DOI: 10.1534/genetics.109.112391

Likelihood-Free Inference of Population Structure and Local Adaptation in a Bayesian Hierarchical Model

Eric Bazin,*,[1] Kevin J. Dawson[†] and Mark A. Beaumont*,[2]

**School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6BX, United Kingdom and [†]Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, United Kingdom*

Manuscript received November 24, 2009
Accepted for publication March 29, 2010

ABSTRACT

We address the problem of finding evidence of natural selection from genetic data, accounting for the confounding effects of demographic history. In the absence of natural selection, gene genealogies should all be sampled from the same underlying distribution, often approximated by a coalescent model. Selection at a particular locus will lead to a modified genealogy, and this motivates a number of recent approaches for detecting the effects of natural selection in the genome as "outliers" under some models. The demographic history of a population affects the sampling distribution of genealogies, and therefore the observed genotypes and the classification of outliers. Since we cannot see genealogies directly, we have to infer them from the observed data under some model of mutation and demography. Thus the accuracy of an outlier-based approach depends to a greater or a lesser extent on the uncertainty about the demographic and mutational model. A natural modeling framework for this type of problem is provided by Bayesian hierarchical models, in which parameters, such as mutation rates and selection coefficients, are allowed to vary across loci. It has proved quite difficult computationally to implement fully probabilistic genealogical models with complex demographies, and this has motivated the development of approximations such as approximate Bayesian computation (ABC). In ABC the data are compressed into summary statistics, and computation of the likelihood function is replaced by simulation of data under the model. In a hierarchical setting one may be interested both in hyperparameters and parameters, and there may be very many of the latter—for example, in a genetic model, these may be parameters describing each of many loci or populations. This poses a problem for ABC in that one then requires summary statistics for each locus, which, if used naively, leads to a consequent difficulty in conditional density estimation. We develop a general method for applying ABC to Bayesian hierarchical models, and we apply it to detect microsatellite loci influenced by local selection. We demonstrate using receiver operating characteristic (ROC) analysis that this approach has comparable performance to a full-likelihood method and outperforms it when mutation rates are variable across loci.

THE study of the effects of natural selection at the genomic level has the potential to uncover hidden aspects of the causal pathways that relate genotype to phenotype and the environment (Sabeti *et al.* 2007). A challenge for any such research program is to distinguish signals of selection from those of a myriad other processes (McVean and Spencer 2006), particularly those related to the demographic history of the population. The study of individual candidate loci or regions of the genome, in isolation, and without regard to the (generally unknown) demographic history of the population is unlikely to be fruitful because selection can generally be mimicked by demographic processes (Teshima *et al.* 2006), and, indeed, this forms the basis of many methods of simulating loci under selection (Spencer and Coop 2004). As a consequence most recent studies concentrate on large-scale surveys of genomic regions, looking for genes that are discrepant (Teshima *et al.* 2006). Within this framework there are two broad strands. One set of approaches is based around the idea of a "selective sweep" in which an allele increases in frequency, as a result of either a single novel mutation or a change in environment, leading to reduced diversity at linked sites (Kaplan *et al.* 1989). Another modeling framework is centered around the idea of "local selection" (Charlesworth *et al.* 1997) in which alternative alleles are favored in different environments. Unlike the selective-sweep scenario where the time of onset of the sweep is an important parameter, the local selection framework is essentially ahistorical: the allele frequencies within a deme are typically modeled by assuming migration–selection–drift balance (Wright 1931; Petry 1983).

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.109.112391/DC1.

[1]*Present address:* Centre de Biologie et de Gestion des Populations (CBGP), Campus International de Baillarguet, CS 30 016, 34988 Montferrier/Lez Cedex, France.

[2]*Corresponding author:* School of Biological Sciences, University of Reading, Whiteknights, PO Box 68, Reading RG6 6BX, United Kingdom. E-mail: m.a.beaumont@reading.ac.uk

It is unclear at this stage which of the two forms of selection are most common. Certainly, the selective sweep scenario is commonly studied and this is unsurprising because the two most intensely surveyed species, humans and *Drosophila melanogaster*, both have demographic histories, the invasion of novel environments, that are conducive to selective sweeps. The model of local selection envisions a rather static view of the world, whereas the commonly held perception is of constantly changing environments and population restructuring. As always, probably the truth lies in between these two extremes, and the aim of this study is to continue the development of methods for detecting local selection, while recognizing the utility of the selective sweep paradigm under many evolutionary scenarios.

Current methods of detecting local selection from gene-frequency information tend to be based around F_{ST}, which has a variety of interpretations (Weir and Hill 2002; Balding 2003). Here, it is defined to be the probability that two gene copies share a common ancestor within the deme in which they are sampled without either of their lineages migrating (Crow and Kimura 1970, p. 105; Vitalis *et al.* 2001). Methods based on F_{ST} have a long history (Cavalli-Sforza 1966; Lewontin and Krakauer 1973). The early methods were based on moment estimators of F_{ST}, as in the above two studies and also, for example, in Beaumont and Nichols (1996), Vitalis *et al.* (2001), and Weir *et al.* (2005). More recently, likelihood-based approaches have been developed (Beaumont and Balding 2004; Riebler *et al.* 2008; Foll and Gaggiotti 2008; Guo *et al.* 2009). These latter approaches are based on a theory for the sampling distribution of genes in the infinite-island or continent-island model of structured populations. The same distribution, multinomial Dirichlet in form (a.k.a. Pólya distribution or Dirichlet compound multinomial), can be derived either from the diffusion theory of Sewall Wright (Wright 1931; Rannala and Hartigan 1996) or from coalescent theory (Balding and Nichols 1994). The key insight that lies behind the use the multinomial-Dirichlet distribution in the detection of local selection is the following result. Marker loci, linked with recombination rate r to loci in which locally deleterious alleles segregate with selection coefficient s, have an effectively reduced migration rate, m, approximated in Petry (1983) as $m' = m \times r/(s + r)$ (Barton and Bengtsson 1986; Charlesworth *et al.* 1997). Under the structured coalescent with constant deme size and migration rates, $F_{ST} = 1/(1 + 2Nm)$, and hence under local selection there is expected to be heterogeneity in the estimates of F_{ST} among loci.

The multinomial-Dirichlet framework has the advantage of having a simple likelihood function that is rapidly computed. If the mutation rate is low enough and the number of demes high enough, then we can justify this approach by the "many demes" approximation of Wakeley (1998). Often this may not be an adequate approximation, and the method then has the disadvantage that it assumes a simplified demographic history and cannot easily take into account recombination and mutational processes. Given that likelihoods in more general frameworks are computationally intractable for large numbers of loci and recombination, it is tempting to consider using a likelihood-free approach (Pritchard *et al.* 1999; Beaumont *et al.* 2002; Marjoram *et al.* 2003; Becquet and Przeworski 2007). Typically these methods require that summary statistics are computed in a large number of Monte Carlo simulations and some match is made between simulated and observed summary statistics. A problem arises in that information on whether there is selection comes from considering all the loci jointly, but to decide whether a specific locus is under selection we also need information on that particular locus. Thus a naive approach, given L loci, would be to have L sets of summary statistics. This could lead to thousands of summary statistics for an analysis. The probability of getting a close match for all L simulated loci will be vanishingly small, and consequently such an approach is unlikely to succeed.

In this article we develop a general method for efficiently computing solutions in hierarchical Bayesian models using a likelihood-free approach. We formulate a hierarchical Bayesian model for identifying loci that are subject to local selection and apply our technique, which is relatively efficient and easy to parallelize on a computing cluster. We demonstrate through the use of extensive comparisons that the method approaches the accuracy of the likelihood-based method of Beaumont and Balding (2004) in situations where the assumptions of the latter hold and exceeds it when there is variability in mutation rate among genetic markers. We then apply the method to microsatellite data from chimpanzees.

A HIERARCHICAL APPROACH TO LIKELIHOOD-FREE INFERENCE

The likelihood-free approach implemented in this study uses the regression-based method of conditional density estimation introduced in Beaumont *et al.* (2002). The approximate Bayesian computation (ABC) technique is currently undergoing quite widespread development, and a number of different approaches have been advocated. While recognizing that these recent developments should supersede the method of Beaumont *et al.* (2002), we justify the use of the regression method on the grounds that (a) as shown later, it performs well in a comparison with a full-likelihood method and (b) it has been widely used and its advantages and pitfalls are well understood. However, we note that the algorithms described in this article are particularly amenable to sequential ABC approaches (Sisson *et al.* 2007; Beaumont *et al.* 2009; Toni *et al.*

2009) and improved conditional density estimation (Blum and François 2009).

Briefly, we assume that we have measured a d-dimensional vector of summary statistics $S(x)$ from a data set. Here we make the distinction between the observed data set x and the random variable, X, generated by simulation. We have N random draws of a (scalar) parameter Φ_i ($i = 1, \ldots, N$) and corresponding summary statistics $S(X_i)$ ($i = 1, \ldots, N$) simulated from the joint distribution of parameters and summary statistics $P(S(X), \Phi)$. (The model may have any number of parameters, which can be considered jointly, but the regression adjustment described here is applied to one parameter at a time.) We scale $S(x)$ and $S(X)$ so that each summary statistic in $S(\cdot)$ has unit variance. We assume a linear model in which

$$\Phi_i = \alpha + \beta^T(S(X_i) - S(x)) + e_i,\ i = 1, \ldots, N,$$

where the e_i are drawn from a distribution common to all X_i, with a mean of zero. We use least squares to minimize

$$\sum_{i=1}^{N}\{\Phi_i - \alpha - \beta^T(S(X_i) - S(x))\}^2 K_\varepsilon(\| S(X_i) - S(x) \|), \tag{1}$$

where, assuming the model above,

$$\alpha = E(\Phi \mid S(X) = S(x)),$$

$$\| y \| = \sqrt{\sum_{i=1}^{d} y_i^2},$$

with Epanechnikov kernel

$$K_\varepsilon(t) = \begin{cases} c\varepsilon^{-1}(1 - (t/\varepsilon)^2) & t \le \varepsilon \\ 0 & t > \varepsilon. \end{cases} \tag{2}$$

Given the estimates $\hat{\alpha}$ and $\hat{\beta}$, we can approximate posterior densities by using the assumption (above) that the distribution of errors is constant in the region where $K_\varepsilon(\|S(X_i) - S(x)\|)$ is positive and hence adjust the parameter values as

$$\Phi_i^* = \Phi_i - \hat{\beta}^T(S(X_i) - S(x)) \tag{3}$$

(Beaumont *et al.* 2002). The posterior density for Φ can be approximated using some density estimation method, and in this article the local-likelihood method of Loader (1996) is used, implemented in Locfit under R, weighting the points with $K_\varepsilon(\|S(X_i) - S(x)\|)$ as above. It should be noted that the "tolerance" of the method, as discussed in this article, is not measured directly in terms of the Epanechnikov bandwidth ϵ, but in terms of P_ε, the proportion of simulated points where $\|S(X_i) - S(x)\| \le \varepsilon$.

In the context of the ABC algorithm above the choice of summary statistics and the choice of metric (implicitly Euclidean, in the example above through the use of the Epanechnikov kernel) are intertwined. Ideally one would choose summary statistics that are of low dimension and are also *Bayes sufficient* (Kolmogorov 1942). That is, we want the summary statistics $S(x)$ to satisfy the condition

$$P(\omega \mid x) = P(\omega \mid S(x)) \tag{4}$$

at all points ω in the parameter space, for all priors $P(\omega)$ (so that we are free to choose whatever prior we want). In practice, such statistics are rarely available. Many approaches to ABC (Pritchard *et al.* 1999; Marjoram *et al.* 2003; Sisson *et al.* 2007) are based on the idea of "rejection" (of observations falling outside a small acceptance region centered on the observed data), giving $P(\Phi \mid \rho(S(X), S(x)) <= \varepsilon)$ for some metric $\rho(\cdot)$. Thus, particularly for high-dimensional $S(\cdot)$, consideration should be given as much to the metric as to the summary statistics. Methods that place more emphasis on conditional density estimation (Beaumont *et al.* 2002; Blum and François 2009) aim to estimate $P(\Phi \mid S(X) = S(x))$ more precisely. A goal of such methods is to estimate the density using a larger proportion, possibly all, of the simulated points (Blum and François 2009).

Application of the ABC method to the situation addressed in the present study has a number of difficulties. We wish to make inferences on the demographic history and also on individual loci. This is a problem that is suited to a hierarchical Bayesian approach, and the main contribution of this study is to devise a method for performing hierarchical Bayesian analysis in the likelihood-free framework. In simple models the parameters for each locus are assumed to be identical, and if a likelihood function is available, it is simply multiplied across loci. By contrast, taking mutation rate as an example, in a hierarchical model the L loci each have their own parameter. At one extreme, identical to the simple case above, if the variability in mutation rate among loci is zero, then, in the terminology of hierarchical models, strength is "borrowed" completely between the loci, and each locus has an identical posterior distribution for mutation rate, and this is the same as the posterior distribution for the hyperparameter specifying the mean of the prior for each locus (the prior for each locus, having, in this case, zero variance). This verbal description is made clearer in the examples below. At the other extreme, the mutation rates at each locus are inferred independently—they have independent posterior distributions, and their prior has a high variance. More typically, the situation is intermediate.

In a hierarchical model one may be interested only in the posterior distribution of the hyperparameters (What is the mean mutation rate among loci? Is there evidence of nonzero variance in mutation rate among loci?). It is possible to compute summary statistics that are invariant to the ordering of loci such as means, variances,

and so forth. We refer to these as *symmetric* summary statistics. This is a typical use of the ABC method for data with many loci (*e.g.*, Pritchard *et al.* 1999 and subsequent likelihood-free articles). The use of means and variances of summary statistics among loci for the ABC analysis allows straightforward inference of the hyperparameters.

By contrast, the focus of the present study is to make inferences on locus-specific parameters, as well as inferring the hyperparameters. This leads to difficulty because one needs summary statistics for each locus. The problem of a plethora of summary statistics has been noted in the Introduction. More fundamental is that, in the absence of missing data, the loci simulated under the model are exchangeable (their ordering or labeling is irrelevant to the likelihood). Thus there is no preferred ordering of the sample loci when compared with those generated by simulation. This problem is intrinsic to any hierarchically structured model and has been encountered before in an ABC setting by Hickerson *et al.* (2006) and Hickerson and Meyer (2008), in which the exchangeable units were taxa (rather than loci, as here). Since the ordering is arbitrary, a naive scheme would simply be to match the summary statistics of the first simulated locus with the those in the first data locus (given an arbitrarily chosen order) and so forth. Although correct in principle, such an approach would be hopelessly inefficient in practice in situations with many loci. Since the ordering is arbitrary, one might find a permutation of the simulated loci that gives the closest match. However, again, with many loci such a procedure is likely to be highly computer intensive, and, without exhaustive searching, not guaranteed to find the optimal match. The method proposed by Hickerson *et al.* (2006) was to rank the taxa by one of the key summary statistics, which makes the problem computationally tractable. However, there is then the problem of which summary statistic to use, and if the statistics are not strongly correlated it may not be very efficient. Similar issues have also been encountered in Sousa *et al.* (2009).

Here our approach is to make use of locus-specific summary statistics together with symmetric summary statistics (those that are invariant to locus ordering) in a computationally efficient way, which we now describe. Suppose that we have a hierarchical model in which there are L loci. For the sake of example we concentrate on loci, but the argument can apply to populations or other repeated units. Each locus has a vector of observations (X_i) and (unobserved) parameter vectors κ_i and λ_i ($i = 1, \ldots, L$). Here, we treat λ_i as a parameter of interest and κ_i as a nuisance parameter. We make this distinction for ease of exposition: it is not fundamental to the treatment below. We assume the vector λ_i is of relatively low dimension, while κ_i may be of high dimension. Let $\kappa = (\kappa_1, \ldots, \kappa_L)$ and $\lambda = (\lambda_1, \ldots, \lambda_L)$. The likelihood function for our model is

$$P(X \mid \kappa, \lambda) = \left[\prod_{i=1}^{L} P(X_i \mid \kappa_i, \lambda_i)\right], \tag{5}$$

where $X = (X_1, \ldots, X_L)$. We assume that, conditional on the hyperparameter α, the priors for each locus are independent, and so

$$P(\kappa, \lambda \mid \alpha) = \prod_{i=1}^{L} P(\kappa_i, \lambda_i \mid \alpha). \tag{6}$$

Thus the joint prior density $P(\alpha, \kappa, \lambda)$ is

$$P(\alpha, \kappa, \lambda) = \left[\prod_{i=1}^{L} P(\kappa_i, \lambda_i \mid \alpha)\right] P(\alpha), \tag{7}$$

with a prior (hyperprior) $P(\alpha)$. Because of conditional independence, it is straightforward to show (appendix) that the joint posterior density can be factorized as

$$P(\alpha, \kappa, \lambda \mid X) = \left[\prod_{i=1}^{L} P(\kappa_i, \lambda_i \mid X_i, \alpha)\right] P(\alpha \mid X), \tag{8}$$

or, marginal to the nuisance parameter κ,

$$P(\alpha, \lambda \mid X) = \left[\prod_{i=1}^{L} P(\lambda_i \mid X_i, \alpha)\right] P(\alpha \mid X). \tag{9}$$

Focusing out attention on a single locus i, the hyperparameter α and the locus-specific parameter λ_i have the joint density

$$P(\alpha, \lambda_i \mid X) = P(\lambda_i \mid X_i, \alpha) P(\alpha \mid X). \tag{10}$$

This factorization suggests that we need to use two distinct types of summary statistics in our approximate Bayesian computation: *symmetric* summary statistics, which are functions of all the loci together (*e.g.*, means, higher moments, ...), $S(X) = S(X_1, \ldots, X_L)$; and unit-specific summary statistics, $U(X_i)$. Rather than insisting that the complete set of summary statistics is Bayes sufficient (see Equation 4), we can now make do with the weaker requirement that $S(X)$ and $U(X_i)$ satisfy

$$P(\alpha, \lambda_i \mid X) = P(\lambda_i \mid U(X_i), \alpha) P(\alpha \mid S(X)), \tag{11}$$

at all points (α, λ_i) for the chosen prior (or family of priors). We want this to hold exactly or at least as an adequate approximation. In the terminology of *marginal sufficiency* introduced by Raiffa and Schlaifer (1961, 2000, p. 35) (see also Basu 1977), the factorization (11) tells us that the summary statistic $S(X)$ is marginally sufficient for the parameter α and that the summary statistic ($S(X)$, $U(X_i)$) is marginally sufficient for the locus-specific parameter λ_i. These points motivate two algorithms.

Algorithm 1:

1. For $k = 1$ to $k = N$ iterations:
   i. Sample (A_k, K_k, Λ_k) from the prior $P(\kappa, \lambda \mid \alpha)P(\alpha)$.
   ii. Simulate data X_k (at L loci) from $P(X_k \mid K_k, \Lambda_k)$.
   iii. For locus $i = 1$ to $i = L$ compute $U(X_{k,i})$.
   iv. Compute $S(X_k)$.
2. Use ABC to condition on $S(X) = S(x)$ (approximately) to obtain a sample of observations A^* from $P(\alpha \mid S(x))$ (marginal to κ, λ).
3. For locus $i = 1$ to $i = L$:

Use ABC to condition on $S(X) = S(x)$ and $U(X_i) = U(x_i)$ (approximately) to obtain a sample of observations Λ_i^* from $P(\lambda_i \mid S(x), U(x_i))$ (marginal to α, κ).

Providing the summary statistics are sufficient, and in the limit that the ABC tolerance $\varepsilon \to 0$, this algorithm should sample from the posterior distribution (9) above without additional approximation. There is, however, a practical problem of computer storage associated with this algorithm. If there are u summary statistics in $U(X_i)$, we would need to store NLu items. For example, with 10^3 loci, 10 summary statistics per locus, 10^6 iterations, and 8 bytes per number, we would have 80 Gb of storage as a binary file or in computer memory—much larger, if stored as text files. Thus, although the algorithm may work well with smaller problems there is a generic problem in scaling up.

The second algorithm is similar to sequential ABC algorithms (Sisson *et al.* 2007; Beaumont *et al.* 2009) in which the problem is attacked in two bites.

Algorithm 2:

Step 1. For $k = 1$ to $k = N$ iterations:

   i. Sample (A_k, K_k, Λ_k) from the prior $P(\kappa, \lambda \mid \alpha)P(\alpha)$.
   ii. Simulate data X_k (at L loci) from $P(X_k \mid K_k, \Lambda_k)$.
   iii. Compute $S(X_k)$.

Condition on $S(X) = S(x)$ using ABC, to obtain a sample of observations A^* from

$$P(\alpha \mid S(x)) \approx P(\alpha \mid x).$$

Step 2. For locus $i = 1$ to $i = L$:
For $k = 1$ to $k = N$ iterations:

   i. Sample $A_k{}^{**}$ from $P(\alpha|S(x)) \approx P(\alpha|x)$ by resampling from the observations A^* generated in step 1.
   ii. Sample $(K_{k,i}{}^{**}, \Lambda_{k,i}{}^{**})$ from the conditional prior $P(\kappa_i, \lambda_i \mid A_k{}^{**})$.
   iii. Simulate data $X_{k,i}$ (at locus i only) from $P(X_{k,i} \mid K_{k,i}{}^{**}, \Lambda_{k,i}{}^{**})$.
   iv. Compute $U(X_{k,i})$.

Condition on $U(X_i) = U(x_i)$ using ABC, to obtain a sample of observations $(A^{***}, \Lambda_i{}^{***})$ from an *approximation* to $P(\lambda_i \mid x_i, \alpha)P(\alpha \mid x)$.

Note that in step 2 above, if sample sizes are identical at each locus (no missing data), then it is necessary to iterate only for one locus, rather than for locus $i = 1$ to $i = L$, because the distribution is the same. The advantage of Algorithm 2 over Algorithm 1 is that it scales easily with increasing numbers of loci. The amount of storage is $1/L$ less than Algorithm 1. The time cost of Algorithm 2 is potentially twice as high, but for, *e.g.*, simulated data or data with equal sample size at each locus it is of the same order as that of Algorithm 1. With a computing cluster of many nodes, the overall execution time may be quite low because step 2 in Algorithm 2 can be performed independently for each locus. An additional advantage is that in the second round of simulation the hyperparameter α is already sampled from an approximation to the posterior distribution, and therefore, as with sequential methods (Sisson *et al.* 2007; Beaumont *et al.* 2009; Toni *et al.* 2009), there is a potential for increased precision in our approximation to the posterior distribution of λ_i, ameliorating that apparent inefficiency of having a second round of simulation. However, a key point to note is that Algorithm 2, in contrast to Algorithm 1, involves an approximation that is in addition to that arising from the use of summary statistics that do not satisfy the marginal sufficiency conditions in (11) and nonzero tolerance ε.

To simplify the explanation of this additional approximation, we assume that we are performing ABC on complete data and that, by whatever means, we can sample α from the true posterior distribution. Then in the two-step algorithm, after step 1, we have a sample from

$$P(\lambda_i \mid \alpha)P(\alpha \mid X = x)$$

(marginal to κ_i), where X_i' is the random variable corresponding to the data simulated in the second round. Using ABC we then condition on $X_i' = x_i$. This gives us a sample of observations $(A^{***}, \Lambda_i{}^{***})$ from

$$\frac{P(X_i' = x_i, \lambda_i \mid \alpha)P(\alpha \mid X = x)}{P(X_i' = x_i \mid X = x)},$$

which is not the same as the desired posterior density $P(\lambda_i, \alpha \mid X = x)$.

By contrast, consider a modification of the two-step algorithm, where we sample from $P(\alpha \mid X_{-i} = x_{-i})$ at step 1 [instead of $P(\alpha \mid X = x)$]. (The subscript $-i$ indicates all the data except that from locus i.) Now we have a sample from

$$P(X_i, \lambda_i \mid \alpha)P(\alpha \mid X_{-i} = x_{-i}).$$

If we condition on $X_i' = x_i$, we obtain a sample of observations $(A^{***}, \Lambda_i{}^{***})$ from

TABLE 1

Model parameters and their prior specification

Parameter	Description	Prior distribution
μ_M	Mean scaled migration rate across populations ($\log_{10}$ scale)	$N(a_1 = 0.869, b_1 = 0.521)$
σ_M	Standard deviation of scaled migration rate across populations ($\log_{10}$ scale)	$N(a_2 = 0, b_2 = 0.2)x > 0$
ρ_Z	Probability that a locus is under selection	$\beta(a_3 = 1, b_3 = 20)$
μ_θ	Mean mutation rate across loci ($\log_{10}$ scale)	$N(a_4 = 0.5, b_4 = 0.2)$
σ_θ	Standard deviation of mutation rate across loci ($\log_{10}$ scale)	$N(a_5 = 0, b_5 = 0.2)x > 0$
θ_i	Scaled mutation rate of the ith locus	$\text{Log}_{10}\text{ Normal}(\mu_\theta, \sigma_\theta)$
Z_i	Indicator that is 0 if the ith locus is neutral and 1 if it is selected	$P(Z_i = 1) = \rho_Z$
M_{ij}	Migration rate of the ith locus in population j	See text

$N(a, b)$ refers to a normal density with mean a and standard deviation b.

$$\frac{P(X_i = x_i, \lambda_i \mid \alpha)P(\alpha \mid X_{-i} = x_{-i})}{P(X_i = x_i \mid X_{-i} = x_{-i})} = P(\lambda_i, \alpha \mid X = x)$$

because

$$\begin{aligned}\frac{P(x_i, \lambda_i \mid \alpha)P(\alpha \mid x_{-i})}{P(x_i \mid x_{-i})} &= \frac{P(x_i, \lambda_i \mid \alpha)}{P(x_i \mid \alpha)} \cdot \frac{P(x_i \mid \alpha)P(\alpha \mid x_{-i})}{P(x_i \mid x_{-i})} \\ &= \frac{P(x_i, \lambda_i \mid \alpha)}{P(x_i \mid \alpha)} \cdot \frac{P(x_i, \alpha \mid x_{-i})}{P(x_i \mid x_{-i})} \\ &= P(\lambda_i \mid x_i, \alpha)P(\alpha \mid x_{-i}, x_i) \\ &= P(\lambda_i \mid x, \alpha)P(\alpha \mid x) \\ &= P(\lambda_i, \alpha \mid x).\end{aligned}$$

When the number of loci L is large, we then expect that any one locus i will make an almost negligible contribution to the information about the hyperparameter α, so that

$$P(\alpha \mid x_{-i}) \approx P(\alpha \mid x_{-i}, x_i) = P(\alpha \mid x).$$

Therefore, in this case our two-step algorithm should differ very little from the modified algorithm that can be demonstrated to provide samples from the correct posterior distribution (with ABC error).

APPLICATION TO GENE FREQUENCY DATA

The model: The primary aim of this study was to model local selection and compare the results of the ABC algorithm with the Bayesian method of Beaumont and Balding (2004), which uses an explicit multinomial-Dirichlet function for the likelihood. We wished to investigate the relative efficiency of both methods, using receiver operating characteristic (ROC) analysis. In one case microsatellite data are simulated with low variation in mutation rate among loci, and in the other it is high. It is expected that the multinomial-Dirichlet likelihood will behave poorly in the latter case because it assumes that all genetic variation is ancestral (*i.e.*, it arises in the "collecting phase" of Wakeley 1998). To keep the models similar, we assume an island model. The multinomial Dirichlet arises under an infinite-island or continent-island case (Balding and Nichols 1994; Rannala and Hartigan 1996), but it is pragmatically easier for the ABC analysis to assume a finite number of demes equal to the number of samples. Unlike Beaumont and Balding (2004) we consider only a model in which positive local selection is modeled.

Variation in mutation rate and migration rate is modeled in a hierarchical Bayesian framework, similar in conception to that described in Storz and Beaumont (2002). We assume that there are D demes. The scaled mutation rate at the ith locus is $\theta_i = 2N\mu_i$, where N is the haploid effective size of the deme and μ_i is the mutation rate at the ith locus. The scaled mutation rate, θ_i, has a prior that is a $\log_{10}$-normal distribution with (on a $\log_{10}$ scale) mean μ_θ and standard deviation σ_θ. We use a Gaussian hyperprior for μ_θ and a truncated Gaussian for σ_θ (Table 1). Note that we do not use an inverse gamma for the σ_θ, following the recommendation of Gelman (2006). Variation among loci in migration rate is modeled in a somewhat different way. The principal idea is that there is an indicator Z_i that takes the value 0 if the ith locus is 'neutral' and 1 if it is subject to local selection. The prior for this is Bernoulli with probability ρ_Z that the locus is under selection—*i.e.*, the prior expected number of loci under selection is $L\rho_Z$. The hyperprior for ρ_Z is beta with parameters given in Table 1. Using the approximation of Petry (1983) that local selection acts to reduce the apparent migration rate, we assume that the ith locus and the jth deme have scaled migration rate $M_{ij} = 2Nm_{ij}$, where

$$M_{ij} = \begin{cases} \mathcal{N}_j & \text{if } Z_i = 0 \\ \mathcal{S}_{ij} & \text{if } Z_i = 1. \end{cases}$$

The neutral migration rate varies among demes with a $\log_{10}$-normal prior having (on a $\log_{10}$ scale) mean μ_M and standard deviation σ_M, with Gaussian hyperprior (Table 1). Note that since we have a constant θ across demes, we implicitly assume that variation in M_n across demes is through m and N is constant. For local directional selection we assume that $\mathcal{S}_{ij}$ has a prior given by

a scaled beta distribution with density $\beta(x/\mathcal{N}_j; 1, 1 \cdot 5)/\mathcal{N}_j$. Thus, for a locus under local directional selection the prior migration rate has a maximum equal to the neutral migration rate, but is more heavily weighted toward lower values. The directed acyclic graph (DAG) for this model is given in Figure 1.

In all our examples below, the point values chosen for the parameters of the hyperpriors, $a_1, \ldots, a_6, b_1, \ldots, b_6$, are given in Table 1.

The likelihood function for our model has the form

$$P(X \mid \lambda, \alpha) = \prod_{i=1}^{L} P(X_i \mid \lambda_i, \alpha), \tag{12}$$

where here we implicitly marginalize over the nuisance parameter, κ. Here $X = (X_1, \ldots, X_L)$ with $X_i = (X_{i1}, \ldots, X_{iD})$, and $\alpha = (\mathcal{N}_1, \ldots, \mathcal{N}_D, \rho_Z, \mu_M, \sigma_M, \mu_\theta, \sigma_\theta)$. The locus-specific parameters are $\lambda_i = (\theta_i, M_{i1}, \ldots, M_{iD}, Z_i)$. The joint prior $P(\lambda, \alpha)$ factorizes as in (7). The factor $P(\alpha)$ (the hyperprior) is now of the form

$$P(\alpha) = \left[\prod_j P(\mathcal{N}_j \mid \mu_M, \sigma_M)\right] P(\mu_M; a_1, b_1) P(\sigma_M; a_2, b_2) \cdot P(\rho_Z; a_3, b_3) P(\mu_\theta; a_4, b_4) P(\sigma_\theta; a_5, b_5), \tag{13}$$

and each factor $P(\lambda_i \mid \alpha)$, of the prior density, is of the form

$$P(\lambda_i \mid \alpha) = \left[\prod_j P(M_{ij} \mid Z_i, \mathcal{N}_j; a_6, b_6)\right] P(Z_i \mid \rho_Z) \cdot P(\theta_i \mid \mu_\theta, \sigma_\theta). \tag{14}$$

Each factor $P(X_i \mid \lambda_i)$ of the likelihood function is of the form

$$P(X_i \mid \lambda_i) = \prod_j P(X_{ij} \mid \theta_i, M_{ij}). \tag{15}$$

For a model of this form, with this choice of prior, the marginal posterior density $P(\alpha, \lambda \mid X)$ has a factorization of the form (9), in which α and λ_i are replaced by the parameters of our genetic model, as specified above. Hence our model is amenable to the use of Algorithms 1 and 2.

Summary statistics: The main aim of the model is to characterize the level of genetic differentiation between populations and differences among loci in their levels of differentiation and genetic variability. The choice of summary statistics has then been based on earlier work relating the expected value of summary statistics to demographic parameters and also to work that has used summary statistics of differentiation to identify loci that are potentially under selection (Beaumont and Nichols 1996; Vitalis *et al.* 2001; Excoffier *et al.* 2009). The strategy has been to compute a set of locus-specific summary statistics $U(X_i)$, and then, for the symmetric

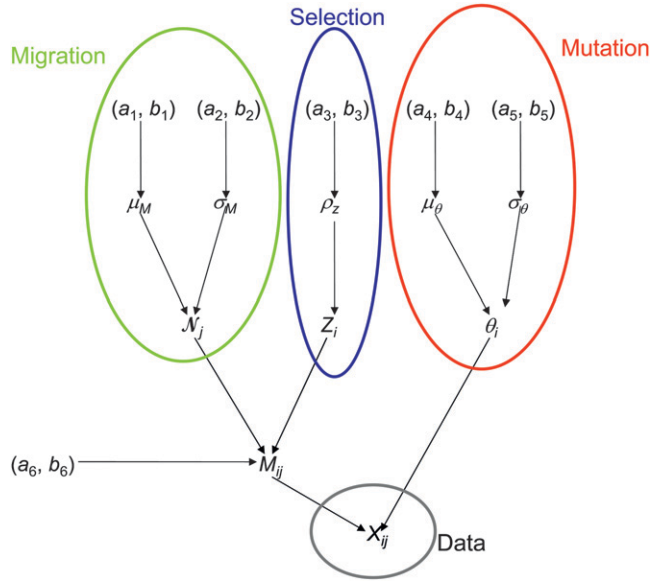

Figure 1.—DAG for the genetic model. See text for details.

summary statistics $S(X)$, the means and other moments of these statistics over loci have been computed.

Locus-specific summary statistics: For each locus we computed the following:

1. The observed probability of nonidentity in state of gene copies between populations, H_B, computed as in Beaumont and Nichols (1996), based on the estimator of Weir and Cockerham (1984).
2. The Weir and Cockerham (1984) estimator of F_{ST}, computed as in Beaumont and Nichols (1996).
3. The logarithm of the variance in microsatellite length between populations. The variance in microsatellite length between populations is $\hat{S}_3$ in Rousset (1996).
4. The statistic $\hat{\rho}_{ST}$ of Rousset (1996), modified from Slatkin (1995), computed as $(\hat{S}_3 - \hat{S}_2)/\hat{S}_3$, where $\hat{S}_2$ is the within-population variance in length, averaged over populations, without weighting for differences in sample size.
5. The variance in the Weir and Cockerham (1984) estimator of F_{ST} estimated for individual alleles (microsatellite lengths). In this case, a locus with K_i alleles was converted into K_i biallelic loci with allele frequencies comprising those of the target allele and all the others combined.
6. In a $K \times D$ table of presence/absence of an allele (microsatellite length) in a population, the proportion of pairwise comparisons between populations in which an allele is observed in at least one of the populations, averaged over alleles. This summary statistic has no previous theoretical basis, but was observed to reduce the mean square error of parameter estimates in simulation tests.
7. The variance of the within-population Weir and Cockerham estimator of F_{ST} (Weir and Hill 2002), computed as in Vitalis *et al.* (2001).

TABLE 2

AUC with 95% C.I. for ABC and BayesFst methods under different scenarios

10^{μ_0}	σ_θ	npop	ρ_Z	N	s	$\bar{F}_{ST}$	AUC (ABC)	AUC (BayesFst)
4	0	6	0.05	400	0.02	0.02	0.498 [0.47, 0.525]	0.499 [0.471, 0.527]
4	0	6	0.05	400	0.02	0.1	0.509 [0.485, 0.533]	0.499 [0.472, 0.525]
4	0	6	0.05	400	0.1	0.02	0.646 [0.615, 0.677]	0.657 [0.624, 0.691]
4	0	6	0.05	4000	0.02	0.02	0.82 [0.79, 0.85]	0.811 [0.781, 0.841]
4	0	6	0.05	400	0.1	0.1	0.887 [0.869, 0.905]	0.887 [0.867, 0.906]
4	0	6	0.05	4000	0.02	0.1	0.935 [0.923, 0.947]	0.94 [0.927, 0.952]
8	0.5	6	0.05	4000	0.1	0.17	0.95 [0.939, 0.96]	0.875 [0.852, 0.897]
4	0	3	0.05	4000	0.1	0.1	0.953 [0.942, 0.964]	0.965 [0.955, 0.974]
8	0.5	6	0.05	4000	0.1	0.1	0.958 [0.947, 0.969]	0.886 [0.863, 0.908]
4	0	6	0.05	4000	0.1	0.1	0.959 [0.948, 0.971]	0.932 [0.917, 0.947]
0.4	0	6	0.05	4000	0.1	0.1	0.962 [0.95, 0.974]	0.972 [0.962, 0.982]
4	0	6	0.01	4000	0.1	0.1	0.974 [0.958, 0.989]	0.983 [0.971, 0.996]
4	0	6	0.05	4000	0.1	0.1	0.974 [0.966, 0.982]	0.981 [0.975, 0.988]
4	0	6	0.1	4000	0.1	0.1	0.977 [0.972, 0.983]	0.985 [0.981, 0.989]
4	0	6	0.05	4000	0.1	0.02	0.989 [0.984, 0.994]	0.992 [0.988, 0.996]

$N\mu$, scaled mutation rate; σ_μ, standard deviation of mutation rate across loci (on $\log_{10}$ scale); npop, number of populations; ρ_Z, proportion of loci under selection; N, subpopulation size; s, selection coefficient. As noted in the text, F_{ST}^j for population j is drawn from a beta distribution with parameters ($a = \bar{F}_{ST}/0.02$, $b = (1 - \bar{F}_{ST})/0.02$). The immigration rate, $\mathcal{N}_j$ in the terminology of our model, is then computed by $\mathcal{N}_j = 1/F_{ST}^j - 1$.

8. The variance of within-population $\hat{\rho}_{ST}$ computed analogously [*i.e.*, as $(\hat{S}_3 - \hat{S}_{2j})/\hat{S}_3$, where $\hat{S}_{2j}$ is computed for each population rather than averaged].

Symmetric summary statistics: To infer hyperparameters we computed 60 symmetric summary statistics $S(X)$, invariant to locus ordering. These included the mean, variance, skew, and kurtosis over loci of the 8 summary statistics above, giving $4 \cdot 8 = 32$ summary statistics, and then the covariance over loci of all 28 pairs of summary statistics.

Transformation of symmetric summary statistics: Previous studies have suggested the use in ABC of transformations, including rotations of the summary statistics (Fagundes *et al.* 2007; Wegmann *et al.* 2009). Because a large number of summary statistics were used, we considered the use of orthogonal transformations of the data to reduce dimensionality. There appear to be two main issues. First, with a large number of summary statistics, many of which are uninformative, a large amount of "noise" is introduced into the computation of distance of simulated data from the observations. Essentially, summary statistics that are unaffected by the parameter values should be weighted out of the distance calculation (Hamilton *et al.* 2005) or not chosen at all (Joyce and Marjoram 2008). Second, there may be a problem of collinearity and resulting instability of the regression once many summary statistics are introduced.

The use of partial least squares (PLS) in an ABC context has been suggested by Wegmann *et al.* (2009). With PLS the orthogonal axes are ordered by decreasing covariance with the independent variable, and it is often used in calibration problems (Gemperline 2007). In our two-step procedure, we need to sample parameters from the joint posterior distribution of hyperparameters, which creates a difficulty because standard PLS assumes a univariate independent variable. A modification of the PLS algorithm exists (PLS-2) for use with a multivariate independent variable. However, we have chosen to use principal component analysis (PCA), also commonly used in calibration and typically producing similar results (Mevik and Wehrens 2007), which orders the axes by decreasing variance. A potential disadvantage of PCA is that axes with small eigenvalues may still have high correlation with the independent variable (here the parameter of interest). To take into account possible correlations between eigenvalues and independent variables, at least marginally, we have defined the following procedure:

The summary statistics sampled from the prior predictive distribution were scaled to have unit variance and rotated (using the R package *Prcomp*).

A Box–Cox transformation was then applied to the resulting eigenvectors.

These were then standardized once more to have unit variance and centered to have zero mean.

The Euclidean distance between these points and the target was computed.

On the basis of the 5% closest points, for the ith component and jth parameter value, the squared correlation coefficient r_{ij}^2 was computed. The components were ranked by the proportion $R_{ij} = r_{ij}^2 / \sum_i r_{ij}^2$. The set of ranked components in which $\sum_i R_{ij} \geq 0.8$ was retained, for each parameter j.

The union was formed over all parameters of the above sets.

The 30 components with the highest eigenvalue were then retained.

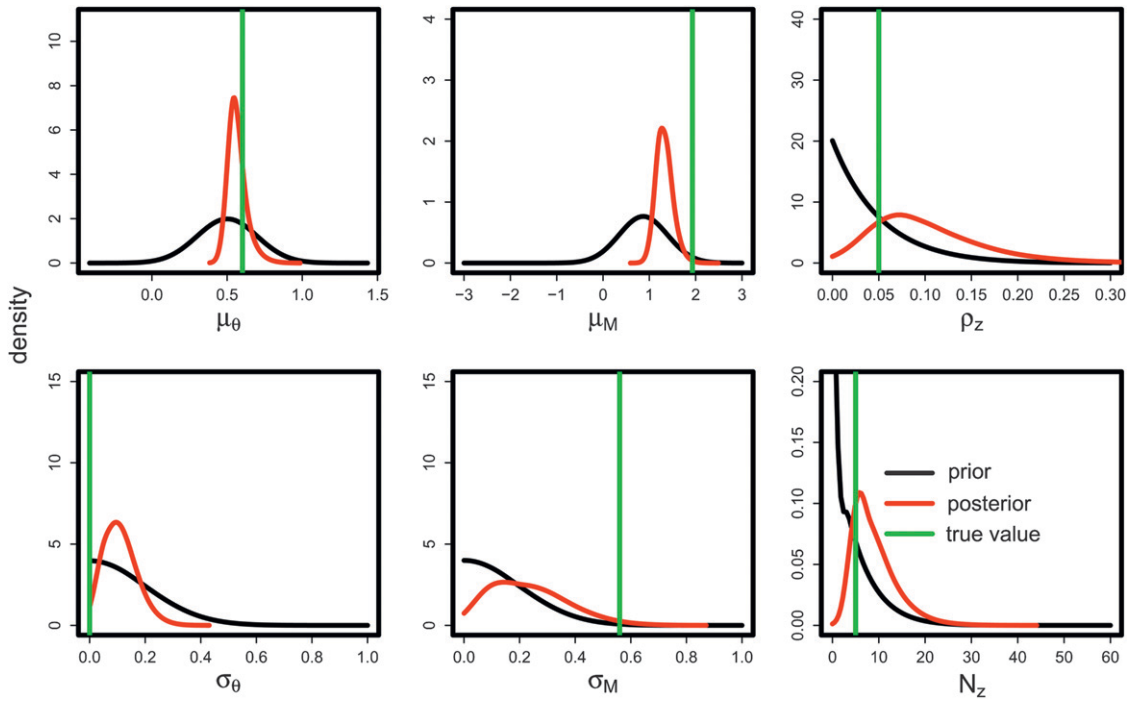

FIGURE 2.—Posterior distribution of genome-wide parameters. The data set contains 100 loci and a sample of 100 gene copies taken from six demes. Five loci are under selection. The data are simulated under the last scenario listed in Table 2.

The regression-based ABC method was then applied (as outlined in Equations 1–3) with $P_\varepsilon = 0.02$.

No claim is made that the above procedure is optimal, and it was obtained through trial and error, on the basis of simulated data with known parameter values. A particular feature of the approach is that there appears to be reduced sensitivity to the addition or removal of summary statistics. The locus-specific summary statistics were used in ABC regression without rotation or further transformation.

The algorithm: Our inference procedure is divided into two steps. We initially approximate the posterior distribution of the higher-level parameters using $S(X)$, and we then approximate the posterior distribution for locus-specific parameters using $U(X)$, as outlined in the following ABC algorithm, based on algorithm 2 above:

1. Compute symmetric summary statistics from the data.
2. Sample the following:
   a. ρ_Z, μ_M, σ_M, μ_θ, σ_θ;
   b. θ_i, Z_i, $\mathcal{N}_j$;
   c. M_{ij}.
3. Run a coalescent simulation of an island model (described in BEAUMONT and NICHOLS 1996; BEAUMONT and BALDING 2004) to obtain data sets X_{ij}.
4. From the simulated data sets, compute the symmetric summary statistics from the X_{ij} in the same way as for step 1 above.
5. Return to step 2 until n sets of summary statistics are obtained.
6. Perform regression ABC (as outlined in the preceding section) to obtain $P_\varepsilon n$ samples from the posterior distribution, where P_ε is the proportion of points accepted.
7. For each locus i:
   a. Compute locus-specific summary statistics from the data for this locus.
   b. In the following order:
      i. Sample with replacement from the $P_\varepsilon n$ samples generated at step 6, ρ_Z, μ_M, σ_M, μ_θ, σ_θ.
      ii. Sample θ_i, Z_i.
      iii. Sample $\mathcal{N}_j$ for $j = 1, \ldots, D$.
      iv. Sample M_{ij} for $j = 1, \ldots, D$.
      v. Sample X_{ij} for $j = 1, \ldots, D$.
      vi. Compute the locus-specific summary statistics as for step 7a above.
      vii. Return to step 7b until n sets of summary statistics are obtained.
   c. Perform ABC one locus at a time (this time measuring locus-specific summary statistics).

PERFORMANCE

To examine the performance of the ABC approach we simulated groups of 100 data sets for 15 different combinations of parameters (scenarios), chosen to vary widely under the assumed prior (Table 1). An archive (suitable only for running on a cluster) containing source code, scripts, and input files for repeating and

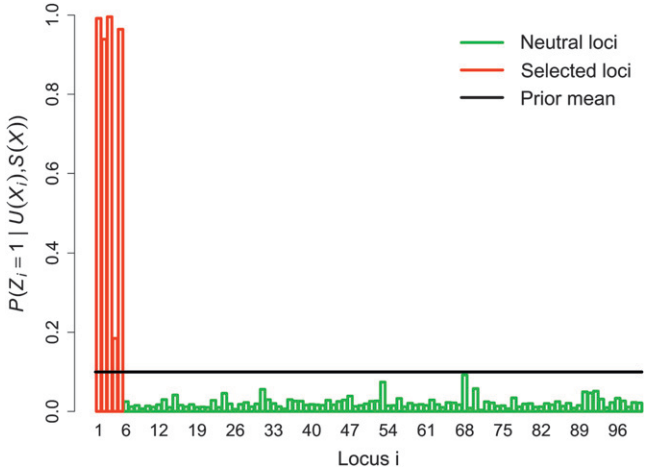

Figure 3.—Estimates of the posterior probability for a microsatellite locus to be under selection, $P(Z_i = 1 \mid U(X_i), S(X))$. The first five loci in red are effectively simulated under selection. The other loci in green are neutral. The data are simulated under the last scenario listed in Table 2.

checking the results presented here is available at http://www.rubic.reading.ac.uk/~mab/stuff/ABCsims.zip. These scenarios included selection coefficients of 0.02 and 0.1. We used ROC analysis, implemented in the ROCR package (Sing *et al.* 2005), as in Riebler *et al.* (2008), to compare the ABC method with BayesFst (Beaumont and Balding 2004). In the case of the ABC method the classifying variable is the posterior probability of locus *i* being under selection, $P(Z_i = 1 \mid U(X_i), S(X))$, while in the case of BayesFst it is a Bayesian *P*-value (Beaumont and Balding 2004). Specifically, for the case here, the *P*-value we use is the posterior probability that the locus effect, α, is less than or equal to zero and hence is a one-tailed *P*-value, for consistency with the ABC model, rather than two tailed as in Beaumont and Balding (2004). We then compute 1 – *P*-value so that values close to 1 indicate selection. In the ROC analysis (see, *e.g.*, Fawcett 2006 for further information) we determine the proportion of false positives and true positives for each value of the threshold that is used to determine whether the classifying variable indicates a locus under selection. This yields a monotonic curve with no positives (true or false) at one end and all positives at the other. If a method has no classification power, the curve should be linear with slope 1, and the area under the ROC curve (AUC) should be 0.5. If a method has perfect classification power, the AUC should be 1.

We simulated data sets using the program that was used to simulate data sets under selection in Beaumont and Balding (2004). This simulates an island model and allows a certain proportion of loci to have alleles that are under selection: either locally positively selected or under balancing selection. We simulated scenarios with six demes (as in Beaumont and Balding 2004) and 100 independent loci and with 100 gene

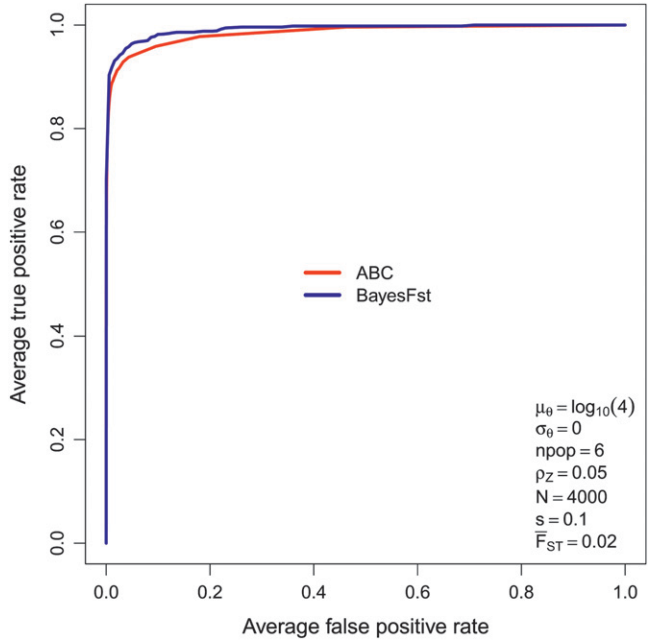

Figure 4.—A comparison of ROC curves for the ABC method (red) and BayesFst (blue). The curves are based on average true positive and false positive rates measured on 100 simulated data sets. The data are simulated under the last scenario listed in Table 2 (parameter values are also shown in legend).

copies taken from each deme. In all simulations the migration rate varied among demes with individual population F_{ST}'s drawn from a beta distribution (see Table 2 legend). This leads to an approximately Gaussian distribution of $\log_{10} \mathcal{N}_j$, as assumed in the model. We

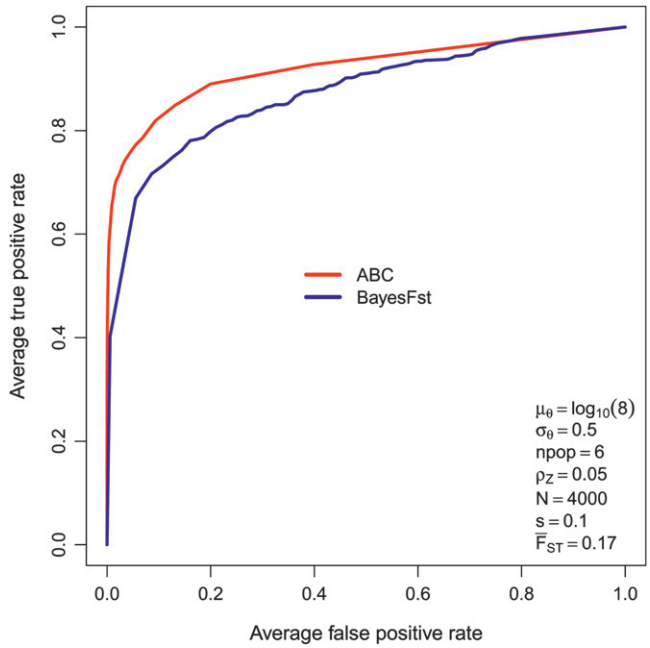

Figure 5.—A comparison of ROC curves for the ABC method (red) and BayesFst (blue). The mutation rate varies across loci. The data are simulated under the 7th scenario listed in Table 2 (parameter values are also shown in legend). Other details are as in Figure 4.

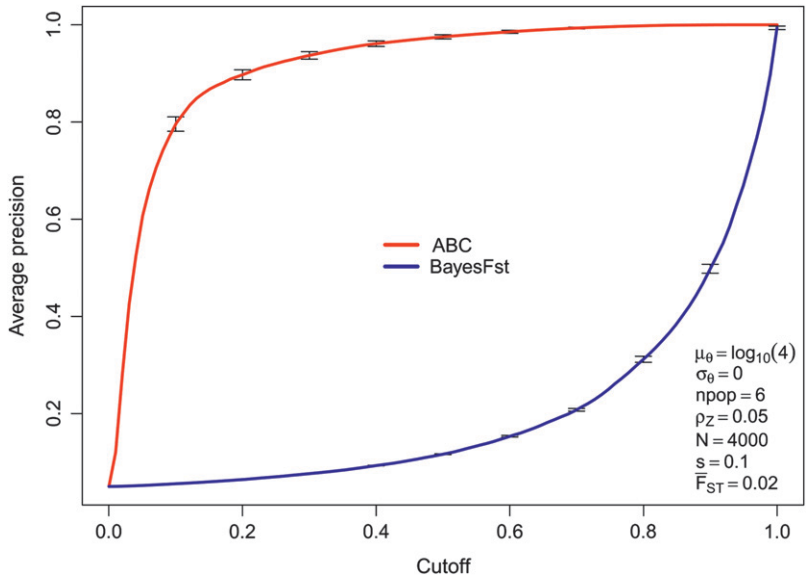

FIGURE 6.—The precision (1 − false discovery rate) is plotted against the classification cutoff (*i.e.*, posterior probability or 1 − *P*-value) used in the ABC and BayesFst method. The data from the last scenario listed are used (see also Figure 4).

tested 15 scenarios (Table 2). Each scenario consisted of 100 replicates (*i.e.*, the total number of simulated loci in Table 2 is 150,000). The data sets were analyzed with the ABC algorithm described above and compared with BayesFst. In the ABC analysis 500,000 iterations were used for both the genome-wide parameter estimation $P(\alpha \mid S(X))$ and the locus-specific parameter estimation $P(\lambda_i \mid U(X_i), S(X))$. For the rejection step, we used the 2% nearest points.

An illustration of the application of the method is given in Figures 2 and 3, which are based on one of the data sets generated for the ROC analysis (scenario 15 in Table 2). Figure 2 shows the posterior distribution of genome-wide parameters and Figure 3 shows the posterior probability $P(Z_i = 1 \mid U(X_i), S(X))$ for each locus. In this example it can be seen that the loci that were simulated to be under selection generally have a higher posterior probability to be under selection, and the posterior mode of the number of loci inferred to be under selection, $\sum Z_i$, is close to the true number of 5, and unsurprisingly ρ_Z has a mode of ~0.05. The demographic parameters are inferred somewhat less well in this example and reflect the influence of the chosen prior. The scaled mutation rate is well estimated, but the inferred value of the scaled migration rate is generally rather too low and weighted toward the prior. The posterior distribution for the variance in mutation rate is broad and tends to follow the prior. The estimated variance among demes in migration rate is rather low and strongly influenced by the prior. The goodness of fit of the model can be examined by seeing how well the symmetric summary statistics $S(X)$ computed from the data fit within the prior predictive distribution (see also RATMANN *et al.* 2009). Since a principal components rotation is used it is relatively straightforward to visualize the fit of the model by plotting the distribution along each axis. An example, using *x*–*y* plots of a selection of axes, is given in supporting information, Figure S1. Unsurprisingly, since the data are simulated from the same model used in the analysis, there is a very good fit.

Overall, in the ROC analysis of the 15 scenarios (Table 2), the performance of the ABC method is quite competitive with BayesFst for both $s = 0.02$ and $s = 0.1$. Although the ABC method often has a slightly lower AUC, the difference is marginal and of the order of the confidence interval. However, in the two scenarios in which there is variability in mutation rate there is superior performance of the ABC method, well beyond the confidence limits of the AUC estimates. Representative numbers, corresponding to rows of Table 2, are given in Figures 4 and 5. The confidence limits are not plotted because they lie close to the estimates. The difference in performance for variable mutation rate arises because the multinomial-Dirichlet model of BayesFst assumes the mutation rate to have a negligible effect on variance in gene frequencies between demes. Thus, when the mutation rate is variable, it contributes to additional variance in gene frequencies between demes, which in BayesFst is attributed to local selection.

Figure 6 shows that, at least for this scenario for the ABC method, the precision, which is 1 – (false discovery rate), initially increases rapidly with the posterior probability that the locus is under selection (the "cutoff") and then smooths off with a false discovery rate of <20% after a posterior probability of ~0.2. By contrast, the BayesFst classifier shows the inverse behavior—it is most sensitive to change in the classification threshold nearer 1. This difference is not surprising, given the different nature of the methods, but suggests that, with the ABC model, posterior probabilities >0.2 are potentially "interesting." It should be noted that unlike, for example, RIEBLER *et al.* (2008), who used a uniform prior, we have explicitly chosen a prior that gives most weight to $P(Z_i = 0)$.

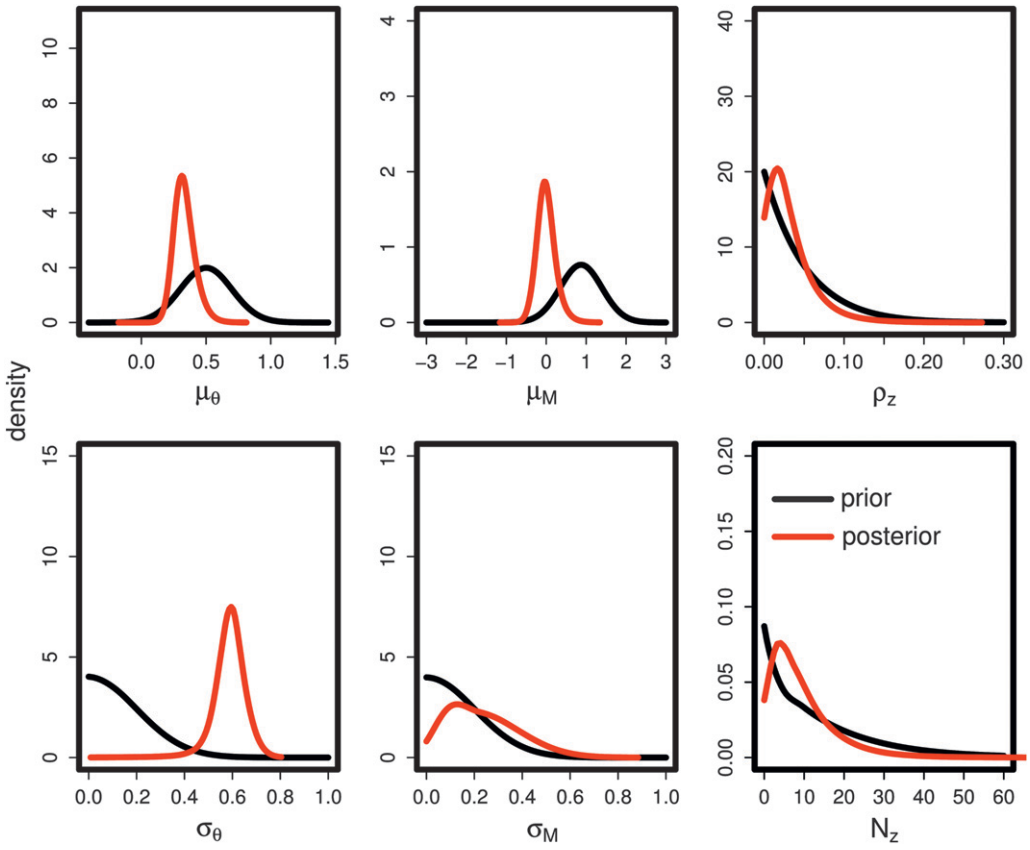

FIGURE 7.—Marginal posterior distributions of hyperparameters for the chimpanzee data.

AN EXAMPLE APPLICATION

We analyzed microsatellite data obtained from a survey of chimpanzee populations from western and central Africa, published by BECQUET *et al.* (2007). These data consist of frequencies sampled in 84 chimpanzees that have been genotyped at 309 microsatellite loci. The study by Becquet *et al.* used clustering methods and identified "western," "central," and "eastern" groups. Of these, we used 64 individuals that had precise designations of location (rather than inferred genetically), giving sample sizes, respectively, of 41, 16, and 7. The gene frequencies were then analyzed using BayesFst and our ABC method (Figures 7 and 8).

The goodness of fit of the ABC simulation can be analyzed by comparing the observed summary statistics to the prior predictive distribution (Figure S2). In this case the data are often on the outer edges of the prior predictive distribution in some projections, but are not markedly outlying. The marginal posterior distributions obtained for the hyperparameters (Figure 7) indicate that gene flow is very low, in line with the conclusions of BECQUET *et al.* (2007). This is a scenario in which it is expected that differences in mutation rate among microsatellites will have a major impact on estimates of F_{ST}. This is indeed observed: the BayesFst analysis yields a large number of positives (Figure 8), which, on the basis of the ROC analysis and theoretical expectations, are likely to be mainly erroneous. By contrast, the ABC analysis suggests that there are possibly two interesting loci, with posterior probabilities >0.2. This conclusion is based on the results from the simulated data sets above (see Figure 6 and related text for rationale). The estimates of posterior probabilities in the ABC analysis shown in Figure 8 have generally low standard errors (on the logit scale), which indicates a reasonable goodness of fit. If the real data are outliers under the model, then the regression step in ABC is an extrapolation, and estimates tend to have very large standard errors. The microsatellites identified by the ABC analysis are GATA81B01 and ATA28C05. The former has not been mapped in *Pan troglodytes*, but is located on the sixth chromosome in *H. sapiens*. The latter has been mapped on the X chromosome of *P. troglodytes* and its nearest ORF is LOC739998 of unknown function.

DISCUSSION

The main contribution of this article has been to demonstrate how one can apply ABC-based models to complex scenarios where the number of summary statistics necessarily scales with the number of parameters in the model. By treating these cases within the hierarchical Bayesian framework, we show how it is possible to deal with quite complicated problems in a computationally feasible way.

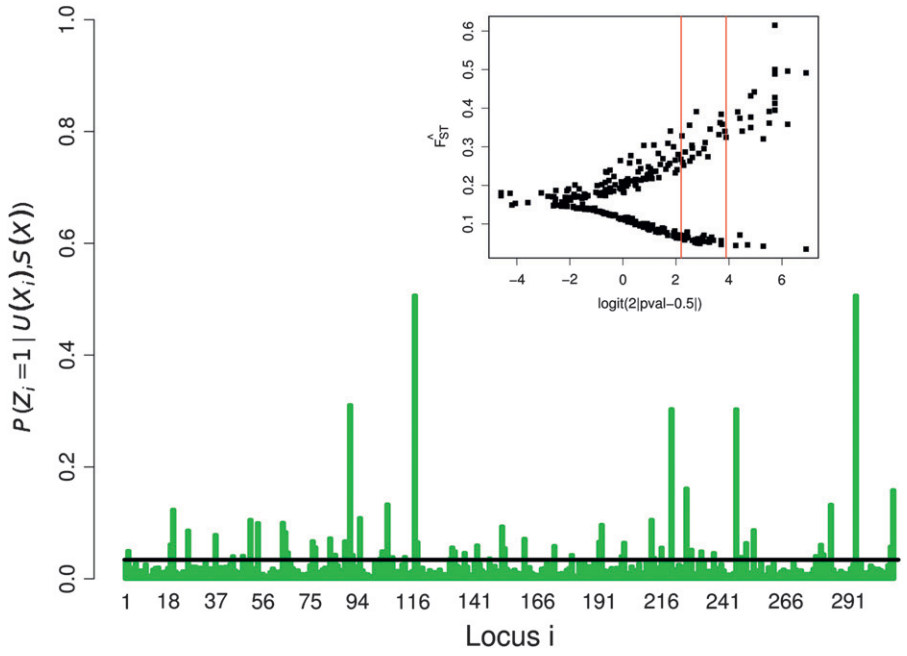

FIGURE 8.—The posterior probability that a locus in the chimpanzee data is under selection, under the ABC model. Inset is the result of an analysis with BayesFst.

We have introduced two algorithms. Both are based on the idea that two types of summary statistics are computed from the data: symmetric summary statistics $S(X)$ used to infer the hyperparameters and those that are unit specific, $U(X_i)$, used to infer parameters. Although, in our treatment, the $S(X)$ are simple functions of the $U(X_i)$, it should be noted that there is no necessity for consistency in, or any formal relationship between, the summary statistics that are used for inferring the hyperparameters and those for inferring the parameters. This distinction is essentially irrelevant providing that the posterior distribution of α is sufficiently accurately approximated. Algorithm 1 is simpler and has the theoretical advantage of sampling from the correct posterior distribution in the limit of zero tolerance and sufficient statistics. However, it suffers from quite significant problems of storage. This may not be an issue in the longer term as computing resources become more extensive. At the present time, storage is certainly an issue to consider when the number of units (loci, individuals, etc.) is >100. Algorithm 2 avoids this storage problem. However, this algorithm involves an approximation (additional to the use of summary statistics in place of the complete data). In Algorithm 2, the second round of simulation will improve the precision in estimates of $P(\lambda_i|U(X_i), S(X)) = \int_\alpha P(\lambda_i|U(X_i), \alpha)P(\alpha|S(X))d\alpha$ because it samples α from $P(\alpha|S(X))$ rather than from $P(\alpha)$. Therefore it may be possible to accept a high proportion of simulated observations, while using a relatively small tolerance. Looking at the problem from the perspective of importance sampling (as in BEAUMONT *et al.* 2009; TONI *et al.* 2009), it is inviting to consider the weight necessary to correct the error in Algorithm 2. It is straightforward to show that the weight is inversely proportional to

$$P(X_i \mid \alpha) = \int_{\lambda_i} P(X_i \mid \lambda_i) P(\lambda_i \mid \alpha) d\lambda_i.$$

That is, if each observation k in step 2 (v) of Algorithm 2 is given a weight that is inversely proportional to the marginal likelihood $P(X_i \mid A_k)$, the resulting weighted sample will be drawn from the correct distribution. Unfortunately the quantity $P(X_i \mid \alpha)$ is not in general easy to compute (otherwise there would be no need to recourse to ABC!). Our main argument in favor of Algorithm 2 is that the approximation will be very slight when the number of units (loci) is large, and scenarios when the number of units is low can be handled by Algorithm 2. The modification, 2a, to Algorithm 2, which is exact, would also be infeasible, requiring separate simulations of step 1 for each locus. Experiments (not shown) with toy simulations based on a beta-binomial model suggest that even with 2 units the approximation in Algorithm 2 is good. With the beta-binomial the ABC can be simulated exactly, the weight above can be computed, and Algorithms 1, 2, and 2a can be easily performed and compared.

One potential criticism of the comparison between our ABC approach and that of BEAUMONT and BALDING (2004) is that one uses a model-choice framework and the other is based on Bayesian *P*-values. Thus it might be argued that we have confounded an intrinsic advantage of the model-choice framework with good performance of ABC. However, with low, nonvariable mutation rates there appears to be relatively little difference in performance of the various approaches to detecting selection that are based on differences in gene frequency. For example, BEAUMONT and BALDING (2004) showed that the difference in performance of the moment-based method of BEAUMONT and NICHOLS (1996) was rela-

tively slight. Riebler *et al.* (2008), who reformulated the model of Beaumont and Balding (2004) into an explicit model-choice framework, demonstrated by means of ROC analysis only a small improvement. Small improvements are also found in Foll and Gaggiotti (2008) and Guo *et al.* (2009). Therefore we argue that the similar performance of BayesFst and the ABC approach with low, nonvariable mutation rates and the better performance of the ABC method with high and variable mutation rates are not biased by an intrinsic superiority of the model-choice framework.

An additional criticism of our model is that we have not included the ability to detect balancing selection, which is present in the methods of Beaumont and Balding (2004), Foll and Gaggiotti (2008), and Riebler *et al.* (2008). Although it would be straightforward to implement, it was not an aim of this study. It is unlikely that by failing to implement a balancing selection component, we have thereby artificially increased the power of the ABC approach in comparison with the multinomial-Dirichlet model. Since the signal of local selection is increased variance in allele frequencies among demes, these would not be placed in a balancing selection category anyway. We note that the attempt to use low F_{ST} as a signal of balancing selection is logically somewhat problematic. If a locus is truly under balancing selection, it is unlikely that the selection coefficients will be identical in each population. Thus we might typically expect the selection coefficients to vary among populations so the equilibria should vary among populations. For populations with relatively high migration rates it is conceivable that loci under balancing selection may have elevated F_{ST} relative to neutral expectation.

By assuming that the scaled mutation rate θ is the same in all demes (while allowing for varying scaled migration rate, $\mathcal{N}_j$), we tacitly assumed constant effective size N in each deme. This may be considered somewhat unrealistic, and a future improvement to the model would allow for variation in deme size. This would be preferable to variable θ because then one could include covariance between $\mathcal{N}$ and θ through shared N. Variability in effective size over time could also be considered. Such improvements may reduce the discrepancies observed in the fit of the model to the chimpanzee data (Figure S2). An advantage of the explicit model-based approach advocated here is that it is relatively easy to examine model discrepancy (see Ratmann *et al.* 2009 for detailed discussion).

In addition to the modeling of potential candidates of balancing selection, further improvements to our demographic model could include, within the island model framework, the number of demes as a parameter to be inferred. This is potentially important when considering the effects of mutation on gene frequencies. For example, in the case of an infinite-allele, finite-island model with F_{ST} defined as in Rousset (1996) we have

$$F_{ST} \approx \frac{1}{1 + (D/(D-1))4Nm + 4N\mu}$$

(for small m and μ). A locus with a higher mutation rate is therefore expected to have reduced F_{ST} but the strength of the effect depends on the deme size, N. Information about the mutation rate is provided by the metapopulation heterozygosity, H_T, which depends *both* on the deme size *and* on the number of demes because

$$H_T = \frac{\theta_M}{1 + \theta_M},$$

where

$$\theta_M \approx 4DN\mu\left(1 + \frac{1}{(D/(D-1)4Nm)}\right).$$

Therefore we expect that very highly heterozygous loci will have reduced F_{ST}, potentially leading to false positives for balancing selection (Beaumont 2008; Excoffier *et al.* 2009), but the amount that F_{ST} is reduced for a given level of heterozygosity depends on the number of demes in the metapopulation. If the number of demes is large, but they have small size, then an elevated mutation rate may have little effect on F_{ST}.

Further extensions of the model may include more general migration matrices, and range expansion, to allow for isolation-by-distance effects [necessary to model human demography (Prugnolle *et al.* 2005)]. It would also be necessary to consider more general mutation models to allow analysis of sequence data. Much of this could be achieved by the use of general-purpose packages (Rambaut and Grassly 1997; Hudson 2002; Laval and Excoffier 2004). To demonstrate the utility of the approach we have applied it to the problem of detecting loci under selection. It is important, however, to emphasize that not all problems can be handled as straightforwardly by a Bayesian hierarchical model, for example, when conditional independence cannot be assumed. There are other areas of application of our ABC method, including population assignment in a more realistic genealogical setting. Its use in fields outside population genetics can also be envisaged.

We are grateful for the constructive critique of two anonymous referees. This work was supported by Biotechnology and Biological Sciences Research Council (BBSRC) grant BBS/B/12776 to M.B. and K.D. and Engineering and Physical Sciences Research Council grant EP/C533550/1 to M.B. E.B. acknowledges grant ANR 07-BDIV-003 (Emerfundis project) for further support. Rothamsted Research receives grant-aided support from the BBSRC.

LITERATURE CITED

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theor. Popul. Biol. **63:** 221–230.

Balding, D. J., and R. A. Nichols, 1994 DNA profile match probability calculations: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci. Int. **64:** 125–140.

Barton, N., and B. Bengtsson, 1986 The barrier to genetic exchange between hybridising populations. Heredity **56:** 357–376.
Basu, D., 1977 On the elimination of nuisance parameters. J. Am. Stat. Assoc. **72:** 355–366.
Beaumont, M., 2008 Selection and sticklebacks. Mol. Ecol. **17:** 3425–3427.
Beaumont, M. A., and D. J. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. **13:** 969–980.
Beaumont, M. A., and R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. Ser. B Biol. Sci. **263:** 1619–1626.
Beaumont, M. A., W. Zhang and D. J. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025–2035.
Beaumont, M. A., J.-M. Cornuet, J.-M. Marin and C. P. Robert, 2009 Adaptive approximate Bayesian computation. Biometrika **96:** 983–990.
Becquet, C., and M. Przeworski, 2007 A new approach to estimate parameters of speciation models with application to apes. Genome Res. **17:** 1505–1519.
Becquet, C., N. Patterson, A. C. Stone, M. Przeworski and D. Reich, 2007 Genetic structure of chimpanzee populations. PLoS Genet. **3:** 10.
Blum, M., and O. François, 2010 Non-linear regression models for approximate Bayesian computation. Stat. Comput. **20:** 63–73.
Cavalli-Sforza, L., 1966 Population structure and human evolution. Proc. R. Soc. Lond. Ser. B **164:** 362–379.
Charlesworth, B., M. Nordborg and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. **70:** 155–174.
Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper & Row, New York.
Excoffier, L., T. Hofer and M. Foll, 2009 Detecting loci under selection in a hierarchically structured population. Heredity **103:** 285–298.
Fagundes, N. J. R., N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano *et al.*, 2007 Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. USA **104:** 17614–17619.
Fawcett, T., 2006 An introduction to roc analysis. Pattern Recognit. Lett. **27:** 882–891.
Foll, M., and O. Gaggiotti, 2008 A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics **180:** 977–993.
Gelman, A., 2006 Prior distributions for variance parameters in hierarchical models. Bayesian Anal. **1:** 515–533.
Gemperline, P. (Editor), 2007 *Practical Guide to Chemometrics*, Ed. 2. Springer, Berlin/Heidelberg, Germany.
Guo, F., D. K. Dey and K. E. Holsinger, 2009 A Bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. J. Am. Stat. Assoc. **104:** 142–154.
Hamilton, G., M. Currat, N. Ray, G. Heckel, M. Beaumont *et al.*, 2005 Bayesian estimation of recent migration rates after a spatial expansion. Genetics **170:** 409–417.
Hickerson, M. J., and C. P. Meyer, 2008 Testing comparative phylogeographic models of marine vicariance and dispersal using a hierarchical Bayesian approach. BMC Evol. Biol. **8:** 322.
Hickerson, M. J., G. Dolman and C. Moritz, 2006 Comparative phylogeographic summary statistics for testing simultaneous vicariance. Mol. Ecol. **15:** 209–223.
Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.
Joyce, P., and P. Marjoram, 2008 Approximately sufficient statistics and Bayesian computation. Stat. Appl. Genet. Mol. Biol. **7:Article 26** .
Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.
Kolmogorov, A. N., 1942 Determination of the centre of dispersion and degree of accuracy for a limited number of observation. Izv. Akad. Nauk. USSR Ser. Mat. **6:** 3–32.
Laval, G., and L. Excoffier, 2004 Simcoal 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics **20:** 2485–2487.
Lewontin, R., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74:** 175–195.
Loader, C. R., 1996 Local likelihood density estimation. Ann. Stat. **24:** 1602–1618.
Marjoram, P., J. Molitor, V. Plagnol and S. Tavaré, 2003 Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **100:** 15324–15328.
McVean, G., and C. C. A. Spencer, 2006 Scanning the human genome for signals of selection. Curr. Opin. Genet. Dev. **16:** 624–629.
Mevik, B. H., and R. Wehrens, 2007 The pls package: principal component and partial least squares regression in R. J. Stat. Softw. **18:** 1–24.
Petry, D., 1983 The effect on neutral gene flow of selection at a linked locus. Theor. Popul. Biol. **23:** 300–313.
Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16:** 1791–1798.
Prugnolle, F., A. Manica and F. Balloux, 2005 Geography predicts neutral genetic diversity of human populations. Curr. Biol. **15:** R159–R160.
Raiffa, H., and R. Schlaifer, 1961 *Applied Statistical Decision Theory*. Harvard University Press, Cambridge, MA.
Raiffa, H., and R. Schlaifer, 2000 *Applied Statistical Decision Theory*. Wiley Classics Library. John Wiley & Sons, New York.
Rambaut, A., and N. Grassly, 1997 Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13:** 235–238.
Rannala, B., and J. A. Hartigan, 1996 Estimating gene flow in island populations. Genet. Res. **67:** 147–158.
Ratmann, O., C. Andrieu, C. Wiuf and S. Richardson, 2009 Model criticism based on likelihood-free inference, with an application to protein network evolution. Proc. Natl. Acad. Sci. USA **106:** 10576–10581.
Riebler, A., L. Held and W. Stephan, 2008 Bayesian variable selection for detecting adaptive genomic differences among populations. Genetics **178:** 1817–1829.
Rousset, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142:** 1357–1362.
Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature **449:** 913–918.
Sing, T., O. Sander, N. Beerenwinkel and T. Lengauer, 2005 ROCR: visualizing classifier performance in R. Bioinformatics **21:** 3940–3941.
Sisson, S. A., Y. Fan and M. M. Tanaka, 2007 Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **104:** 1760–1765.
Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457–462.
Sousa, V. C., M. Fritz, M. A. Beaumont and L. Chikhi, 2009 Approximate Bayesian computation without summary statistics: the case of admixture. Genetics **181:** 1507–1519.
Spencer, C. C. A., and G. Coop, 2004 Selsim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics **20:** 3673–3675.
Storz, J. F., and M. A. Beaumont, 2002 Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. Evolution **56:** 154–166.
Teshima, K. M., G. Coop and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? Genome Res. **16:** 702–712.
Toni, T., D. Welch, N. Strelkova, A. Ipsen and M. Stumpf, 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface **6:** 187–202.
Vitalis, R., K. Dawson and P. Boursot, 2001 Interpretation of variation across marker loci as evidence of selection. Genetics **158:** 1811–1823.
Wakeley, J., 1998 Segregating sites in Wright's island model. Theor. Popul. Biol. **53:** 166–174.

Wegmann, D., C. Leuenberger and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics **182:** 1207–1218.
Weir, B. S., and C. Cockerham, 1984 Estimating f-statistics for the analysis of population structure. Evolution **38:** 1358–1370.
Weir, B. S., and W. G. Hill, 2002 Estimating f-statistics. Annu. Rev. Genet. **36:** 721–750.
Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen and W. G. Hill, 2005 Measures of human population structure show heterogeneity among genomic regions. Genome Res. **15:** 1468–1476.
Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Communicating editor: R. Nielsen

APPENDIX: FACTORIZATION OF THE POSTERIOR DENSITY

In this appendix, we derive the factorization (8), and hence (9), under assumptions that are slightly more general than those set out in (6) and hence (5). In fact we continue to assume that the joint prior $P(\kappa, \lambda, \alpha)$ factorizes as in (6), but we assume that the likelihood function $P(X|\kappa, \lambda, \alpha)$ for our model has the factorization (12). Note that here, α is also a parameter of the model. This formulation covers the special case where the parameter λ_i is simply a function of κ_i and α.

From the factorization (12) of the likelihood function, and the factorization (6) of the prior density, it follows that the joint density $P(\alpha, \kappa, \lambda, X)$ has the factorization

$$P(\alpha, \kappa, \lambda, X) = \left[\prod_{i=1}^{L} P(X_i \mid \kappa_i, \lambda_i, \alpha) P(\kappa_i, \lambda_i \mid \alpha)\right] P(\alpha). \tag{A1}$$

The marginal density $P(\alpha, X)$ is therefore

$$P(\alpha, X) = \left[\prod_{i=1}^{L} P(X_i \mid \alpha)\right] P(\alpha), \tag{A2}$$

where

$$P(X_i \mid \alpha) = \int_{\kappa} \int_{\lambda} P(X_i \mid \kappa_i, \lambda_i, \alpha) P(\kappa_i, \lambda_i \mid \alpha) d\kappa d\lambda. \tag{A3}$$

Dividing (A1) by (A2) we have

$$\begin{aligned} P(\kappa, \lambda \mid \alpha, X) &= \frac{P(\alpha, \kappa, \lambda, X)}{P(\alpha, X)} \\ &= \prod_{i=1}^{L} \left[\frac{P(X_i \mid \kappa_i, \lambda_i, \alpha) P(\kappa_i, \lambda_i \mid \alpha)}{P(X_i \mid \alpha)}\right] \\ &= \prod_{i=1}^{L} P(\kappa_i, \lambda_i \mid \alpha, X_i). \end{aligned} \tag{A4}$$

Substituting the right-hand side of (A4) into the factorization

$$P(\alpha, \kappa, \lambda \mid X) = P(\kappa, \lambda \mid \alpha, X) P(\alpha \mid X), \tag{A5}$$

we obtain the factorizations (8) of the posterior density $P(\alpha, \kappa, \lambda \mid X)$.