, Helsinki, Finland and †Department of Biology, FIN-90014, University of Oulu, Oulu, Finland*

Manuscript received September 6, 2002
Accepted for publication October 4, 2002

ABSTRACT

We introduce a Bayesian method for estimating hidden population substructure using multilocus molecular markers and geographical information provided by the sampling design. The joint posterior distribution of the substructure and allele frequencies of the respective populations is available in an analytical form when the number of populations is small, whereas an approximation based on a Markov chain Monte Carlo simulation approach can be obtained for a moderate or large number of populations. Using the joint posterior distribution, posteriors can also be derived for any evolutionary population parameters, such as the traditional fixation indices. A major advantage compared to most earlier methods is that the number of populations is treated here as an unknown parameter. What is traditionally considered as two genetically distinct populations, either recently founded or connected by considerable gene flow, is here considered as one panmictic population with a certain probability based on marker data and prior information. Analyses of previously published data on the Moroccan argan tree (*Argania spinosa*) and of simulated data sets suggest that our method is capable of estimating a population substructure, while not artificially enforcing a substructure when it does not exist. The software (BAPS) used for the computations is freely available from http://www.rni.helsinki.fi/~mjs.

ONE of the inevitable consequences of genetic drift is that gene frequencies diverge between populations of a common origin when migration and mutation rates are low. In evolutionary science, a lot of effort has therefore been devoted to the development and empirical application of statistical methods for estimation of the degree of population differentiation using molecular marker data. A majority of studies have used statistical measures derived from Wright's *F*-statistics (Wright 1951, 1965), while only recently, more sophisticated methods have been proposed; see, *e.g.*, Holsinger (1999), Edwards and Beerli (2000), Kitada *et al.* (2000), Pritchard *et al.* (2000), and Dawson and Belkhir (2001).

Natural animal and plant populations typically have a nested substructure with respect to their hierarchical spatial pattern, such as sites within riverbeds, riverbeds within a river, or rivers within a river basin (Weir 1996). When sampling individuals from such hierarchical systems, one often follows the substructure (at least implicitly) by collecting the data groupwise from individuals sharing some low level of hierarchy. The traditional analyses have quantified this kind of nested genetic variation by using various statistical measures and conditioning on the fixed preassigned structure. Recently, approaches of Pritchard *et al.* (2000) and Dawson and Belkhir (2001) used Bayesian model-based clustering to assign individuals one at a time to unknown populations. Their main focus was on the situation where the information contained in the sampling design is not available or not imposed, although Pritchard *et al.* (2000) briefly considered also the other case. Here we introduce an approach that is conditioned on the geographical sampling information available about the preassigned groups of individuals. The partition among the groups is treated here as the parameter of main interest, such that all group combinations are considered *a priori* equally likely. The molecular marker data are then used for assessing which substructures are empirically plausible. The actual analysis is performed using a systematic Bayesian approach, where a Markov chain Monte Carlo (MCMC) estimation is used whenever the number of possible partitions is too large to be handled with exact calculations.

The posterior distribution of the population substructure and population-specific parameters also enables the estimation and uncertainty assessment for any related quantities that might be of interest, such as the *F*-statistics familiar to most evolutionary biologists. Our method is applicable to several types of codominant markers [*e.g.*, allozymes, single-nucleotide polymorphisms (SNPs), and microsatellites], on the basis of assumptions of Hardy-Weinberg equilibrium (HWE) and linkage equilibrium between loci within each observed population. We also discuss possible extensions of the method-

[1]*Corresponding author:* Rolf Nevanlinna Institute, Research Institute of Mathematics, Statistics and Computer Science, P.O. Box 4, FIN-00014, University of Helsinki, Helsinki, Finland.
E-mail: mjs@rolf.helsinki.fi

ology to higher-dimensional hierarchies and an alternative way of handling the situation where the HWE assumption seems empirically unjustified.

The proposed Bayesian model is described in the following section, whereas the computational details are given in the APPENDIX. Investigation of genetic separation among populations is considered thereafter. To illustrate the methodology we use the Moroccan argan tree (*Argania spinosa*) data from PETIT *et al.* (1998). Results of sensitivity studies using simulated data are also presented, and finally, some possibilities for further extensions of the method are discussed.

BAYESIAN MODELING OF ALLELE FREQUENCIES IN A GEOGRAPHICALLY STRUCTURED POPULATION

We consider a sampling design where individuals are gathered from N_P distinct populations on the basis of available prior knowledge concerning their geographical separation. Assume that genotypes are observed at N_L independent (unlinked) marker loci, where at each locus j there are $N_{A(j)}$ possible alleles to be distinguished. To be adequate sources of information about population substructure, these markers should be neutral and their mutation frequency should be reasonably low. Furthermore, the unlinked genetic markers are assumed to be in HWE within each observed population.

Since the true underlying population substructure is unknown, the number of populations with differing allele frequencies is treated here as a parameter ν_P, having the range of reasonable values $[1, N_P]$, where the upper bound is directly given by the sampling design. At locus j, the unobserved probability of observing allele A_{jk} (allele frequency) in population i is represented by p_{ijk} $[i = 1, \ldots, \nu_P; j = 1, \ldots, N_L; k = 1, \ldots, N_{A(j)}]$. To simplify the notation, θ is used as a generic symbol jointly for the allele frequencies (θ_i for population i), and similarly n represents jointly the observed marker allele counts n_{ijk}. Missing alleles are simply ignored among observations, since they do not contribute in the model under HWE assumption. Note here that p_{ijk} depends on ν_P, and consequently, n_{ijk} may be a sum of several allele counts calculated from the original populations. The partition of the original populations can be represented by a $N_P \times N_P$ population structure parameter matrix S, with elements defined as

$$S_{mr} = \begin{cases} 1, & \text{if } \theta_m = \theta_r, \\ 0, & \text{otherwise,} \end{cases}$$

where m and r take values in the range $[1, N_P]$. The joint distribution of the observed marker allele counts and the model parameters is specified by

$$\begin{aligned} \pi(\theta, \nu_P, S, n) &= \pi(n|\theta, \nu_P, S)\pi(\theta|\nu_P, S)\pi(S|\nu_P)\pi(\nu_P) \\ &\propto \prod_{i=1}^{\nu_P}\prod_{j=1}^{N_L}\prod_{k=1}^{N_{A(j)}}[p_{ijk}^{n_{ijk}}\pi(p_{ijk})]\pi(S|\nu_P)\pi(\nu_P), \end{aligned} \quad (1)$$

where $\pi(n|\theta, \nu_P, S) \propto \prod_{i=1}^{\nu_P}\prod_{j=1}^{N_L}\prod_{k=1}^{N_{A(j)}}p_{ijk}^{n_{ijk}}$ is the multinomial likelihood, $\pi(\theta|\nu_P, S) = \prod_{i=1}^{\nu_P}\prod_{j=1}^{N_L}\prod_{k=1}^{N_{A(j)}}\pi(p_{ijk})$ is the prior density of θ, and $\pi(S|\nu_P)\pi(\nu_P)$ is the joint prior of the structure parameters. When the allele frequencies of two populations are equal, their observed counts in n can be summed together in the likelihood. It is worth noting that *under the assumptions of HWE and linkage equilibrium* the above model arises naturally from the basic modeling principles of the Bayesian framework; see, *e.g.*, BERNARDO and SMITH (1994).

In a multinomial setting, a common choice as a prior $\pi(\theta|\nu_P, S)$ for the allele frequencies (see RANNALA and MOUNTAIN 1997; HOLSINGER 1999; PRITCHARD *et al.* 2000; ANDERSON and THOMPSON 2002) is the Dirichlet(λ) distribution with hyperparameter vector λ, where each element λ_k represents the prior mass on the allele k (at some arbitrary locus). As a reference assumption we prefer an invariant noninformative prior with $\lambda_{ijk} = 1/N_{A(j)}$, which can be interpreted to relatively contain as much information as a likelihood with a single observation. This particular prior was also suggested in ANDERSON and THOMPSON (2002) in a related context. It is further assumed that $\pi(S|\nu_P)\pi(\nu_P)$ is a uniform distribution in the finite space of distinct values of (ν_P, S). A strategy enabling joint estimation of the parameters (θ, ν_P, S) in model (1) is described in the APPENDIX, and the given noninformative priors are used in all subsequently reported analyses of real and simulated data.

MEASURING OF GENETIC SEPARATION AMONG POPULATIONS

A wide diversity of evolutionary measures of population differentiation is available in the genetic literature (see WEIR 1996; NAGYLAKI 1998; TOMIUK *et al.* 1998; YANG 1998; EXCOFFIER 2001; ROUSSET 2001). Simple statistical point estimates of such parameters can be obtained, but this requires conditioning on a known population structure. Quantification of the uncertainty about the estimates is much more tedious and resampling methods (like bootstrap) are often applied in the estimation of confidence intervals. However, resampling methods may provide biased estimates when based on hierarchical data sets (PETIT and PONS 1998).

Given the posterior of the allele frequencies and population structure (θ, ν_P, S), it is possible to derive the posterior distribution also for any function of these parameters, such as the familiar F-statistic F_{ST} (in examples we have used the formula given in NEI 1977). Our approach enables Bayesian model-averaged estimation of evolutionary measures, by accounting for the uncertainty related to the unknown population structure. For a general discussion of Bayesian model averaging, see BALL (2001) and SILLANPÄÄ and CORANDER (2002). To aid in interpretation of the genetic marker data with respect to separation among populations, we emphasize

the importance of studying the visual appearances of the posterior distributions of all parameters of interest. Using the posterior distribution of structure parameters, it is possible to give a measure of the *uncertainty* concerning whether any particular population pair among the original N_P populations can in fact be regarded as samples from a single population. However, our model cannot readily be used to empirically verify whether one has collected individuals that are originally from several different populations within a single geographical region. The uncertainty can be presented as an $N_P \times N_P$ matrix with the (mr)th element defined as the posterior probability

$$P(\theta_m = \theta_r | n), \tag{2}$$

which can be calculated by summing the posterior probabilities of such partitions where the two populations (m and r) are merged together (see the APPENDIX). However, when the amount of data increases, differences can be detected on a finer scale. Consequently, it may be that the posterior probability (2) approaches zero, although the allele distributions are rather close to each other in some metric. In addition to the probability (2), one can technically measure the discrepancy between allele *distributions* of two populations over different loci by using the Kullback-Leibler divergence (KULLBACK and LEIBLER 1951; KULLBACK 1968; ANDERSON and THOMPSON 2002). However, as was pointed out to us by a referee, the evolutionary meaning of this quantity is not known and needs to be further investigated.

EXAMPLE ANALYSES

Real data: To illustrate the proposed methodology, we used the Moroccan argan tree (*A. spinosa*) data from PETIT *et al.* (1998), which has previously also been analyzed in HOLSINGER (1999). Due to implementation, Holsinger's analysis was based on preprocessing of multiallelic data to a biallelic form, and therefore, his results are comparable to ours only under the same restriction.

The original data consist of allele measurements at 12 isozyme loci (two to five alleles) for 12 different populations with 20–50 individuals in each. We use the same abbreviated notation for the population names as PETIT *et al.* (1998). For $N_P = 12$ there are 4,213,597 possible partitions of the populations, so that the exact analysis may not in this case be considered feasible for a routine analysis. Nevertheless, we performed the exact analysis and the differences in pairwise probabilities (2) appeared to be negligible when compared to the MCMC approximation (based on a Markov chain of length 10^5 after a discarded "burn-in" period of 10^4 iterations). In all investigations reported subsequently we have used the MCMC approach with the same chain length for the burn-in. Mixing properties of the chains were monitored visually using various tools (*e.g.*, cumulative occupancy plot; see UIMARI and SILLANPÄÄ 2001), and our algorithm seems to perform well in this respect. Note that successive realizations of allele frequencies are independent, and consequently, values for quantities depending on allele frequencies (such as F_{ST}) do not have any autocorrelation (see the APPENDIX).

TABLE 1

Posterior probabilities of different groupings of population samples for the *Argania spinosa* data

Population groupings	Posterior probability
(MI, SI, TE)	0.999
(AR, TT)	0.874
(AD, AR)	0.093
Others	0.000

Simulated data: In addition to the example analysis with real data, we applied our approach also to data sets that were simulated from population models with or without substructure. This enables investigation of whether one has a sufficient probability of detecting differences among allele frequencies while still maintaining a low probability of imposing a structure artificially, when such does not exist. From a theoretical point of view it is clear that the given Bayesian model will *a priori* support the simplest partition with no separation of populations, since the conditional distribution of the marker frequencies has then the smallest possible number of parameters.

We simulated data sets from distributions with 10 different alleles, some of which were considerably rare. In the first setting alleles were generated for a single locus with frequencies [0.3, 0.3, 0.2, 0.1, 0.05, 0.015, 0.015, 0.01, 0.005, 0.005]. Samples of 10, 20, and 50 diploid individuals from this single population were then randomly assigned into five different populations. An analogous setting with the same allele frequencies was also used to generate observations from five independent loci simultaneously. In the second scheme alleles were generated from two populations with different allele frequencies, one having the frequencies in the previous example and one with frequencies [0.15, 0.15, 0.15, 0.15, 0.1, 0.1, 0.05, 0.05, 0.05, 0.05]. The same sample sizes and numbers of loci (one and five) as in the first scheme were used. All sampling configurations were replicated 10,000 times and the posterior distributions were analytically calculated for each replicate.

Results, real data: From the posterior of structure S based on the real data (see Table 1), samples from populations Mijji (MI), Sidi Ifni (SI), and Tensif (TE) are all considered to originate from a single population with probability 0.999. Furthermore, given the abbreviations Argana (AR), Tizint'est (TT), and Ademine (AD),

population samples in pairs (AR, TT) and (AD, AR) are considered to have equal origins with probabilities 0.874 and 0.093, respectively. All the remaining combinations of populations are estimated to have corresponding probability equal to zero. For comparison the posterior mean of F_{ST} equals 0.273 (95% credible interval being [0.251, 0.296]). Figure 1 shows the posterior density of F_{ST} and Figure 2 illustrates the rapid convergence of the particular chain with respect to ν_P in a form of cumulative occupancy probabilities. The posterior estimate of F_{ST} is rather distinct from the value obtained in Holsinger (1999), and therefore, we repeated our analysis in this respect, using a biallelic transform of the original data following Holsinger (1999). The resulting estimate is 0.172 (with 95% credible interval being [0.148, 0.196]), and the comparable estimate and credible interval given in Holsinger (1999) are 0.192 and [0.177, 0.206], respectively. The slightly lower value of our estimate is expected, since we are accounting for the equality of certain populations.

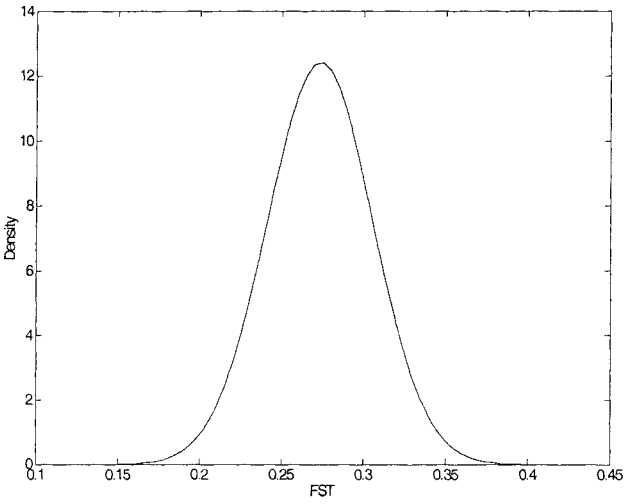

Figure 1.—Kernel-estimated posterior distribution of F_{ST} for the *Argania spinosa* data.

To investigate sensitivity and the effects of individual loci, we reanalyzed the data using only a single locus at a time. In Table 2, only counts of loci for which the pairwise posterior probabilities $P(\theta_m = \theta_r|n)$ for populations (m and r) that exceed 0.75 are shown. It can be seen that most populations have concordant allele frequencies at many loci; however, concordant loci vary among the populations.

The estimated posterior means of Kullback-Leibler divergences are used in a three-dimensional multidimensional scaling plot of the populations (Figure 3) to visualize their distinction from each other. The estimated distances among populations MI, SI, and TE are equal to zero, and therefore, the population labels are overlapping in the plot. The populations Beni-Snassen (BS) and Oued Grou (OG) seem to locate far from the other populations, which is in concordance with the results of Petit *et al.* (1998). When Figure 3 is compared to the geographical map given in Petit *et al.* (1998), one can conclude that some genetic distances coincide relatively closely with the geographical distances, whereas some pairs of genetically similar populations are very distant from each other.

Results, simulated data: For the simulated data sets lacking population substructure, results are summarized in Figure 4. Histograms in the figure show the empirical distribution (over replications) in different settings for the posterior probability of the event that any two populations are equal. The panels correspond to the case with one locus only; for data sets with five loci the posterior

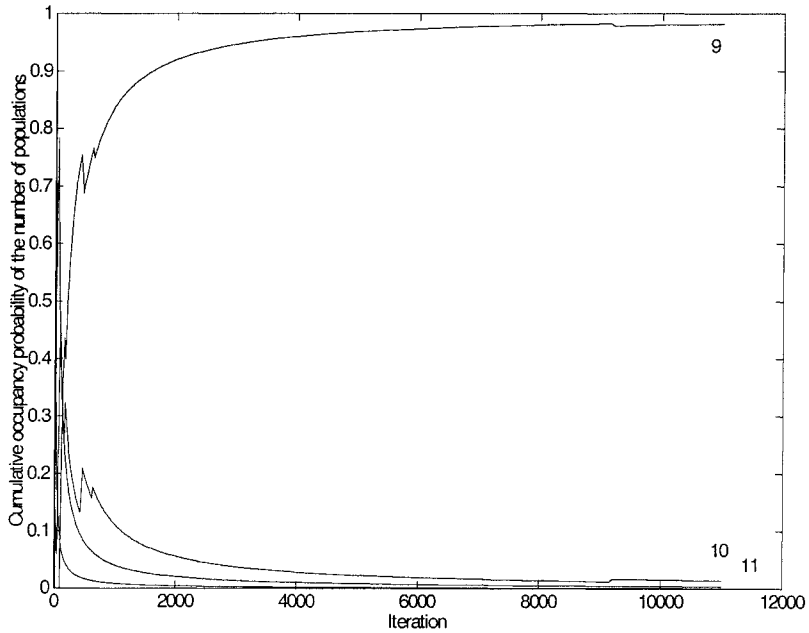

Figure 2.—Cumulative probabilities of different values of ν_P at each MCMC iteration cycle in an early phase (≤11,000 cycles) of the chain.

TABLE 2

Counts of loci of the *Argania spinosa* data for which pairwise posterior probabilities of populations being equal exceed 0.75

	AB	AD	AR	BS	GO	MI	OG	SI	TA	TE	TM	TT
AB		7	7	2	4	6	6	5	5	5	5	7
AD			8	5	4	7	8	7	7	6	6	9
AR				5	5	8	8	7	7	7	8	10
BS					4	7	6	7	5	7	5	5
GO						6	5	6	4	6	5	4
MI							8	8	7	8	7	7
OG								8	7	8	8	8
SI									6	8	7	7
TA										7	8	8
TE											6	8
TM												8
TT												

probability was equal to unity for all replicates. The analysis illustrates clearly that our method will support merging of populations if the data do not provide enough evidence against the similarity hypothesis. Results for the configurations where the underlying structure consists of two distinct populations are presented in Figure 5, analogously to the previous example. As expected, the empirical power to detect the correct underlying structure increases with the sample size.

DISCUSSION

We have presented a Bayesian method for estimating hidden population substructure using multilocus molecular markers. Underlying model assumptions concerning HWE and linkage equilibrium within the populations imply that each individual contributes two *independent* alleles to the likelihood at each locus. To check the validity of these assumptions, one may use, for instance, the methods introduced in Ayres and Balding (1998, 2001), respectively. When the populations are in significant departure from HWE, the data are effectively assumed to contain too much information about the allele frequencies, and consequently, the level of uncertainty concerning the parameters will become

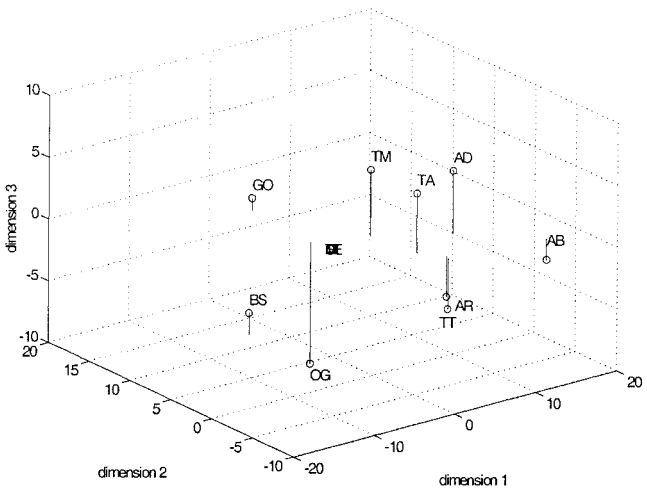

Figure 3.—A multidimensional scaling plot of the estimated posterior means of Kullback-Leibler divergences among *Argania spinosa* populations (MI, SI, and TE have zero distances).

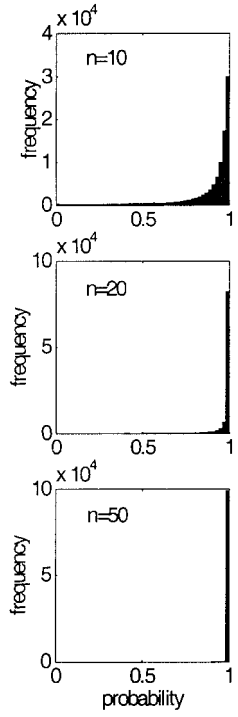

Figure 4.—No substructure. Shown are empirical distributions (based on 10,000 replicates) of the posterior probability of the event that any two of five simulated populations are equal on the basis of single-locus data. All underlying populations have equal allele frequencies. Sample size (n) is indicated at each top left corner.

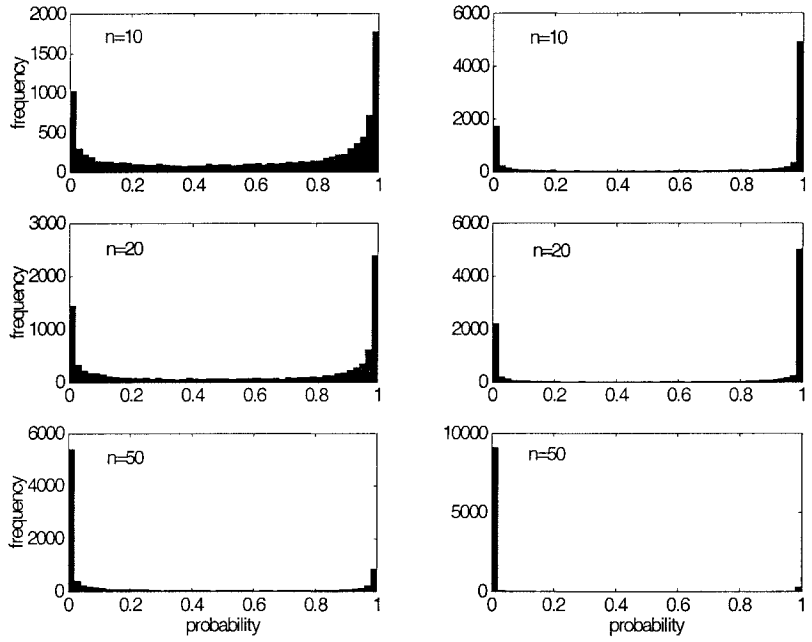

FIGURE 5.—Two underlying populations. Shown are empirical distributions (based on 10,000 replicates) of the posterior probability of the event that the two simulated populations are equal on the basis of single-locus (left) and five-loci (right) data. The underlying populations have different allele frequencies. Sample size (n) is indicated at each top left corner.

underestimated. One potential remedy for this is to parameterize the model using genotype frequencies instead of allele frequencies. Such a model avoids HWE assumption and allows for any form of dependence between alleles. This would also enable the use of commonly available dominant markers such as randomly amplified polymorphic DNAs and amplified fragment length polymorphisms in our analysis. For another Bayesian approach to the analysis of dominant markers, see HOLSINGER *et al.* (2002). In the genotype model with codominant markers it is possible to take into account missing allelic data through data augmentation (SCHAFER 2000). However, the genotype model may become infeasible when the data are scarce or when the number of alleles at different loci is high.

When the aim of modeling the marker data is investigation of neutral evolution, one should bear in mind the assumption of a relatively slow rate of mutation of the alleles. In this respect conclusions with respect to differentiation are most well suited for allozymes and SNPs on low-mutating genome regions. One should be more careful concerning inferences about genetic drift when using microsatellite alleles, since they fluctuate more randomly over generations.

We have here concentrated on utilization of the geographical information available in a two-level hierarchy, since it corresponds to commonly used sampling designs. Occasionally, sampling designs may enable the use of information even from higher-dimensional hierarchies (typically, at three levels). Such designs can be taken into account by defining the hyperparameters in the prior as random coefficients depending on some parameter indexing the nested population substructure (*cf.* HOLSINGER 1999). Although the exact form of the posterior may then not be available even for a small number of populations, the MCMC approach presented here can be modified to handle more general settings, by suitably changing the mechanism of generating proposals.

The general Bayesian approach applied here is very flexible, and it would be valuable to incorporate information from phenotypes, different mutation models, spatial distances, and demographic parameters in the future. In conclusion, we have shown that the Bayesian model is a powerful tool for inference about the genetic population structure. However, as the simulation results with an underlying population structure illustrate, one cannot expect to obtain conclusive evidence for separation among populations when the numbers of sampled individuals and loci are small, unless the observed allele frequencies are considerably different. This feature represents common sense in statistical inference and protects against exaggerated interpretations concerning differences caused by random fluctuations in allele frequencies over generations.

Our analysis shows the favorable feature of combining information from several loci into a single probability model, as opposed to the simple averaging used in a traditional F_{ST} analysis. One special advantage of the proposed MCMC sampling scheme is that tuning problems related to the choice of proposal and prior distributions seem to be minimized. This reflects the positive effect of analytically integrating out relative allele frequency parameters from the posterior expression of the structure. A major advantage of the approach as a whole compared to most earlier methods is that the number of populations is treated here as an unknown parameter. Hence, we can avoid the labeling problems of populations that occur with high levels of gene flow. In other words, what is considered as two genetically distinct

populations, either recently founded or connected by considerable gene flow, would be considered as one panmictic population with a certain probability in our approach.

The authors thank two anonymous referees whose suggestions and comments significantly improved the original manuscript. This work was supported by research grant nos. 52457 and 47201 from the Academy of Finland and by the Centre of Population Genetic Analyses, University of Oulu, Finland.

LITERATURE CITED

Aarts, E. H. L., and J. Korst, 1989 *Simulated Annealing and Boltzmann Machines*. Wiley, Chichester, UK.

Abramovitz, M., and I. A. Stegun (Editors), 1969 *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York.

Anderson, E. C., and E. A. Thompson, 2002 A model-based method for identifying species hybrids using multilocus genetic data. Genetics **160:** 1217–1229.

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

Ayres, K., and D. J. Balding, 2001 Measuring gametic disequilibrium from multilocus data. Genetics **157:** 413–423.

Ball, R. D., 2001 Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. Genetics **159:** 1351–1364.

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

Dawson, K. J., and K. Belkhir, 2001 A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet. Res. **78:** 59–77.

Edwards, S. V., and P. Beerli, 2000 Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution **54:** 1839–1854.

Excoffier, L., 2001 Analysis of population subdivision, pp. 271–308 in *Handbook of Statistical Genetics*, edited by D. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

Gilks, W., S. Richardson and D. Spiegelhalter, 1996 *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

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

Holsinger, K. E., P. O. Lewis and D. K. Dey, 2002 A Bayesian approach to inferring population structure from dominant markers. Mol. Ecol. **11:** 1157–1164.

Kitada, S., T. Hayashi and H. Kishino, 2000 Empirical Bayes procedure for estimating genetic distance between populations and effective population size. Genetics **156:** 2063–2079.

Kullback, S., 1968 *Information Theory and Statistics*. Wiley, New York.

Kullback, S., and R. A. Leibler, 1951 On information and sufficiency. Ann. Math. Stat. **22:** 79–86.

Nagylaki, T., 1998 Fixation indices in subdivided populations. Genetics **148:** 1325–1332.

Nei, M., 1977 F-statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. **41:** 225–233.

Petit, R. J., and O. Pons, 1998 Bootstrap variance of diversity and differentiation estimators in a subdivided population. Heredity **80:** 56–61.

Petit, R. J., A. El Mousadik and O. Pons, 1998 Identifying populations for conservation on the basis of genetic markers. Conserv. Biol. **12:** 844–855.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

Rannala, B., and J. L. Mountain, 1997 Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA **94:** 9197–9201.

Rousset, F., 2001 Inferences from spatial population genetics, pp. 239–270 in *Handbook of Statistical Genetics*, edited by D. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

Schafer, J. L., 2000 *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC Press, Boca Raton, FL.

Sillanpää, M. J., and J. Corander, 2002 Model choice in gene mapping: what and why. Trends Genet. **18:** 301–307.

Tomiuk, J., B. Guldbrantsen and V. Loeschcke, 1998 Population differentiation through mutation and drift—a comparison of genetic identity measures. Genetica **102/103:** 545–558.

Uimari, P., and M. J. Sillanpää, 2001 Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. Genet. Epidemiol. **21:** 224–242.

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

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

Wright, S., 1965 The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution **19:** 395–420.

Yang, R.-C., 1998 Estimating hierarchical F-statistics. Evolution **52:** 950–956.

Communicating editor: J. B. Walsh

APPENDIX: ESTIMATION OF MODEL PARAMETERS

To enable Bayesian inference jointly about parameters (θ, ν_P, S) in general, the standard Metropolis-Hastings MCMC algorithm (*e.g.*, Gilks *et al.* 1996) is utilized to obtain an approximation to the posterior distribution. However, for a sufficiently small number of collected populations N_P, the limited size of the parameter space enables the posterior distribution of (ν_P, S) to be calculated by complete enumeration, such that the marginal likelihood of a particular partition value (ν_P, S) is divided by the sum of marginal likelihoods over all possible partitions. Conditional on this distribution, one can generate a suitable number of independent posterior realizations of θ explicitly (see below).

The number of distinct values of S (*i.e.*, partitions of the finite set $\{1, \ldots, N_P\}$) equals the sum $\sum_{\nu_P=1}^{N_P} \sigma_{\nu_P}^{N_P}$, where $\sigma_{\nu_P}^{N_P}$ is the Stirling number of the second kind (see, *e.g.*, Abramovitz and Stegun 1969). In routine applications, the number of distinct partitions is practically small for at least $N_P \leq 10$ (where $\sum_{\nu_P=1}^{N_P} \sigma_{\nu_P}^{N_P} \leq 115{,}975$), to allow for enumerative calculations. For larger N_P we use an MCMC algorithm to generate samples from the posterior distribution of (θ, ν_P, S) with two distinct move types: (1) randomly *split* a population into two distinct parts or *merge* two different populations into a single one and (2) update allele frequencies θ with Gibbs sampling conditional on the data and the current value of parameters (ν_P, S). Since the model is specified at the population level, split and merge moves are restrictively proposed only within the range of populations that were present in the original sampling configuration.

In typical applications the value N_P is small enough (say, at most 30–50) so that the time required for the convergence of the MCMC approach is presumably acceptable for practical purposes. In Dawson and Belkhir (2001) partitioning at the individual level (as opposed to the population level used here) using MCMC was still performing well for 200 entities. However, in

very complicated problems with a really large number of populations it may be sensible to approximate only the mode of the posterior distribution. In such cases it is preferable to use a formulation in terms of a combinatorial optimization problem, such as those solved by simulated annealing (Aarts and Korst 1989).

The posterior distribution of the population structure is proportional to the analytically calculated integral (see Rannala and Mountain 1997) according to

$$\pi(\nu_P, S|n) \propto \int \prod_{i=1}^{\nu_P} \prod_{j=1}^{N_L} \prod_{k=1}^{N_{A(j)}} p_{ijk}^{n_{ijk}+\lambda_{ijk}} d\theta = \prod_{i=1}^{\nu_P} \prod_{j=1}^{N_L} \frac{\Gamma(\sum_k \lambda_{ijk})}{\Gamma(\sum_k (\lambda_{ijk} + n_{ijk}))} \prod_{k=1}^{N_{A(j)}} \frac{\Gamma(\lambda_{ijk} + n_{ijk})}{\Gamma(\lambda_{ijk})}, \quad \text{(A1)}$$

where $\Gamma(\cdot)$ is the gamma function.

The acceptance ratio for the Metropolis-Hastings step, where current populations given in (ν_P, S) are split or merged to form a proposal (ν_P^*, S^*), equals

$$\frac{\pi(\nu_P^*, S^*|n)}{\pi(\nu_P, S|n)} \times \frac{q((\nu_P, S)|(\nu_P^*, S^*))}{q((\nu_P^*, S^*)|(\nu_P, S))}, \quad \text{(A2)}$$

where $q(\cdot|\cdot)$ is the conditional probability of proposing a population substructure from a given one, calculated explicitly at each iteration. The proposed structures are generated uniformly from the set of possible splits or mergings at a given configuration. The prior ratio of the structure parameters equals one for all possible values and cancels therefore from (A2).

Given the previously specified priors, the full conditional distribution of θ is a product of Dirichlet distributions, given by

$$\pi(\theta|\nu_P, S, n) \propto \prod_{i=1}^{\nu_P} \prod_{j=1}^{N_L} \prod_{k=1}^{N_{A(j)}} p_{ijk}^{n_{ijk}+\lambda_{ijk}}, \quad \text{(A3)}$$

from which values can be drawn explicitly. Note that the full conditional distribution given a specific partition remains unchanged during the simulation. In many analyses the used prior gives an equal mass to all alleles, although it would also be possible to incorporate knowledge from previous studies into λ. In the approach presented we prefer the theoretically derived reference choice of $\lambda_{ijk} = 1/N_{A(j)}$, which was also used in Anderson and Thompson (2002; for a theoretical derivation see Bernardo and Smith 1994). As discussed in Anderson and Thompson (2002), other choices with larger λ_{ijk} lead to a prior containing a substantial amount of information when the number of alleles is large. Only the suggested prior has the property of containing as much information as a likelihood with a single observation regardless of the number of alleles, which makes it a reasonable reference choice.