Detecting and Measuring Selection from Gene Frequency Data

Renaud Vitalis,*,†,1 Mathieu Gautier,*,† Kevin J. Dawson,‡ and Mark A. Beaumont§

*Institut National de la Recherche Agronomique, Unité Mixte de Recherche CBGP, (Inra, Ird, Cirad, Montpellier-SupAgro) 34988 Montferrier-sur-Lez Cedex, France, †Institut de Biologie Computationnelle, 34095 Montpellier Cedex, France, ‡Cancer Genome Project, The Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, United Kingdom, §Department of Mathematics and School of Biological Sciences, University of Bristol, Bristol BS8 1TW, United Kingdom

ABSTRACT The recent advent of high-throughput sequencing and genotyping technologies makes it possible to produce, easily and cost effectively, large amounts of detailed data on the genotype composition of populations. Detecting locus-specific effects may help identify those genes that have been, or are currently, targeted by natural selection. How best to identify these selected regions, loci, or single nucleotides remains a challenging issue. Here, we introduce a new model-based method, called SelEstim, to distinguish putative selected polymorphisms from the background of neutral (or nearly neutral) ones and to estimate the intensity of selection at the former. The underlying population genetic model is a diffusion approximation for the distribution of allele frequency in a population subdivided into a number of demes that exchange migrants. We use a Markov chain Monte Carlo algorithm for sampling from the joint posterior distribution of the model parameters, in a hierarchical Bayesian framework. We present evidence from stochastic simulations, which demonstrates the good power of SelEstim to identify loci targeted by selection and to estimate the strength of selection acting on these loci, within each deme. We also reanalyze a subset of SNP data from the Stanford HGDP–CEPH Human Genome Diversity Cell Line Panel to illustrate the performance of SelEstim on real data. In agreement with previous studies, our analyses point to a very strong signal of positive selection upstream of the *LCT* gene, which encodes for the enzyme lactase–phlorizin hydrolase and is associated with adult-type hypolactasia. The geographical distribution of the strength of positive selection across the Old World matches the interpolated map of lactase persistence phenotype frequencies, with the strongest selection coefficients in Europe and in the Indus Valley.

IN the new era of population genomics, surveys of genetic polymorphism ("genome scans") offer the opportunity to distinguish locus-specific from genome-wide effects at many loci (Black *et al.* 2001). Reliable inference of demography and phylogenetic history depends on being able to identify putative neutral regions of the genome, which are assumed to be influenced by genome-wide effects only (Ross *et al.* 1999). Conversely, detecting locus-specific effects may help identify those genes that have been, or still are, targeted by natural selection (Luikart *et al.* 2003). Such genes may be involved, for example, in the adaptation to new environments or in the arms race with pathogens (Nielsen 2005). The applications for population genomic analyses therefore cover a wide range of disciplines. Although some progress has been made (Nielsen 2001), the problem of how best to identify regions, loci, or single nucleotides that have been, or are currently, targets of selection remains challenging.

Tests of selective neutrality have been developed for samples drawn from single populations. Most of them are based on the comparison of some summary statistics of the site-frequency spectrum (*i.e.*, the observed distribution of gene frequencies) to their expected distribution from diffusion theory under an infinitely many sites mutation model (Bustamante *et al.* 2001; Payseur *et al.* 2002; Nielsen *et al.* 2005b; Williamson *et al.* 2005). Accounting for different classes of markers (*e.g.*, selected and neutral) is achieved by a Poisson random field (PRF) approximation (Sawyer and Hartl 1992), which assumes independent mutation and selection parameters across sites (see, *e.g.*, Kim and Stephan 2002; Bustamante *et al.* 2003). In particular, Bustamante *et al.* (2003) developed a hierarchical

Copyright © 2014 by the Genetics Society of America
doi: 10.1534/genetics.113.152991
Manuscript received May 8, 2013; accepted for publication December 9, 2013; published Early Online December 20, 2013.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.152991/-/DC1
[1] Corresponding author: Centre de Biologie pour la Gestion des Populations, Campus International de Baillarguet, CS 30 016, 34988 Montferrier-sur-Lez Cedex, France. E-mail: vitalis@supagro.inra.fr

PRF model that allows the estimation of selection coefficients at a set of DNA polymorphisms sampled in a single population (see also Nielsen *et al.* 2005a). Williamson *et al.* (2005) used a similar approach to infer selection in a nonequilibrium demographic model. Yet, they assume *a priori* which mutations are selectively neutral and which are not. The putatively neutral class of markers is then used to infer demographic parameters and, given these estimates, inferences regarding selection are performed on the other class of markers.

Other tests of selective neutrality are based on haplotype structure. Focusing on haplotypes at a locus of interest (referred to as "core haplotypes"), Sabeti *et al.* (2002) analyzed the decay of gene identity as a function of distance from the core, as measured by the extended haplotype homozygosity (EHH). Core haplotypes that have both a high population frequency and a high EHH are evidence of recent positive selection. Deriving the expected distribution of the EHH requires making strong assumptions about the underlying population history, however, which makes it difficult to evaluate the significance of observed values. Several extensions have therefore been proposed and adapted to genome-wide scans of single nucleotide polymorphism (SNP) data (reviewed in Gautier and Vitalis 2012), based on the empirical distribution of EHH-like statistics either for single populations (Voight *et al.* 2006) or for pairs of differentiated populations (Tang *et al.* 2007). Such approaches therefore rely on the assumption that most SNPs behave neutrally, so that the observed distribution of the statistics provides a proxy to the null distribution.

When markers are genotyped across multiple populations, it has been advocated that signatures of natural selection may simply be identified in the extreme tails of the empirical distribution of F_{ST} estimates (Goldstein and Chikhi 2002). The rationale is that loci that are involved in adaptation to local environmental conditions are expected to show unusually high levels of differentiation among populations. Conversely, loci that are under balancing selection are expected to show unusually low levels of differentiation. These model-free approaches are highly computationally efficient and, therefore, have early been applied to large data sets of tens to hundreds of thousands of SNPs, such as the Perlegen (Hinds *et al.* 2005) and the HapMap (International HapMap Consortium 2003, 2005) data in humans (see, *e.g.*, Akey *et al.* 2002; Weir *et al.* 2005; Barreiro *et al.* 2008). Model-free approaches implicitly assume that most of the markers analyzed are selectively neutral, however, and the choice of a decision criterion to characterize "outlier loci" is fairly subjective. These methods are intended to be immune to arbitrary assumptions about the (unknown) demographic history of the sample. Dependence upon the unknown demography (including the geographic and historical relationship among populations) was indeed a severe criticism of Lewontin and Krakauer's model-based test of selective neutrality (Lewontin and Krakauer 1973), which relies on the sampling distribution of the parameter F_{ST} (Robertson 1975; Nei and Maryuyama 1975).

Refinements of this controversial test (see, *e.g.*, Beaumont and Nichols 1996) were based on the properties of gene genealogies in structured populations (Beaumont 2005), which, in many cases, tend toward a simple structure (Nordborg 1997; Wakeley 1999). In this simple structure, the recent genealogy of genes in each local population can be separated from the ancestral genealogy of the whole metapopulation (Wakeley 2004). If this separation-of-timescale approximation holds, the gene frequency distribution in each local population may be approximated as a Dirichlet-multinomial distribution (Balding and Nichols 1995; Balding 2003). Otherwise, it may be necessary to model demography explicitly (Nielsen *et al.* 2009), to restrict the analysis to pairwise comparisons (Vitalis *et al.* 2001), or to account for the covariance matrix of the population allele frequencies (Bonhomme *et al.* 2010; Coop *et al.* 2010; Frichot *et al.* 2013; Günther and Coop 2013; Guillot *et al.* 2014). Several likelihood-based approaches that take advantage of the Dirichlet-multinomial distribution of gene frequencies have been proposed, generally within a Bayesian framework. In particular, Beaumont and Balding (2004) proposed decomposing F_{ST} into locus-, population-, and locus-by-population components in a hierarchical model. Their model offered a new way to formulate the problem of inferring which loci are targets of selection. This framework was further extended by Riebler *et al.* (2008), Foll and Gaggiotti (2008), Guo *et al.* (2009), and Gompert and Buerkle (2011). Yet, these approaches (as most F_{ST}-based methods aimed at looking for locus-specific effects on F_{ST} estimates; see Beaumont and Nichols 1996) are typically not designed to identify population-specific selection. Furthermore, many species have a hierarchical structure, where a subset of populations share a recent ancestry, or exchange more migrants, relative to the full species range. For such complex structures, extensions of the Beaumont and Nichols (1996) test or the classical Lewontin and Krakauer (1973) test have been proposed (see Excoffier *et al.* 2009; Bonhomme *et al.* 2010, respectively). As an alternative to models based on the Dirichlet-multinomial distribution of gene frequencies, Coop *et al.* (2010) developed an extension to the truncated Gaussian model proposed by Nicholson *et al.* (2002), which accounts for the pattern of covariance in allele frequencies between populations (see also Frichot *et al.* 2013; Günther and Coop 2013; Guillot *et al.* 2014). The resulting multivariate normal distribution of gene frequencies allows testing a linear effect of some environmental variable on the allele frequency at some loci (Hancock *et al.* 2010, 2011; Frichot *et al.* 2013).

A major limitation of the methods based on comparisons among populations, however, is that they do not quantify selection. Rather, they are constructed as tests of departure from selective neutrality (Gautier *et al.* 2010). One exception, though, lies in Bazin *et al.* (2010), where the effective migration rate at a marker locus is expressed as a function of the selection coefficient at a positively selected locus and the recombination rate between the two (see Petry 1983). While neutrality is a convenient null hypothesis, a proper interpretation of the observed patterns of variability, in particular the extent to which the neutral model is applicable, requires methods that rely on nonneutral models (see, *e.g.*, Donnelly *et al.* 2001). Furthermore, proper tests of selection should provide estimates of the parameters of interest, *i.e.*, the strength and the type of selection acting on segregating polymorphisms.

Here, we provide a new method, called SelEstim, to distinguish neutral from selected polymorphisms and estimate the intensity of selection at the latter. Our model accounts explicitly for positive selection, and we suppose that all marker loci in the data set are responding to selection, to some extent. The method is based on a diffusion approximation for the distribution of allele frequency in a population subdivided in a number of demes that exchange migrants (*i.e.*, an island model; see Wright 1931). The framework for statistical inference from this model consists in a hierarchical Bayesian model (see Gelman *et al.* 2004). We use a componentwise Markov chain Monte Carlo (MCMC) algorithm to sample from the joint posterior distribution of the model parameters. We then evaluate the performance of SelEstim, by means of stochastic simulations. Last, we reanalyze a subset of SNP data from the Centre d'Etude du Polymorphisme Humain (CEPH) Human Genome Diversity Panel (HGDP) (Cann *et al.* 2003) to illustrate the applicability of SelEstim on real data.

Model

Assumptions

We consider an infinite island model where the ith deme consists of N_i diploid individuals and receives immigrants from the whole population at rate m_i. We define the scaled migration parameter in the ith deme as $M_i \equiv 4N_im_i$. We consider biallelic markers, *i.e.*, that only two alleles (denoted by A and a) may occur at a given locus. We denote by p_{ij} the frequency of allele A in deme i at locus j and by π_j the frequency of allele A at the jth locus in the whole population. Since we consider that the population as a whole is made of an infinite number of islands, π_j gives the frequency of allele A in the pool of migrant individuals. The following notations are used hereafter: the vector of allele frequencies in deme i at locus j is $\mathbf{p}_{ij} \equiv (p_{ij},\ 1 - p_{ij})$, and the vector of allele frequencies at locus j among migrants is $\boldsymbol{\pi}_j \equiv (\pi_j,\ 1 - \pi_j)$. We consider a simple genic model of selection where, at each locus, the allele A provides a selective advantage. The homozygote individuals AA and the heterozygotes Aa have a relative increase of fitness of $1 + s_{ij}$ and $1 + s_{ij}/2$, respectively, as compared to the aa homozygotes. We define the scaled coefficient of selection in deme i at locus j as $\sigma_{ij} \equiv 2N_is_{ij}$. We define the indicator variable κ_{ij}, which takes the value $\kappa_{ij} = 0$ if allele A is selected for, and $\kappa_{ij} = 1$ if allele a is selected for. Therefore, the frequency of the selected allele in deme i at locus j reads $\tilde{p}_{ij} \equiv \kappa_{ij}(1 - p_{ij}) + (1 - \kappa_{ij})p_{ij}$.

The data consist of individuals collected in a set of n_d demes and genotyped at L loci. We denote by n_{ij} the total number of genes sampled in the ith deme at the jth locus, out of which x_{ij} have allelic state A. The vector of allele counts in deme i at locus j therefore reads $\mathbf{n}_{ij} \equiv (x_{ij}, n_{ij} - x_{ij})$.

Given the frequencies p_{ij} of allele A, the conditional distribution of allele counts $\mathbf{n}_{ij}$ in population i at locus j is binomial:

$$\mathcal{L}\left(p_{ij};\mathbf{n}_{ij}\right) = \binom{n_{ij}}{x_{ij}} p_{ij}^{x_{ij}}\left(1-p_{ij}\right)^{n_{ij}-x_{ij}}. \tag{1}$$

In the limit of large deme size, as $N_i \rightarrow \infty$, and assuming that selection and random genetic drift are of comparable strength (*i.e.*, that M_i and σ_{ij} have a finite limit as $N_i \rightarrow \infty$), the distribution of the $\mathbf{p}_{ij}$ may be approximated by the stationary density of a diffusion process, which has the form

$$\psi\left(p_{ij};M_i,\sigma_{ij},\kappa_{ij},\boldsymbol{\pi}_j\right) = C^{-1}\exp\left(\sigma_{ij}\tilde{p}_{ij}\right)p_{ij}^{M_i\pi_j-1}\left(1-p_{ij}\right)^{M_i\left(1-\pi_j\right)-1} \tag{2}$$

(Wright 1949; Barton and Turelli 1987; Ethier and Nagylaki 1988; Bürger 2000). In Equation 2, C is the constant that ensures that the distribution integrates to 1. This constant can be evaluated as

$$\begin{aligned} C &= \int \exp\left(\sigma_{ij}\tilde{p}_{ij}\right)p_{ij}^{M_i\pi_j-1}\left(1-p_{ij}\right)^{M_i\left(1-\pi_j\right)-1} d\mathbf{p}_{ij} \\ &= {}_1F_1\left(M_i\tilde{\pi}_{ij};M_i;\sigma_{ij}\right)\frac{\Gamma\left(M_i\pi_j\right)\Gamma\left(M_i\left(1-\pi_j\right)\right)}{\Gamma\left(M_i\right)}, \end{aligned} \tag{3}$$

where ${}_1F_1(a; b; z)$ is the confluent hypergeometric, or Kummer's, function (see, *e.g.*, Abramowitz and Stegun 1965, p. 504), and $\tilde{\pi}_{ij} \equiv \kappa_{ij}(1 - \pi_j) + (1 - \kappa_{ij})\pi_j$.

Given the model specified in Equations 1 and 2, we are interested in evaluating the parameters of interest $\mathbf{M} \equiv (M_1, \ldots, M_i, \ldots, M_{n_\mathrm{d}})$, $\boldsymbol{\pi} \equiv (\pi_1, \ldots, \pi_j, \ldots, \pi_L)$, $\boldsymbol{\sigma} \equiv (\sigma_{11}, \ldots, \sigma_{ij}, \ldots, \sigma_{n_\mathrm{d}L})$, and $\boldsymbol{\kappa} \equiv (\kappa_{11}, \ldots, \kappa_{ij}, \ldots, \kappa_{n_\mathrm{d}L})$, from the observed allele counts $\mathbf{n}$ over all sampled demes and loci. The directed acyclic graph (DAG) for this hierarchical-Bayesian model is shown in Figure 1.

We assume a Bernoulli prior distribution for the parameters κ_{ij}, *i.e.*, $\kappa_{ij} \sim \mathrm{Ber}(0.5)$, and a uniform prior for the $\boldsymbol{\pi}_j$'s, that is $\boldsymbol{\pi}_j \sim \mathrm{Beta}(1,\ 1)$. We further assume a log-uniform prior for the M_i's with support from 0.001 to 10,000; *i.e.*, the priors of the M_i's are uniform in log scale: $\log(M_i) \sim \mathcal{U}(\log(10^{-3}), \log(10^4))$. The prior distributions for the selection coefficients σ_{ij} (at each locus, in each deme) are modeled hierarchically (see, *e.g.*, Gelman *et al.* 2004, pp. 124–125). In particular, we assume that σ_{ij} has an exponential prior distribution $f(\sigma_{ij}|\delta_j) \sim \exp(\delta_j^{-1})$ that depends upon the locus-specific hyperparameter δ_j, which represents the average effect of selection at locus j (over all demes). We further assume that this hyperparameter δ_j has an exponential prior distribution $f(\delta_j|\lambda) \sim \exp(\lambda^{-1})$ that depends, in turn, upon the hyperparameter λ, which represents the genome-wide effect of selection over all demes and loci. Last, we assume that the prior distribution of λ is $f(\lambda) \sim \exp(\Lambda^{-1})$, with $\Lambda = 1.0$, in what follows. Assuming independence of allele frequencies

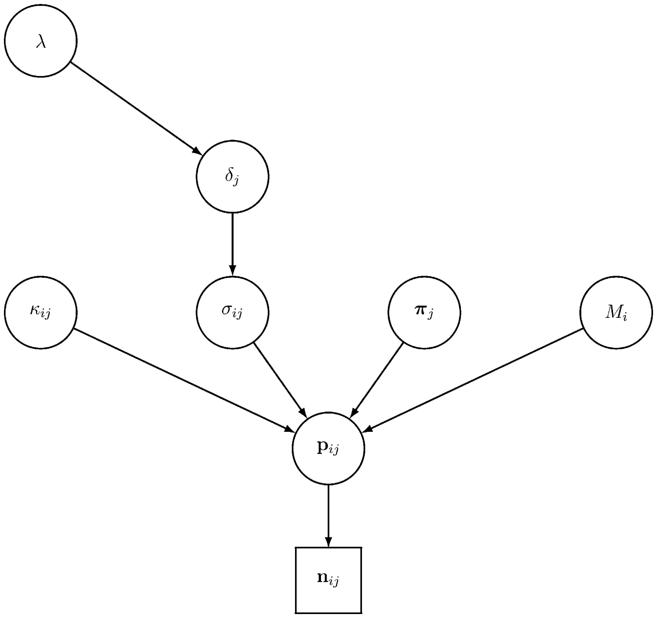

Figure 1 Directed acyclic graph (DAG) of the hierarchical Bayesian model.

among loci and populations, the posterior distribution of the parameters $f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\sigma}, \boldsymbol{\kappa}, \boldsymbol{\delta}, \lambda \,|\, \mathbf{n})$, *i.e.*, the conditional distribution of the parameters $\mathbf{M}$, $\boldsymbol{\pi}$, $\boldsymbol{\kappa}$, $\boldsymbol{\sigma}$, $\boldsymbol{\delta}$, and λ given the data $\mathbf{n}$, depends upon the prior distributions of the parameters and the data as

$$f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\kappa}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \lambda|\mathbf{n}) \propto \prod_{i=1}^{n_\mathrm{d}} \prod_{j=1}^{L} \mathcal{L}\left(p_{ij}; \mathbf{n}_{ij}\right) \psi\left(p_{ij}; M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij}\right) \times f(\mathbf{M}) f(\boldsymbol{\pi}) f(\boldsymbol{\kappa}) f(\boldsymbol{\sigma}|\boldsymbol{\delta}) f(\boldsymbol{\delta}|\lambda) f(\lambda). \tag{4}$$

Although in Equation 4 the likelihood from Equation 1 can be integrated analytically over the distribution of unknown population frequencies given by Equations 2 and 3, we found that it increases the computational burden significantly. This is so because additional gamma and confluent hypergeometric functions are then introduced in the likelihood function $\mathcal{L}'(M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij}, \mathbf{n}_{ij})$.

We implemented a single-component Metropolis–Hastings (or Metropolis within Gibbs) algorithm (see, *e.g.*, Ntzoufras 2009) to sample from the joint posterior distribution of $f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\kappa}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \lambda \,|\, \mathbf{n})$, which is specified by Equation 4. In practice, we therefore update one parameter at each time, iteratively, as detailed in Supporting Information, File S1. The proposal distributions for each of the $\mathbf{M}$, $\boldsymbol{\pi}$, $\boldsymbol{\kappa}$, $\boldsymbol{\sigma}$, $\boldsymbol{\delta}$, and λ parameters are adjusted by means of 25 short pilot runs of 1000 iterations to obtain acceptance rates between 0.25 and 0.40 (see, *e.g.*, Gilks *et al.* 1996).

The software package implementing the model described here, called SelEstim, is available at http://www1.montpellier.inra.fr/CBGP/software/selestim. All the postprocessing statistical analyses were performed using the R software environment for statistical computing, v. 3.0.1 (R Core Team 2013).

Identifying loci under selection from the MCMC outputs

Because the model assumes that each and every locus in a data set is selected to a certain extent, we are particularly interested in the posterior densities of the locus-specific hyperparameters δ_j: we expect the density to be shifted toward zero for neutral markers and to positive values for (presumably) selected loci (see Figure 2). Yet, given the hierarchical structure of our model, it would not be sufficient to simply test whether, at a particular locus, the posterior distribution of δ_j departs from zero. This approach would neglect the genome-wide effects of selection. Since we assume in our model that the δ_j's are drawn independently from a common hyperdistribution with parameter λ (which represents the genome-wide effect of selection), it is indeed more appropriate to compare the posterior distributions of the locus-specific coefficients of selection with the "centering" distribution derived from the hyperdistribution of the genome-wide effect of selection. This centering distribution is a good choice, rather than, for example, a point mass at zero, because it leads to consistent estimates (the hyperprior will shrink toward zero with more data). Furthermore, as becomes apparent, it provides some, albeit very weak, power to identify loci under balancing selection.

Following Guo *et al.* (2009), we consider the following steps to detect outlier loci in a data set: (i) approximate the posterior distributions of the locus-specific selection hyperparameters (δ_j); (ii) compare each of these distributions to a "centering" distribution derived from the hyperdistribution with parameter λ that describes the among-locus variation in the locus-specific effect of selection; and last (iii) measure the mean of the posterior distributions of σ_{ij} for outlier loci over sampled populations to summarize the distribution of selection effects across sampling locations.

Our preliminary analyses suggested that the posterior distribution of the parameters δ_j is unimodal. Furthermore, its support is on $[0, \infty)$, by definition of δ_j. For subsequent analyses, we approximate the posterior distributions of the δ_j's by a gamma distribution with the same mean and variance as estimated from the MCMC sample. In doing so, we assume that slight discrepancies in higher-order moments will not affect the comparison of these distributions with their centering distribution. In the following, we therefore consider that the posterior distribution of δ_j is $\Gamma(k_0, \theta_0)$ with $k_0 = \bar{x}^2_{\delta_j}/s^2_{\delta_j}$ and $\theta_0 = s^2_{\delta_j}/\bar{x}_{\delta_j}$, where $\bar{x}_{\delta_j}$ and $s^2_{\delta_j}$ are the mean and the variance, respectively, of the posterior distribution of δ_j, as estimated from the MCMC outputs. Since we assumed an exponential prior distribution for the hyperparameters δ_j, *i.e.*, $f(\delta_j) \sim \exp(\lambda^{-1})$, or equivalently $f(\delta_j) \sim \Gamma(1, \lambda)$, we defined the centering distribution of δ_j as $\Gamma(1, \theta_1)$, where $\theta_1 = \bar{x}_\lambda$ is the posterior mean of λ, as estimated from the MCMC outputs. We expect that the stronger the intensity of selection at locus j, the larger the departure of the posterior distribution of δ_j from the centering distribution. We use the

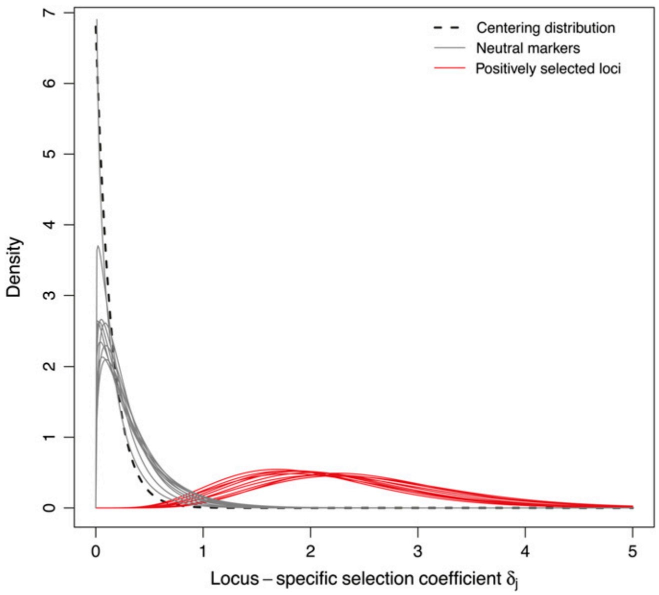

Figure 2 Examples of posterior densities of the locus-specific hyperparameter δ_j for neutral markers (in gray) and positively selected loci (in red). The "centering" distribution (dashed line) is derived from the hyperdistribution of the genome-wide effect of selection, λ, and is defined as an exponential distribution $\sim \exp(\bar{x}_\lambda^{-1})$, where $\bar{x}_\lambda$ is the posterior mean of λ, as estimated from the MCMC.

Kullback–Leibler divergence (KLD) to measure the distance of the posterior distribution of δ_j from the centering distribution. The KLD of a distribution with density $f(x)$ from a distribution with density $g(x)$ is defined as

$$\mathrm{KLD}[f(x)\|g(x)] = \int_{-\infty}^{\infty} f(x)\log\left[\frac{f(x)}{g(x)}\right]\mathrm{dx}. \tag{5}$$

With a little algebra, one can show that the KLD of a gamma distribution with shape and scale parameters (k_0, θ_0) from a gamma distribution with shape and scale parameters (1, θ_1) is given by

$$\mathrm{KLD}[\Gamma(k_0,\theta_0)\|\Gamma(k_1,\theta_1)] = \log\left[\frac{\theta_1}{\Gamma(k_0)\theta_0^{k_0}}\right] + k_0\frac{\theta_0-\theta_1}{\theta_1} + (k_0-1)[\log(\theta_0)+\digamma(k_0)], \tag{6}$$

where $\digamma(x) \equiv \Gamma'(x)/\Gamma(x)$ is the digamma function. Note that as an alternative to these computations, one might use other KLD estimators, such as those based on nearest neighbors (see, *e.g.*, Pérez-Cruz 2008).

To provide a decision criterion for discriminating between neutral and selected markers, we calibrate the KLD using simulations from a predictive distribution based on the observed data set. The motivation here is to generate a set of loci equivalent to those that we observe in their levels of diversity and genetic variation. Note that generating a full posterior predictive distribution for KLD is unhelpful here because extreme *P*-values would indicate poor model fit rather than give evidence of selection *per se*. A key requirement in our calibration is that we generate a distribution of KLD under a null model in which the δ_j for each locus are drawn from their centering distribution. For this reason we cannot use a strictly neutral model (*i.e.*, the beta-binomial parameterized by M_i). A further assumption that we make, for otherwise the approach would be computationally intractable, is that the loci are exchangeable. Thus the KLD computed for each locus in our data set is compared to the simulated distribution of KLD among all the loci generated from the predictive distribution. As described in File S2, we parameterize our predictive distribution using the estimated posterior means for M_i, π_j, κ_{ij}, and λ. We show below that, although the method is somewhat conservative, the calibration based on these posterior means is generally good. Thus, in practice, for each data set and each analysis, we use the algorithm detailed in File S2 to generate pseudo-observed data (POD). We then analyze that POD, using the same MCMC parameters (number and length of pilot runs, burn-in, chain length, etc.) as for the analysis of the original data set. The KLD values computed for each simulated locus are then combined to obtain an empirical distribution. The quantiles of this empirical distribution are computed and are used to calibrate the KLD observed for each locus in the original data: *e.g.*, the 99% quantile of the KLD distribution from the POD analysis provides a 1% threshold KLD value, which is then used as a decision criterion to discriminate between selection and neutrality.

Materials and Methods

Simulation study

We evaluated the performance of the method by simulating artificial data sets for fixed parameter values. The simulations were performed according to an island model with 50 demes, each consisting of $N = 250$ diploid individuals. Following Beaumont and Balding (2004), we simulated allele counts data from a Wright–Fisher model with migration and selection.

Initialization was achieved by means of a Pólya urn scheme simulation of the coalescent (Donnelly and Tavaré 1995) with scaled mutation parameter $\theta \equiv 4N\mu = 1$. This amounts to considering selection acting on standing variation and makes this simulation model similar in spirit to the models considered by Innan and Kim (2004) and Przeworsky *et al.* (2005). At each generation (generations were discrete and nonoverlapping), each individual produced a random number of offspring drawn from a Poisson distribution with mean 100. Mutations then occurred at rate 2×10^{-5}. Dispersal of the (diploid) offspring then occurred at rate m, with dispersing individuals reaching necessarily a distinct deme. Selection of the offspring surviving to adulthood was then achieved, according to the scheme detailed below. A number N of adults was drawn from among the offspring, except if the number of offspring in a deme was $<N$, in which case all

offspring survived. This life cycle was repeated for 25,000 generations. Samples were then taken, but only if the minimum allele frequency (the frequency of the least frequent allele) was >0.01. All loci were considered as independent, so that each multilocus data set was made of independent realizations of that process.

To account for the possibility of positive selection to local environmental conditions, the demes were arbitrarily provided with attributes (blue, red, or uncolored), which were assigned at random, independently for each selected locus. For positively selected loci, one allele B was considered as advantageous in a blue deme (and neutral in a red deme), while the other allele R was considered as advantageous in a red deme (and neutral in a blue deme). Both alleles were considered as neutral in uncolored demes. Therefore, BB homozygotes had fitness $(1 + s)$ in blue demes and 1 in red and uncolored demes; RR homozygotes had fitness $(1 + s)$ in red demes and 1 in blue and uncolored demes; BR heterozygotes had fitness $1 + s/2$ in red and blue demes and 1 in uncolored demes. For loci under balancing selection, only the heterozygote genotypes were selected for in the blue and red demes, with relative fitness $(1 + s)$. Homozygote genotypes were neutral (relative fitness 1) in all demes, as were the heterozygote genotypes in uncolored demes.

A total of 18 data sets were generated using the Wright–Fisher model described above (see Table 1). For each locus, 50 diploid individuals (100 genes) per deme were sampled. The details that distinguish the different data sets are given in Table 1. For example, for data sets 1–9, the samples were taken in 6 demes: 2 blue demes, 2 red demes, and 2 uncolored demes, and each simulated data set consisted in 10,000 SNPs, with 8000 neutral markers, 1000 positively selected loci, and 1000 loci under balancing selection. For data sets with selected loci (data sets 1–11), we assumed that 30% of all demes were blue demes, 30% were red demes, and 40% were uncolored demes. For data sets with neutral markers only (data sets 12–18), 50,000 SNPs were simulated and all demes were uncolored. Table 1 further gives the combinations of F_{ST} and σ values used for the simulations.

An additional set of four simulations was performed to test for the robustness to departure from the island model assumptions. To that end, a classical island model, a hierarchical island model (see, *e.g.*, Excoffier *et al.* 2009), a stepping-stone model in one dimension, and a pure drift model (whereby nine populations diverge sequentially) were simulated (see Figure S1). The sample characteristics (number of individuals, number of sampled demes) were the same for all simulations, and the model parameters were tuned to achieve an overall realized F_{ST} of ≈ 0.24.

For each of these 22 data sets the MCMC algorithm was run to sample from the joint posterior distribution of the model parameters. For each Markov chain, 100,000 updating steps were completed after 25 short pilot runs of 1000 iteration and a burn-in of 25,000 steps. Samples were collected for all the model parameters every 25 steps (thinning) to reduce autocorrelations, yielding 4000 observations.

The data sets 1–11 (see Table 1) were analyzed using BayeScan v. 2.1 (Foll and Gaggiotti 2008) with default option values (except for the MCMC parameters, see below). BayeScan is based on the Dirichlet-multinomial model for allele frequencies in an island model of population structure. At each locus, the distribution of allele frequency in each subpopulation depends on the allele frequency in the common pool of migrants and a subpopulation-specific F_{ST}^{ij} parameter. In BayeScan, as in Beaumont and Balding's model (Beaumont and Balding 2004), the logit of F_{ST}^{ij} is decomposed additively into a locus-specific component (α_i) shared by all populations, and a population-specific component (β_j) shared by all loci. Significantly positive (resp., negative) values of α_i are taken as evidence of positive (resp., balancing) selection. BayeScan is based on a reversible-jump MCMC algorithm, which estimates the posterior probabilities of two alternative models, a purely neutral one ($\alpha_i = 0$), and one including selection ($\alpha_i \neq 0$). Each Markov chain was run for 100,000 updating steps, after 25 short pilot runs of 1000 iteration each and a burn-in of 25,000 steps. Samples were collected every 25 steps (thinning), yielding 4000 observations. For each output, we computed the Bayes factor (BF) for the model including selection ($\alpha_i \neq 0$), assuming prior odds of 10 for the neutral model ($\alpha_i = 0$).

The comparison of the relative efficiency of the two methods was achieved by means of receiver operating characteristic (ROC) analysis (see, *e.g.*, Fawcett 2006, for further information), using the R software environment for statistical computing, v. 3.0.1 (R Core Team 2013).

Application to human data

We applied SelEstim on the Stanford HGDP–CEPH Human Genome Diversity Cell Line Panel (Cann *et al.* 2003) SNP genotyping data, which consist of genotypes at more than 650,000 SNPs (ftp://ftp.cephb.fr/hgdp_supp1). Because we were interested, for illustrative purpose, in measuring the genetic signature of selection in the lactase gene, we considered only the 53,765 SNPs mapping to chromosome 2 (HSA2) and incorporated the genotyping data of the two SNPs reported to be tightly associated with lactase persistence (−13910C → T and −22018G → A) as published by Bersaglieri *et al.* (2004). All the populations with <15 genotyped individuals were discarded from the data set. Furthermore, we removed seven populations from Oceania and Southern America, as well as three populations from Sub-Saharan Africa (the Biaka Pygmies, Mbuti Pygmies, and the Mandenka) that were absent from Bersaglieri *et al.*'s data set (Bersaglieri *et al.* 2004). The two Bantu populations (from Kenya and South Africa) were merged, as in Bersaglieri *et al.* (2004). Finally, we discarded all the SNPs with a minimum allele frequency (MAF) below 0.01 across the populations retained. The final data set consisted in 52,633 SNPs characterized in 23 populations from Africa and Eurasia.

Table 1 Parameters of simulated data sets

						Markers			Sampled demes	
Data set	N	$M \equiv 4Nm$	F_{ST}	s	$\sigma \equiv 2Ns$	Pos.	Bal.	Neut.	No.	Categories
1	250	18.067	0.05	0.02	10	1,000	1,000	8,000	6	(2,2,2)
2	250	18.067	0.05	0.05	25	1,000	1,000	8,000	6	(2,2,2)
3	250	18.067	0.05	0.10	50	1,000	1,000	8,000	6	(2,2,2)
4	250	8.664	0.10	0.02	10	1,000	1,000	8,000	6	(2,2,2)
5	250	8.664	0.10	0.05	25	1,000	1,000	8,000	6	(2,2,2)
6	250	8.664	0.10	0.10	50	1,000	1,000	8,000	6	(2,2,2)
7	250	3.858	0.20	0.02	10	1,000	1,000	8,000	6	(2,2,2)
8	250	3.858	0.20	0.05	25	1,000	1,000	8,000	6	(2,2,2)
9	250	3.858	0.20	0.10	50	1,000	1,000	8,000	6	(2,2,2)
10	250	8.664	0.10	0.05	25	1,000	1,000	8,000	3	(1,1,1)
11	250	8.664	0.10	0.05	25	1,000	1,000	8,000	12	(4,4,4)
12	250	84.604	0.01	–	–	0	0	50,000	6	(0,0,6)
13	250	44.619	0.02	–	–	0	0	50,000	6	(0,0,6)
14	250	18.067	0.05	–	–	0	0	50,000	6	(0,0,6)
15	250	8.664	0.10	–	–	0	0	50,000	6	(0,0,6)
16	250	5.468	0.15	–	–	0	0	50,000	6	(0,0,6)
17	250	3.858	0.20	–	–	0	0	50,000	6	(0,0,6)
18	250	2.888	0.25	–	–	0	0	50,000	6	(0,0,6)

All the simulations were performed according to an island model with n_d = 50 demes, each made of N = 250 diploid individuals (500 genes). Fifty diploid individuals (100 genes) were sampled per deme. The migration rate m was fixed to achieve the desired value of F_{ST}, using Equation 6 in Rousset (1996). The number of loci simulated under positive selection (Pos.), balancing selection (Bal.) and neutrality (Neut.) is indicated. The total number of sampled demes is also given, together with the composition of the sample: (2,2,2) indicates that the sample of six demes consisted of two blue demes, two red demes, and two uncolored demes.

Results

Evaluating the performance of SelEstim:

To asses the KLD calibration procedure, we first evaluate the method using simulated data sets made of neutral markers only (data sets 12–18; see Table 1). Figure 3A shows the false-positive rate, *i.e.*, the proportion of markers that are incorrectly classified as under selection, as a function of various KLD thresholds. For each data set, KLD thresholds based on the quantiles of the KLD distributions from the POD analyses were computed and used as a decision criterion for discriminating between neutral and selected markers. Figure 3B represents the false-positive rate as a function of the quantile probability (comprised between 0 and 1). Figure 3B shows that, for the data sets considered here, the false-positive rate at any KLD threshold is always less than the corresponding quantile probability. This suggests that our calibration procedure is "conservative," at least for the range of F_{ST} values considered here (ranging from 0.01 to 0.25).

Figure 4 shows the performance of the method on data set 5 (see Table 1), which corresponds to $F_{ST} = 0.10$ and $\sigma \equiv 2Ns = 25$. Figure S2, Figure S3, Figure S4, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, Figure S10, and Figure S11 provide the same outputs for simulated data sets 1–4 and 6–11. Figure 4A shows that the distribution of KLD measures for positively selected loci departs from that of the neutral markers and the loci under balancing selection. This is essentially true for the data sets for which $F_{ST} \geq 0.05$ and $\sigma \geq 25$, as can be seen from Figure S2, Figure S3, Figure S4, Figure S5, Figure S6, Figure S7, Figure S8, and Figure S9. Not surprisingly, large KLD values correspond to large F_{ST} estimates (Figure 4B). This is so because for positively selected loci, one allele is selected for in blue populations and the other in red populations, which tends to exacerbate differentiation. A close examination of Figure 4C further shows that the false-positive rate (the proportion of neutral markers that exhibit a signature of selection) at any KLD threshold is always less than the corresponding KLD quantile probability (as in Figure 3B). This strengthens the notion that the KLD thresholds should not be interpreted as frequentist P-values. All else being equal, increasing the number of sampled demes improves the discrimination of neutral and positively selected markers (compare Figure 4, Figure S10, and Figure S11). Last, Figure 4 shows that SelEstim has some weak statistical power to identify loci under balancing selection. This arises because markers with very low divergence will tend to have posteriors for δ that are pushed further toward zero in comparison with the centering distribution and, hence, will also have a higher KLD. Yet, although the KLD for loci under balancing selection is indeed slightly higher, on average, than for neutral markers (Figure 4, Figure S2, Figure S3, Figure S4, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, Figure S10 and Figure S11), it remains very low. This result is not surprising, though, since the selection scheme considered in our model of inference accounts only for positive genic selection. Furthermore, previous simulation studies have also shown that, in the absence of an explicit model of selection, similar methods generally lack power to detect balancing selection (Beaumont and Balding 2004; Foll and Gaggiotti 2008; Riebler *et al.* 2008).

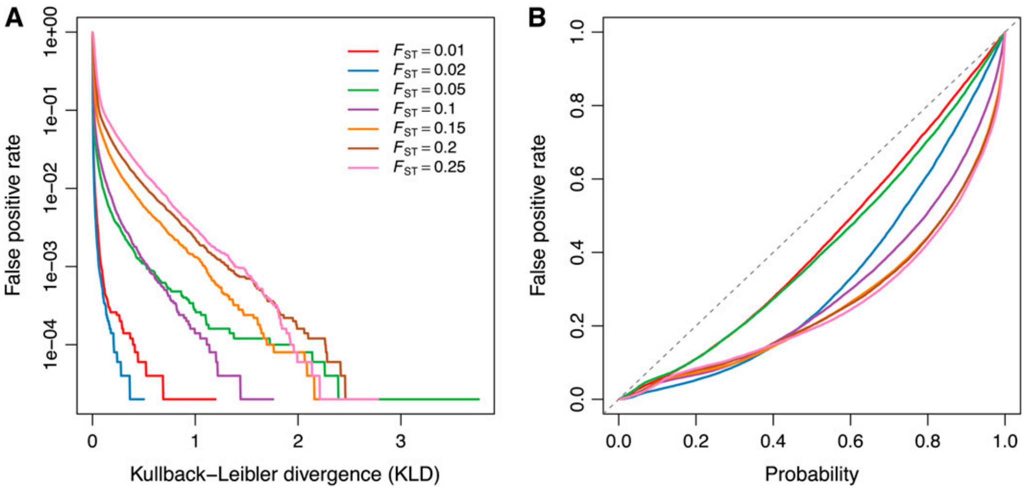

Figure 3 (A) False-positive rate (neutral loci detected as outliers) as a function of the Kullback–Leibler divergence (KLD) threshold, for data sets 12–18. (B) False-positive rate, as a function of the quantile probability. For each data set analysis, pseudo-observed data (POD) are generated from the inference model with hyperparameters λ, π_j, and M_i set to their respective posterior means, using a rejection-sampling algorithm (see File S2). The POD is then analyzed, using the same MCMC parameters (number and length of pilot runs, burn-in, chain length, etc.) as for the analysis of the original data. Each quantile probability defines a KLD threshold, which is used as a decision criterion for discriminating between neutral and selected markers.

Comparison with BayeScan:

Figure 4D shows the relationship between BayeScan BF and the KLD for each and every locus from data set 5. According to Jeffreys' rule (Jeffreys 1961; Kass and Raftery 1995), a BF >10 ($\log_{10}(\text{BF}) \geq 1$) provides "strong" evidence for selection, and for a BF >100 ($\log_{10}(\text{BF}) \geq 2$) the evidence is "decisive." However, it should be noted that, by definition, the BF depends on the prior odds for the neutral model. For this set of simulated data, there is a good agreement between BayeScan BF criterion and the KLD, since markers with high KLD show decisive evidence of selection from BayeScan BF criterion. For BayeScan analyses, whenever the posterior probability of the model including selection ($\alpha_i \neq 0$) was equal to 1, we arbitrarily defined the $\log_{10}$(posterior odd) as $\log_{10}(3999.5/4000) - \log_{10}(0.5/4000)$ to account for the chain length (4000 iterations). By construction, the maximum value that the BF can take ($\log_{10}(\text{BF}) = 4.857$; see Figure 4D) therefore depends on the MCMC length and the prior odds for the neutral model (here equal to 10).

The ROC analysis (Figure 5) further shows a slight power gain of SelEstim over BayeScan (Foll and Gaggiotti 2008). In the ROC analysis, the proportion of false positives and true positives is computed for each possible value of the threshold that is used to classify a locus under selection (see, *e.g.*, Fawcett 2006, for further information). For SelEstim, the classifying variable was the KLD of the posterior distribution of the locus-specific coefficient of selection δ_j from its centering distribution, while in the case of BayeScan it was the Bayes factor. The ROC analysis 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 curve should perfectly superimpose to the left-hand and top sides of the unit square, and the AUC should be 1. Considering positively selected loci first, the area under the ROC curve for SelEstim is slightly larger, and closer to 1, than that obtained for BayeScan (see Figure 5). In particular SelEstim appears to have progressively improved relative performance with decreasing values of σ. As for loci under balancing selection, SelEstim seems slightly better than BayeScan based on the ROC analysis, although both methods lack statistical power in these sets of simulated data.

The analysis of data sets 1–11 (see Table 1) took 12,175 sec on average per data set (SD = 6921) with BayeScan and 12,633 sec (SD = 6149) with SelEstim. SelEstim is therefore 3.85% slower than BayeScan, based on the same MCMC parameters (number and length of pilot runs, burn-in, chain length, etc.) and using the same number of processor cores. Note, however, that the KLD calibration procedure of SelEstim comes at the cost of up to a doubled computing time, due to the additional analysis of the POD.

Robustness to model misspecification

We analyzed simulated data departing from the island model assumptions (see Figure S1). The rate of false positives detected by SelEstim at a given threshold was higher for data simulated from a hierarchical island model or a stepping-stone model, as compared to data simulated from a nonhierarchical model with the same overall F_{ST} (Figure S12). However, for all scenarios considered, the false-positive rate at any KLD threshold was less than or nearly equal to the corresponding quantile probability (Figure S12F), which suggests that our calibration procedure is "conservative," even for strong departures from the island model assumptions. Furthermore, the rates of false positives for these scenarios were much lower as compared to BayeScan analyses of the same data (Figure S13).

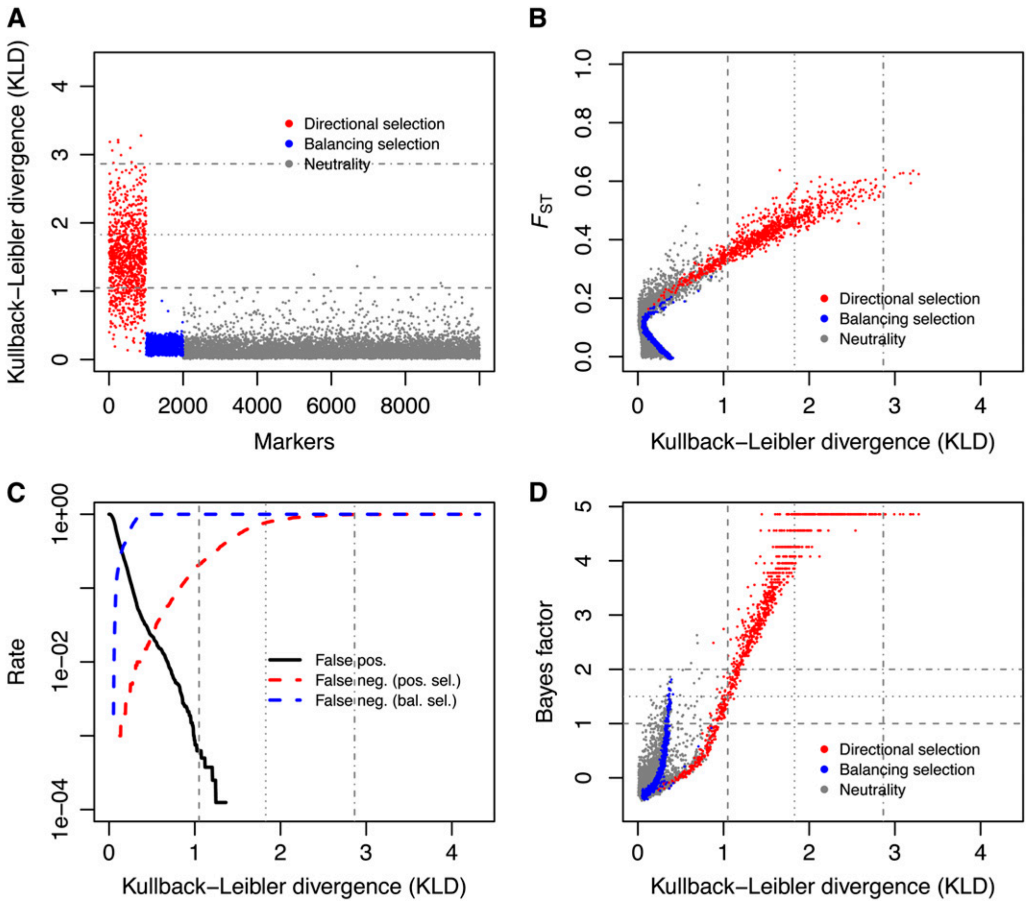

Figure 4 Analysis of the allele count data from data set 5. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in gray. (B) F_{ST} as a function of the KLD measure for all loci. (C) False-positive (neutral loci detected as outliers) and false-negative (selected loci not detected as outliers) rates as a function of the KLD threshold. (D) Relationship between BayeScan Bayes factor $\log_{10}$(BF) and the KLD for all markers in data set 5. Positively selected markers are in red, loci under balancing selection are in blue, and neutral markers are in gray. The horizontal lines in A and the vertical lines in B–D indicate the KLD thresholds corresponding to the 95%, 99%, and 99.9% quantile of the KLD distribution from the POD analysis of data set 5. In D, the horizontal dotted lines indicate the $\log_{10}$(BF) = 1, log_10(BF) = 1.5 and $\log_{10}$(BF) = 2 thresholds, which correspond to strong, very strong and decisive support following Jeffreys's (1961) scale of evidence for selection, respectively.

Inference of selection coefficients

For data set 5 (see Table 1), we examined the distributions of the posterior means of the parameters κ_{ij} (which indicate which allele is selected for), for the 1000 positively selected loci. Here, from the hypotheses of our simulation model, $\kappa_{ij} = 0$ indicates that the blue allele is selected for, and $\kappa_{ij} = 1$ indicates that the red allele is selected for. Figure 6A shows the distributions of the posterior means of κ_{ij} in each sampled deme. Consistent with our expectation, it is apparent from Figure 6A that the posterior means of κ_{ij} in demes 1 and 2 (blue demes) are shifted toward zero and that the posterior means of κ_{ij} in demes 3 and 4 (red demes) are shifted toward one. Alleles of the right color are therefore selected for in the right deme. It is also reassuring to see that in demes 5 and 6 (uncolored demes), the posterior means of κ_{ij} are centered around 0.5, which is consistent with the fact that neither allele should be selected for in these demes.

For the same 1000 positively selected loci, we further examined the posterior means of the scaled coefficients of selection $\sigma_{ij} \equiv 2N_i s_{ij}$, conditionally on κ_{ij}. By doing so, we estimate the coefficient of selection associated with the allele being effectively targeted by selection. Figure 6B shows that the posterior means of $f(\sigma_{ij}|\kappa_{ij} = 0)$ in blue demes and the posterior means of $f(\sigma_{ij}|\kappa_{ij} = 1)$ in red demes are very close to the simulated values ($\sigma \equiv 2Ns$ =25 in data set 5; see Table 1). By contrast, in uncolored demes, the posterior means of σ_{ij} that were not conditioned upon κ_{ij}, are much lower and closer to the prior distribution of the hyperparameter λ (which represents the genome-wide effect of selection over all demes and loci). Figure S14, Figure S15, Figure S16, Figure S17, and Figure S18 reproduce the same outputs as in Figure 6 for data sets 1–4 and 6–11. The posterior means of the scaled coefficients of selection σ_{ij} conditionally on κ_{ij} are very close to the simulated values, for $F_{ST} \geq 0.05$ and $\sigma \geq 25$ (data sets 2–3, 5–6, and 8–9) and $F_{ST} = 0.2$ and $\sigma = 10$ (data set 7). All else being equal, increasing the number of sampled demes improves the estimation of the scaled coefficients of selection σ_{ij} conditionally on κ_{ij} (compare Figures 6 and Figure S18).

Last, we examined the distributions of the posterior means of κ_{ij} for the 8000 neutral markers in data set 5. Figure 7A shows that the posterior means of κ_{ij}, which do not depend on the color of the sampled demes, are all centered around 0.5. This result is consistent with the fact that neither allele should be selected for in these demes. Furthermore, the distributions of the posterior means of κ_{ij} for neutral markers are narrower, as compared to the posterior means of κ_{ij} for selected loci in uncolored demes (see Figure 6A). The distributions of the

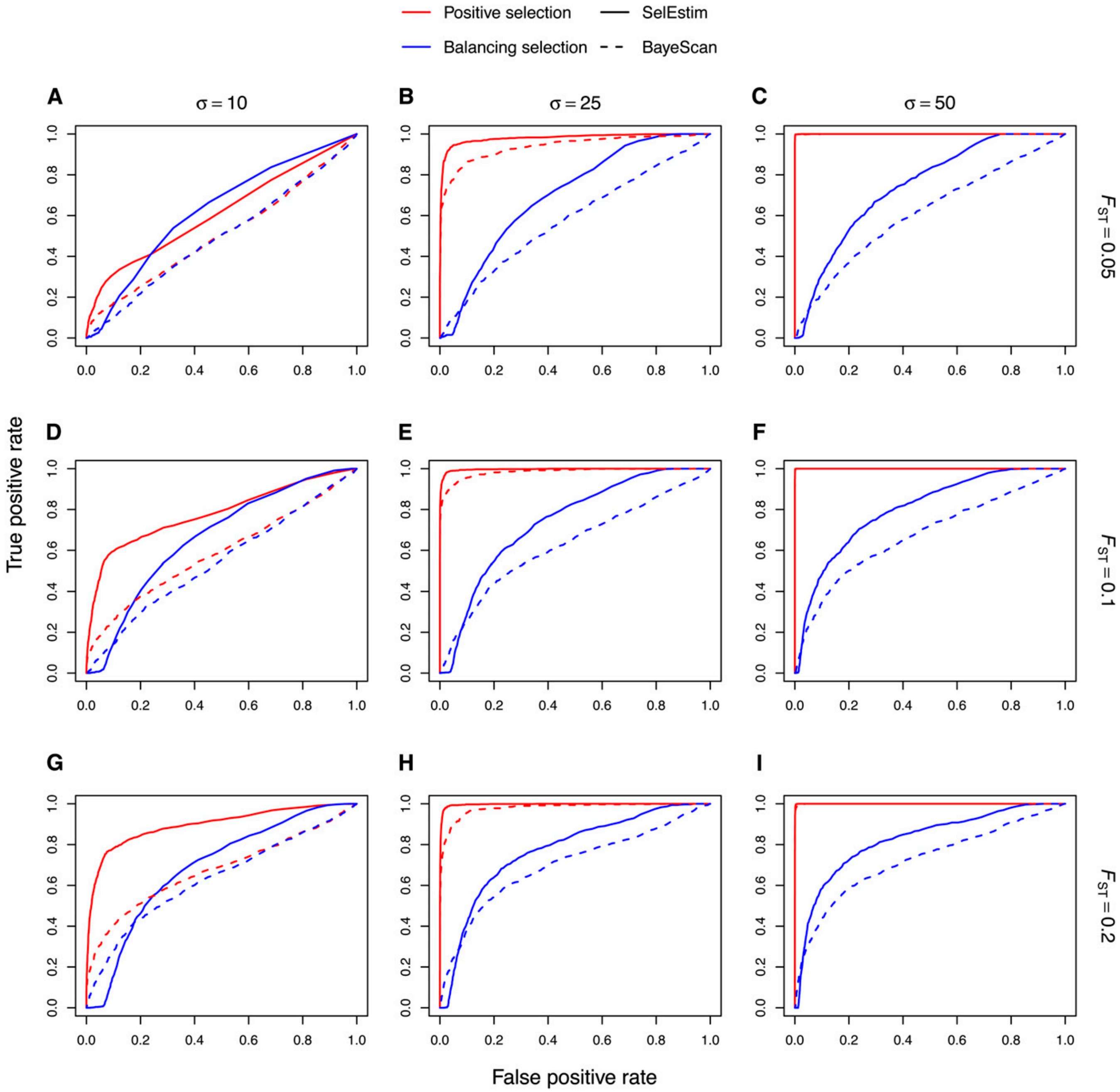

Figure 5 Receiver operating characteristic (ROC) analysis for the data sets 1 (A), 2 (B), 3 (C), 4 (D), 5 (E), 6 (F), 7 (G), 8 (H) and 9 (I). In the ROC analysis, the proportion of false positives and true positives is computed for each possible value of the threshold that is used to classify a locus under selection. For SelEstim, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and its centering distribution, while in the case of BayeScan it was the Bayes factor.

posterior means of κ_{ij} for neutral markers are therefore closer to their prior distribution. Yet, the distributions are still wider than expected from the prior distribution, since the mean over 4000 independent samples (which corresponds to the MCMC length) from a Ber(0.5) distribution should be approximately normally distributed with mean 0.5 and standard deviation 0.008. This extra variance may be due to the hierarchical structure of the model, which produces a correlation between the parameters. In addition, Figure S19A shows that, in the absence of selection (data set 16; see Table 1), the distribution of the posterior means of κ_{ij} is centered around 0.5 and narrower as compared to data sets that include positively selected loci (compare with Figure 7A). Therefore, the higher variance of the distributions of κ_{ij} for selected loci in uncolored demes (as compared to neutral markers) certainly stems from the influence of selection occurring for the same loci in blue and red demes, through the prior on the locus-specific hyperparameter δ_j. The posterior means of σ_{ij} for neutral markers

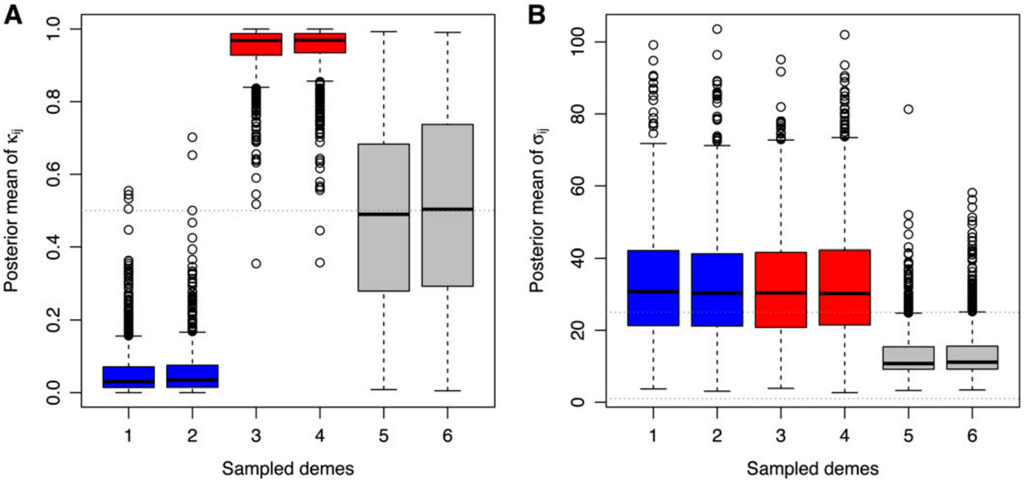

Figure 6 Analysis of the allele count data from data set 5. (A) Boxplot representation of the posterior means of the parameters κ_{ij} (which indicate which allele is selected for) for the 1000 positively selected loci in blue demes (1–2), red demes (3–4), and uncolored demes (5–6). (B) Boxplot representation of the posterior means of the selection coefficients σ_{ij} for positively selected loci in data set 5. For blue demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the blue allele being selected for (κ_{ij} = 0). For red demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the red allele being selected for (κ_{ij} = 1). For uncolored demes, the posterior means of the selection coefficients σ_{ij} are unconditional. The horizontal dashed line gives the true (simulated) value of σ_{ij} = 25.

(unconditionally upon κ_{ij}) are very low and close to the posterior mean of the hyperparameter λ. In the absence of selection (data set 16), the posterior means of σ_{ij} for neutral markers (unconditionally upon κ_{ij}) are also very low and not different from the prior mean of the hyperparameter λ (Figure S19B).

Implicitly, Figure 6 and Figure 7 demonstrate that SelEstim is able to give accurate measures of the scaled coefficient of selection at one locus in different demes and therefore to provide evidence of local adaptation. This paves the way for the inference of the distribution of selection strength across populations in a landscape as is illustrated in the next section.

Last, we analyzed a set of 11 simulations using the same parameters as for data set 5 (see Table 1), but varying the proportion of selected loci from 10 to 5000 of 10,000 markers (hence, from 0.1 to 50%). Interestingly, we found a strong correlation between the posterior mean of the genome-wide coefficient of selection λ and the number of positively selected loci (see Figure S20). However, as the number of positively selected loci increases, the performance of SelEstim weakens (>20% of selected markers), although less markedly than BayeScan (see Figure S21).

Application to human data

We ran three independent MCMC analyses on a subset of the Stanford HGDP–CEPH Human Genome Diversity Cell Line Panel (Cann *et al.* 2003) SNP Genotyping Data. The data consisted in 52,631 SNPs from the HGDP–CEPH data, and two SNPs (−13910C → T and −22018G → A) known to be tightly associated with lactase persistence (Bersaglieri *et al.* 2004), genotyped in 23 populations from Africa and Eurasia. After 25 pilot runs of 1000 iterations, each MCMC was run for 100,000 updating steps, after a burn-in period of 25,000 steps. Samples were collected from the Markov chains for all the model parameters every 25 steps (thinning) to reduce autocorrelations, yielding 4000 samples for each parameter.

Convergence was assessed by computing the multivariate extension of Gelman–Rubin's diagnostic (Brooks and Gelman 1998) on the three independent Markov chains.

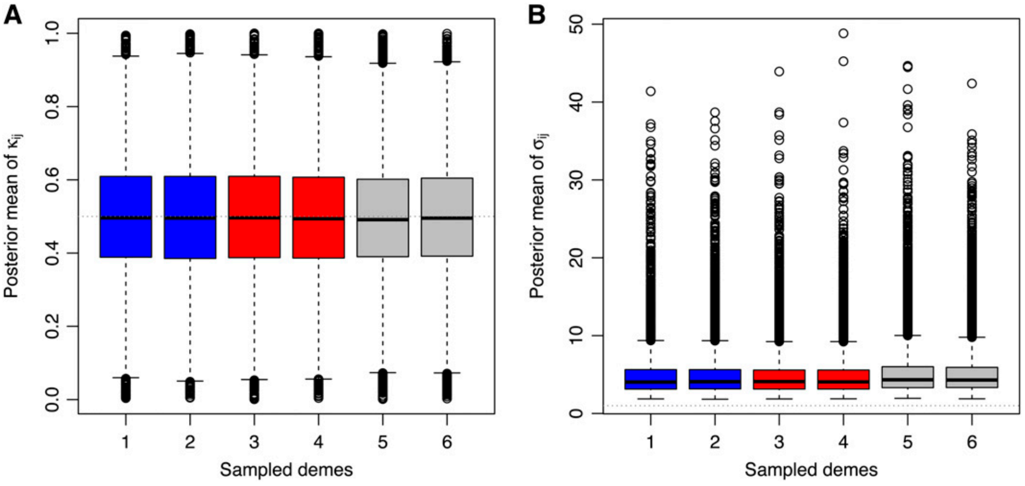

Figure 7 Analysis of the allele count data from data set 5. (A) Boxplot representation of the posterior means of the parameters κ_{ij} (which indicate which allele is selected for) for the 8000 neutral markers in blue demes (1–2), red demes (3–4), and uncolored demes (5–6). (B) Boxplot representation of the posterior means of the selection coefficients σ_{ij} for neutral markers in data set 5. The posterior means of the selection coefficients σ_{ij} are unconditional.

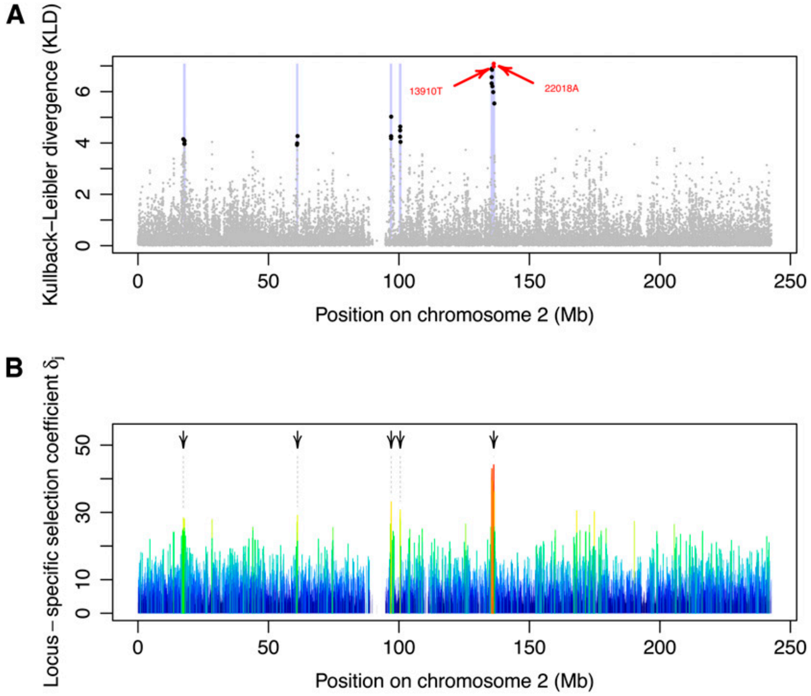

Figure 8 Analysis of the HGDP–CEPH data. (A) KLD between the posterior of δ_j and its centering distribution. The genomic regions highlighted in light purple show the strongest signatures of selection in the HGDP–CEPH data. They correspond to sets of 1-Mb windows (centered around each marker) containing at least three SNPs above the critical KLD value of 3.924 (corresponding to the 99.9% quantile of the KLD distribution from the POD analysis). For each region, the SNPs with KLD $\geq$ 3.924 are shown in black. The alleles −13910T and −22018A associated with lactase persistence are shown in red. (B) Locus-specific selection coefficient δ_j along chromosome 2. The color (from blue to red) and the width of each segment is proportional to the strength of selection. The arrows highlight the five outstanding genomic regions highlighted in A. The closest gene from each of these regions is provided in Table 2.

Gelman–Rubin's diagnostic is based on the computation of the ratio of the pooled-chains variance over the within-chain variance and was calculated using the coda package, v. 0.16-1, (Plummer *et al.* 2006) as implemented for R (R Core Team 2013). Gelman–Rubin's diagnostic was equal to 1.09 for the hyperparameter λ and to 1.07 for the parameters M_i, which indicates that the chains converge satisfactorily to the target distribution. The following analyses are based on the outputs from one of the three Markov chains.

To identify the genomic regions showing the strongest signatures of selection in the HGDP–CEPH data, 1-Mb windows were constructed for each marker by including all markers that were ≤500 kb from that marker. The average number of markers per window was ~248. Outstanding regions were then defined as the windows containing at least three SNPs above the critical KLD value of 3.924 (corresponding to the 99.9% quantile of the KLD distribution from the POD analysis). Figure 8A shows the distribution of the KLD for each SNP along HSA2. The two SNPs that are tightly associated with lactase persistence (−13910C → T and −22018G → A) are highlighted. These two SNPs have the two largest KLD values. Furthermore, the nine SNPs with the largest KLD values were located 3.7 kb and 1.0 Mb upstream of the *LCT* gene, at <805.2 kb from −13910C → T and <813.4 kb from −22018G → A. Figure 8B represents the distribution of the posterior means of the locus-specific selection parameter δ_j, along HSA2. This figure therefore represents the variation of the strength of selection along the chromosome and depicts a very strong signal of positive selection in the vicinity of the *LCT* gene (located from base pair 136,545,414 to 136,594,749), which encodes for the enzyme lactase–phlorizin hydrolase and is associated with adult-type hypolactasia. In addition to the *LCT* region, we found four other outstanding genomic regions, which are indicated by arrows in Figure 8B. The closest gene from each region was determined from the UCSC Genome Browser (http://genome.ucsc.edu/), using the Genome Reference Consortium GRCh37 assembly (hg19). Table 2 provides the list of these genes, along with their functions.

Figure 9 shows the distribution of the scaled coefficients of selection σ_{ij} (conditionally on κ_{ij} indicating allele −13910C → T to be targeted by selection) across African and Eurasian populations. The map from Figure 9 was extrapolated by kriging using the R package fields (Fields Development Core Team 2006), v. 6.8. It is obvious from Figure 9 that the intensity of selection is very strong in Europe and around the Indus valley and attains similar levels in both geographic regions.

Discussion

We developed a hierarchical Bayesian model that considers explicitly the effect of genic selection on the distribution of allele frequencies at SNP loci. SelEstim extends previous methods based on the Dirichlet-multinomial distribution of allele frequencies (which reduces to the beta-binomial distribution for SNP data) that arises as the diffusion approximation of genetic drift in the migration-drift equilibrium island model (see, *e.g.*, Beaumont and Balding 2004; Riebler *et al.* 2008; Foll and Gaggiotti 2008; Guo *et al.* 2009; Gautier *et al.* 2010). The beta-binomial model has been argued to be robust to the vagaries of demographic history (Beaumont and Nichols 1996; Beaumont 2005) because in

Table 2 List of the four genomic regions from chromosome 2 (besides the *LCT* region) showing the strongest signatures of selection

Region	SNP ID	Position	KLD	Closest gene (position)	Function
1	rs7355461	17,538,316	4.148	*RAD51AP2* (chr2:17,691,986–17,699,706)	Unknown function[a]
2	rs1177279	61,295,122	4.269	*KIAA1841* (chr2:61,293,006–61,316,639)	Unknown function
3	rs1256991	97,669,386	5.023	*FAM178B* (chr2:97,541,619–97,652,301)	Unknown function
4	rs1519662	101,143,618	4.636	*NMS* (chr2:101,086,944–101,099,742)	Neuropeptide signaling pathway

To identify these regions (indicated by arrows in Figure 8B), 1-Mb windows were constructed for each marker by including all markers that were ≤500 kb from that marker. Outstanding regions were then defined as the windows containing at least three SNPs above the critical KLD value of 3.924 (corresponding to the 99.9% quantile of the KLD distribution from the POD analysis). For each region, the SNP with the highest KLD value is indicated, along with its position. The closest gene from each SNP, determined from the University of California—Santa Cruz Genome Browser (http://genome.ucsc.edu/) using the Genome Reference Consortium GRCh37 assembly (hg19), is provided along with its function.

[a] Close to *VSNL1* (chr2:17,720,393–17,838,285) in Pickrell *et al.* (2009) (top 10 XP–EHH signal).

many situations, the genealogy of genes in a metapopulation divides into a scattering phase, which represents the recent genealogy of each deme, and a collecting phase, which represents the ancestral genealogy of the whole metapopulation (Nordborg 1997; Wakeley 2004). With this separation of timescales, there are no mutations in the scattering phase and the distribution of gene frequencies in each deme depends upon the frequencies in the pool of migrants (*i.e.*, the collecting phase) and the deme-specific F_{ST}. However, the assumption that each deme receives migrants from a unique migrant pool may not hold if populations share a history of successive divergences (Gaggiotti and Foll 2010). In that case indeed, gene frequencies may be correlated among closely related populations, which violates the assumption that populations are independent (Robertson 1975; Excoffier *et al.* 2009; Bonhomme *et al.* 2010; Gompert and Buerkle 2011). SelEstim should therefore be used with caution on populations that are known to be hierarchically structured (but see Figure S12). To conclude on a more positive note, we would argue that although violations from the island model assumptions certainly inflate the overall variance of the M_i parameters, it should not generate artificially correlated signals across closely linked SNPs as observed, *e.g.*, in Figure 8. From a practical point of view, using sliding windows to identify genomic regions of interest may therefore constitute a valuable approach (see, *e.g.*, Gautier *et al.* 2009). Accounting explicitly for the correlation of gene frequencies across populations due to shared history and gene flow might be achieved by considering the multivariate generalization of the Gaussian approximation of the gene frequency distribution (Coop *et al.* 2010; Günther and Coop 2013), which was recently extended to infer population splits and mixtures (Pickrell and Pritchard 2012). A Gaussian approximation for the distribution of allele frequencies (as suggested by Nicholson *et al.* 2002) can be justified whenever the deterministic equilibrium is located away from the boundaries (fixation of any one of the alleles). However, it is a poor approximation when the deterministic equilibrium is close to one of the boundaries (Barton and Rouhani 1987; Gautier and Vitalis 2013).

Our model introduces two major improvements over the methods based on the Dirichlet-multinomial or the beta-binomial distribution of gene frequencies. First, instead of being conceived as tests of departure from a neutral model (see, *e.g.*, Gautier *et al.* 2010), SelEstim incorporates an explicit selection model, which allows selection strength to be inferred among a set of markers. Second, SelEstim provides the distribution of selection strength across populations, which allows identifying the local population(s) where selection is acting. This is so because the hierarchical structure of our model improves the estimation of locus- and population-specific coefficients of selection (σ_{ij}) by borrowing strength across multiple populations.

Detecting selection among a set of markers

At a given SNP, j, the locus- and population-specific parameters of selection σ_{ij} depend upon a locus-specific hyperparameter δ_j that gives the population-wide effect of selection at a particular locus. It is therefore natural to use the posterior distribution of the hyperparameters δ_j as a means to discriminate between neutral and selected markers. We indeed expect the posterior density of δ_j to be shifted toward zero if the jth marker is neutral and toward positive values if the jth marker is targeted by selection. To operate this classification, it would have been possible to follow Beaumont and Balding (2004) and adopt a simple informal criterion assuming that δ_j is significantly different from zero at some critical level P whenever its equal-tailed $100(1 - P)$% credible interval excludes zero. Yet, this approach would neglect the genome-wide effect of selection, which in our model is driven by the hyperparameter λ. We therefore proposed comparing the posterior distributions of the locus-specific coefficients of selection δ_j with the centering distribution derived from the hyperdistribution with parameter λ. To that end, we used the KLD to measure the divergence between these two distributions. We calibrated this measure using simulations from a predictive distribution based on the observed data set. We found that the false-positive rate at any KLD threshold is always less than the corresponding quantile probability (Figure 3, Figure 4, Figure S1, Figure S2, Figure S3, Figure S4, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, and Figure S10), which suggests that such calibration is conservative, even for strong departures from the island model assumptions (Figure S12F).

McCulloch (1989), Peng and Dey (1995), and Guo *et al.* (2009) proposed an alternative calibration, based on the

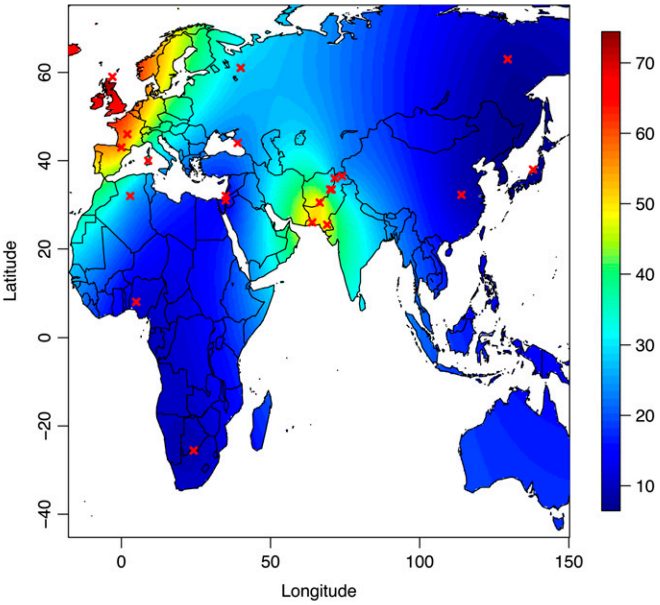

Figure 9 Extrapolated spatial distribution of the selection coefficient σ_{ij} at locus −13910C → T, conditionally on allele −13910T being selected for, across African and Eurasian populations.

following argument: consider flipping a "fair" coin with equal probability 0.5 for heads and tails *vs.* flipping a biased coin with probability ν for heads. Then the KLD between these two Bernoulli distributions denoted by Ber(0.5) and Ber(ν), respectively, can be computed and used as a threshold for any value of ν. In our case, however, we must divide ν in half, since we look for only unusually large KLD values. Therefore, the KLD threshold can be defined as $\text{KLD}[\text{Ber}(0.5), \text{Ber}(\nu)] = -\log[\nu(2-\nu)]/2$. For example, the KLD between two Bernoulli distributions corresponding to flipping a fair coin and a biased coin that gives heads or tails with probability 0.05 (resp. 0.01) is equal to 1.164 (resp. 1.959). However, we found that this calibration procedure was overly conservative (see Table S1 and Table S2).

Last, we used ROC analyses, as in Riebler *et al.* (2008), to compare our model with BayeScan (Foll and Gaggiotti 2008). We found that SelEstim performed slightly better than BayeScan (Figure 5). Since BayeScan was shown to outperform Beaumont and Balding's approach (Beaumont and Balding 2004), as well as some other popular moment-based methods using dominant markers (Pérez-Figueroa *et al.* 2010), we may therefore conclude that SelEstim represents an appreciable improvement to the population genomicist's toolbox.

As an alternative to the KLD, it could have been possible to implement Bayesian model selection using, *e.g.*, Bayes factors to discriminate between neutral and selected loci. Foll and Gaggiotti (2008) proposed using reversible-jump MCMC to estimate, for each marker, the posterior probabilities of two alternative models: a purely neutral one and one including selection. Riebler *et al.* (2008) also proposed an elegant reparameterization of Beaumont and Balding's model (Beaumont and Balding 2004), by introducing a Bernoulli-distributed auxiliary variable to indicate whether a locus is targeted by selection. This parameterization was later shown to facilitate the computation of Bayes factors (Gautier *et al.* 2009). Both approaches (reversible-jump MCMC and auxiliary variable) are actually straightforward to implement in our hierarchical-Bayesian model (not shown). Nevertheless, the KLD has the practical advantage that, provided the MCMC sampler converges and a large enough sample is drawn from the δ_j and the λ posterior distributions, its computation does not depend upon the length of the MCMC. As an illustration example, we ran a BayeScan analysis of the 52,633 SNPs from the HGDP–CEPH data, using the same MCMC parameters (number and length of pilot runs, burn-in, chain length, etc.) as with SelEstim, assuming prior odds of 1000 for the neutral model (Figure S22A). It is clear from this figure that a substantial number of markers for which the Kullback–Leibler divergence provides no evidence of selection have $\log_{10}(\text{BF}) \geq 2$ (Figure S22, B and C). This number increases with decreasing prior odds (not shown). Furthermore, since the maximum value that the BF can take is bounded by the MCMC length, we may observe a "saturation" effect with many of the outliers sharing the same evidence of selection (see Figure S22A). From a practical point of view, this may prevent the visual identification of genomic regions potentially targeted by selection (as, *e.g.*, in Figure 8), unless very long MCMC are performed (to achieve an effective sample size of the order of the number of markers). In addition, we found that the outputs of BayeScan vary with the prior odds, which depend on the user's prior belief for the proportion of presumably neutral SNPs. Our results therefore argue in favor of using KLD in empirical studies since it allows ranking the SNPs in order of the divergence between locus-specific and genome-wide selection strength, which indicates the degree of evidence that a locus is under selection. However, we concur with Coop *et al.* (2010) that making statements about the statistical significance of outlier loci might be hazardous. In particular, we refrained from defining *P*-values from KLD measures.

Inferring selection strength across populations

In an early analysis of population differentiation using the HapMap data set, Weir *et al.* (2005) already showed the utility of estimating population-specific F_{ST} values (Weir and Hill 2002). In particular, concentrating their analyses on chromosome 2, they did not find any outstanding peak of population average F_{ST} around the *LCT* gene, although there was a clear elevation of the population-specific F_{ST} values for Caucasians of European descent and European Americans. Yet, in Weir *et al.*'s study (Weir *et al.* 2005), the characterization of "exceptional regions" was based on the greatest difference between population-specific F_{ST} values (averaged over 5-Mb windows) being larger than 3 SD, which does not provide a definitive statistical criterion to decide which loci are outliers of the empirical, genome-wide distribution of F_{ST}.

Like Beaumont and Balding (2004), Riebler *et al.* (2008), Foll and Gaggiotti (2008), and Guo *et al.* (2009), who considered population-specific effects on F_{ST}, we considered in our model that the distribution of allele frequency depends upon population-specific parameters (M_i). Since we defined a parameter that indicates which allele is selected for, the selected allele need not to be the same in all the sampled demes. Furthermore, the strength of selection need not to be the same in all demes. SelEstim therefore accounts for situations where selection is acting in some populations, but not all, possibly in opposite direction (with alternative alleles being selected for in different environments). It is therefore particularly relevant to detect the signatures of local adaptation in subdivided populations.

Not surprisingly, we found that SelEstim has weak statistical power to identify loci under balancing selection (see Figure 5). Since our genic selection model allows for only positive selection, this was somewhat expected, but by using a centering distribution we are able, in principle, to identify loci with support for unusually low values of δ. Beaumont and Balding (2004) concluded from simulations that their method could not easily identify loci under balancing selection, even for very strong selection. Although Foll and Gaggiotti (2008) showed that microsatellites could be used to detect balancing selection, especially with data sets containing a large number of sampled populations, they needed 10 populations with SNPs to achieve the same rate of detection (Foll and Gaggiotti 2008). In principle, however, it should be possible to scan for SNPs targeted by balancing selection using a modified version of our model, in particular Equation 2, that would account for overdominance with population-specific selection pressures (see, *e.g.*, Equation 13.60 in Wright 1969, p. 371). This strategy could be valuable for improving statistical power to identify loci under balancing selection.

Because our model accounts explicitly for positive selection, it cannot only be used to detect the genomic signatures of selection, but also to measure the strength of selection along the genome. As mentioned above, contrary to previous approaches that approximated selection as a locus-specific effect in a logistic regression model (Beaumont and Balding 2004) or a reduction in effective migration rate (see, *e.g.*, Bazin *et al.* 2010), we introduced explicitly a scaled coefficient of selection $\sigma_{ij} \equiv 2N_i s_{ij}$ for locus j in deme i, where s_{ij} represents the relative gain in fitness brought by a positively selected allele. We found that the posterior means of the scaled coefficients of selection σ_{ij} (conditionally on κ_{ij}) were close to the simulated value for positively selected loci, although slightly overestimated (Figure 6, Figure S14, Figure S15, Figure S16, Figure S17, and Figure S18). We also found that the variation of σ_{ij} across populations with different selection regimes was remarkably well inferred, with selected loci exhibiting large coefficients of selection in the colored demes and small coefficients of selection in uncolored demes (Figure 6, Figure S14, Figure S15, Figure S16, Figure S17, and Figure S18).

The strong correlation between the posterior mean of the genome-wide coefficient of selection λ and the number of positively selected loci (Figure S20) would tend to suggest that the parameter λ provides some information about the extent of selection acting on the genome. This must be nuanced, however, at least for two reasons. First, as the number of positively selected loci increases, the performance of SelEstim weakens (see Figure S21). Second, we have observed that the parameter λ also depends on demography, and particularly on departures from the island model assumptions (not shown). This identifiability problem therefore prevents the comparison of λ estimates (to infer the overall effect of selection) across species with different population structures. We note that this identifiability problem is somewhat avoided in BayeScan with the Gaussian prior (zero mean and standard deviation of 1) put on the locus-specific component α_i (which, therefore, provides no information whatsoever on the extent of selection acting on the genome).

Application example at the LCT gene

To illustrate how the inference of selection strength may provide new insights into the characterization of local adaptation, we investigated the well-studied and clear-cut example of the evolution of lactase persistence in humans (see Gerbault *et al.* 2011, for a review). The region around the *LCT* gene that allows lactose tolerance to persist into adulthood is indeed a very-well-known example of selection in humans (Sabeti *et al.* 2006). The first causative polymorphism described was the −13910C → T mutation (Enattah *et al.* 2002), which lies in the *cis*-acting regulatory element located in the 13th intron of a neighboring gene, *MCM6*. Although this single mutation of purported western Eurasian origin accounts for much of observed lactase persistence outside Africa, multiple independent mutations in the same region upstream of the *LCT* gene have been associated with this trait in pastoralists from Saudi Arabia (Enattah *et al.* 2008) and Africa (Tishkoff *et al.* 2007). The lactase persistence allele at the *LCT* locus lies on a haplotype that is common in Europeans but that extends largely undisrupted for >1 Mb, much farther than is typical for an allele of that frequency (Bersaglieri *et al.* 2004). More recently, Romero *et al.* (2012) found that the −13910C → T mutation also explains a substantial proportion of lactase persistence in the Indian subcontinent. Most interestingly, they showed that the −13910C → T mutation in India is identical by descent to the European allele and is associated with the same extended haplotype in both populations, which strongly suggests that the origin of the −13910C → T mutation is shared in Europe and India. These results are consistent with the high levels of present-day milk consumption in India and with archeological and genetic evidence for the independent domestication of cattle in the Indus valley ca. 7000 years ago (Romero *et al.* 2012).

In agreement with these studies, our analyses pointed to a very strong signal of positive selection between 3.7 kb and 1.0 Mb upstream of the *LCT* gene (Figure 8). The strongest evidence of selection (in terms of KLD) was found for the two SNPs that are tightly associated with lactase persistence (−13910C → T and −22018G → A). Building on the fact that our model is able to give accurate measures of the scaled coefficient of selection at each locus in different demes, we further examined the distribution of the strength of positive selection at the −13910C → T SNP across the 23 populations analyzed. We found the strongest selection coefficients in Europe and in the Indus Valley (Figure 9), which matches the interpolated map of lactase persistence phenotype frequencies in the Old World (Itan *et al.* 2010). More precisely, we found that the coefficients of selection ($\sigma \equiv 2Ns$) at the −13910T allele ranged from 8.73 (Sardinians) to 101.66 (Orcadians) in Europe and from 4.94 (Kalash) to 77.50 (Balochi) in Central/South Asia. There have been previous attempts to measure the strength of selection acting at the *LCT* gene, although most of them relied on strong assumptions on the demographic and adaptive history of the studied populations. For example, Aoki (1986) predicted that a selection coefficient $s > 5\%$ would be necessary to explain the observed allele frequency of the −13910T allele, assuming that this mutation appeared 6000 years ago in a population of effective size 500, which would give $\sigma \equiv 2Ns = 50$. Bersaglieri *et al.* (2004) estimated the coefficient of selection s to be 1–15% for a new mutation arising in a population of effective size of 500–5000. More recently, Tishkoff *et al.* (2007) estimated selection intensity by matching simulated data under a coalescent framework to the observed centimorgan span and the observed frequency of the allele targeted by selection. They found extremely recent and strong positive selection in many African populations ($\sigma \equiv 2Ns$ ranging from 800 to 1940 assuming an effective population size N of 10,000). Modeling a geographical structuring of selection pressure by latitude, Gerbault *et al.* (2009) found selection coefficients in the range between 0.8 and 1.8% (Gerbault *et al.* 2011), also assuming a carrying capacity of 10,000. However, assuming an effective population size N of 10,000 may largely overestimate $\sigma \equiv 2Ns$ (see Tenesa *et al.* 2007, for more accurate estimates of effective size based on measures of linkage disequilibrium). Last, using a spatially explicit model and approximate Bayesian computation (Beaumont *et al.* 2002), Itan *et al.* (2009) estimated coefficients of selection to lie in the range of 5.2–15.9%. The difficulty in comparing these values is that strong hypotheses about the effective population size need to be made. It is clear from the stationary density of the diffusion process in Equation 2 that the two parameters s and N are not identifiable. Estimating s therefore requires informative priors on N. Furthermore, the population size considered in our model is the local effective size of a deme, not the effective size of the total population. Therefore, considering the scaled coefficient of selection ($\sigma \equiv 2Ns$) might be more appropriate for interpreting the variation of the strength of selection exerted at different loci or at one locus in different populations.

Perspectives

SelEstim provides a new tool with which to detect signatures of selection from genome-wide scan studies and, perhaps most importantly, to infer the intensity of selection across loci and populations. However, as most F_{ST}-based methods aimed at looking for locus-specific effects on F_{ST} estimates, SelEstim assumes that molecular markers are independent from each other. There are few exceptions, though. For example Guo *et al.* (2009) introduced a conditional autoregressive model to incorporate the local correlation among SNPs. Gompert and Buerkle (2011) proposed an extension of the models developed by Beaumont and Balding (2004), Riebler *et al.* (2008), and Foll and Gaggiotti (2008), which incorporates genetic distances among haplotypes (ϕ-statistics; see Excoffier *et al.* 1992) in measures of genetic differentiation. More recently, Fariello *et al.* (2013) developed a haplotype-based method, which uses a multipoint linkage disequilibrium model (Scheet and Stephens 2006) that regroups individual chromosomes into local haplotype clusters. The reconstructed haplotypes are then used to measure differentiation between populations (see also Browning and Weir 2010). Handling conditional dependencies of markers along the genome would therefore be an essential step forward in future developments of SelEstim. In the meantime, we recommend potential users to view this method as a first step toward identify genomic regions of interest, which should then be characterized more specifically in further studies.

Acknowledgments

We acknowledge two anonymous reviewers for helpful comments, which led to a significantly improved version of this manuscript. We also warmly thank Pierre Pudlo and Christian P. Robert for insightful discussions on KLD calibration. This work was supported by a Biotechnology and Biological Sciences Research Council grant (BBS/B/12776) awarded to M.A.B. and K.J.D. M.A.B. was supported by a Natural Research Council Advanced Fellowship (NER/J/S/2001/00792). R.V. was supported by the French National Research Agency (ANR) programme NUTGENEVOL 07-BLAN-0064. Both R.V. and M.G. were supported by the ANR programme EMILE 09-BLAN-0145-01 and by the Institut National de la Recherche Agronomique, through the support of the starting group "IG-GiPop" (2010–2012).

Literature Cited

Abramowitz, M., and I. A. Stegun, 1965 *Handbook of Mathematical Functions*. Dover, New York.

Akey, J. M., G. Zhang, L. Jin, and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12: 1805–1814.

Aoki, K. A., 1986 Stochastic model of gene-culture coevolution suggested by the 'culture historical hypothesis' for the evolution of adult lactose absorption in humans. Proc. Natl. Acad. Sci. USA 83: 2929–2933.

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

Balding, D. J., and R. A. Nichols, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96: 3–12.

Barreiro, L. B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci, 2008 Natural selection has driven population differentiation in modern humans. Nat. Genet. 40: 340–345.

Barton, N., and S. Rouhani, 1987 The frequency of shifts between alternative equilibria. J. Theor. Biol. 125: 397–418.

Barton, N., and M. Turelli, 1987 Adaptive landscapes, genetic distance and the evolution of quantitative characters. Genet. Res. 49: 157–173.

Bazin, E., M. A. Beaumont, and K. J. Dawson, 2010 Likelihood-free inference of population structure and local adaptation in a bayesian hierarchical model. Genetics 185: 587–602.

Beaumont, M. A., 2005 Adaptation and speciation: What can F_{ST} tell us? Trends Ecol. Evol. 20: 435–440.

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. 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.

Bersaglieri, T., P. Sabeti, N. Patterson, T. Vanderploeg, S. Schaffner *et al.*, 2004 Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74: 1111–1120.

Black, W. C., C. F. Baer, M. F. Antolin, and N. M. DuTeau, 2001 Population genomics: genome-wide sampling of insect populations. Annu. Rev. Entomol. 46: 441–469.

Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah *et al.*, 2010 Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics 186: 241–262.

Brooks, S., and A. Gelman, 1998 General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Statist. 7: 434–455.

Browning, S., and B. Weir, 2010 Population structure with localized haplotype clusters. Genetics 185: 1337–1344.

Bürger, R., 2000 *The Mathematical Theory of Selection, Recombination Mutation*. Wiley, Chichester, England.

Bustamante, C. D., J. Wakeley, S. A. Sawyer, and D. L. Hartl, 2001 Directional selection and the site-frequency spectrum. Genetics 159: 1779–1788.

Bustamante, C. D., R. Nielsen, and D. L. Hartl, 2003 Maximum likelihood and bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. Theor. Popul. Biol. 63: 91–103.

Cann, H. M., C. de Toma, L. Cazes, M.-F. Legrand, V. Morel *et al.*, 2003 A human genome diversity cell line panel. Science 296: 261–262.

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics 185: 1411–1423.

Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29: 401–421.

Donnelly, P., M. Nordborg, and P. Joyce, 2001 Likelihoods and simulation methods for a class of nonneutral population genetics models. Genetics 159: 853–867.

Enattah, N. S., T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen *et al.*, 2002 Identification of a variant associated with adult-type hypolactasia. Nat. Genet. 30: 233–237.

Enattah, N. S., T. G. Jensen, M. Nielsen, R. Lewinski, M. Kuokkanen *et al.*, 2008 Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. Am. J. Hum. Genet. 82: 57–72.

Ethier, S. N., and T. Nagylaki, 1988 Diffusion approximations of markov chains with two time scales and application to population genetics, II. Adv. Appl. Probab. 20: 525–545.

Excoffier, L., P. E. Smouse, and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among dna haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479–491.

Excoffier, L., T. Hofer, and M. Foll, 2009 Detecting loci under selection in a hierarchically structured population. Heredity 103: 285–298.

Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal, and B. Servin, 2013 Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics 193: 929–941.

Fawcett, T., 2006 An introduction to ROC analysis. Pattern Recognit. Lett. 27: 882–891.

Fields Development Core Team, 2006 *fields: Tools for Spatial Data*. National Center for Atmospheric Research, Boulder, Colorado.

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.

Frichot, E., S. D. Schoville, G. Bouchard, and O. François, 2013 Testing for associations between loci and environmental gradients using latent factor mixed models. Mol. Biol. Evol. 30: 1687–1699.

Gaggiotti, O., and M. Foll, 2010 Quantifying population structure using the *F*-model. Mol. Ecol. Res. 10: 821–830.

Gautier, M., and R. Vitalis, 2012 rehh: An r package to detect footprints of selection in genome-wide snp data from haplotype structure. Bioinformatics 28: 1176–1177.

Gautier, M., and R. Vitalis, 2013 Inferring population histories using genome-wide allele frequency data. Mol. Biol. Evol. 39: 654–668.

Gautier, M., L. Flori, A. Riebler, F. Jaffrézic, D. Laloé *et al.*, 2009 A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. BMC Genomics 10: 550.

Gautier, M., T. D. Hocking, and J.-L. Foulley, 2010 A Bayesian outlier criterion to detect SNPs under selection in large data sets. PLoS ONE 5: e11913.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2004 *Bayesian Data Analysis*, Ed. 2. Chapman & Hall, New York.

Gerbault, P., C. Moret, M. Currat, and A. Sanchez-Mazas, 2009 Impact of selection and demography on the diffusion of lactase persistence. PLoS ONE 4: e6369.

Gerbault, P., A. Liebert, Y. Itan, A. Powell, M. Currat *et al.*, 2011 Evolution of lactase persistence: an example of human niche construction. Philos. Trans. R. Soc. Lond. B Biol. Sci. 366: 863–877.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, 1996 *Markov Chain Monte Carlo in Practice*, Ed. 2. Chapman & Hall, New York.

Goldstein, D. B., and L. Chikhi, 2002 Human migrations and population structure: what we know and why it matters. Annu. Rev. Genomics Hum. Genet. 3: 129–152.

Gompert, Z., and C. A. Buerkle, 2011 A hierarchical bayesian model for next-generation population genomics. Genetics 187: 903–917.

Guillot, G., R. Vitalis, A. le Rouzic, and M. Gautier, 2014 Detecting correlation between allele frequencies and environmental variables as a signature of selection: a fast computational approach for genome-wide studies. Spatial. Stat. (in press).

Günther, T., and G. Coop, 2013 Robust identification of local adaptation from allele frequencies. Genetics 195: 205–220.

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.

Hancock, A. M., D. B. Witonsky, E. Ehler, G. Alkorta-Aranburu, C. Beall *et al.*, 2010 Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proc. Natl. Acad. Sci. USA 107: 8924–8930.

Hancock, A. M., D. B. Witonsky, G. Alkorta-Aranburu, C. M. Beall, A. Gebremedhin *et al.*, 2011 Adaptations to climate-mediated selective pressures in humans. PLoS Genet. 7: e1001375.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin *et al.*, 2005 Whole-genome patterns of common DNA variation in three human populations. Science 307: 1072–1079.

Innan, H., and Y. Kim, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. USA 101: 10667–10672.

International HapMap Consortium, 2003 The international HapMap project. Nature 426: 789–796.

International HapMap Consortium, 2005 A haplotype map of the human genome. Nature 437: 1299–1320.

Itan, Y., A. Powell, M. A. Beaumont, J. Burger, and M. G. Thomas, 2009 The origins of lactase persistence in europe. PLOS Comput. Biol. 5: e1000491.

Itan, Y., B. L. Jonesand, C. J. E. Ingram, D. M. Swallow, and M. G. Thomas, 2010 A worldwide correlation of lactase persistence phenotype and genotypes. BMC Evol. Biol. 10: 36.

Jeffreys, H., 1961 *Theory of Probability*, Ed. 3. Oxford University Press, Oxford.

Kass, R. E., and A. E. Raftery, 1995 Bayes factors. J. Am. Stat. Assoc. 90: 773–795.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765–777.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. Genetics 74: 175–195.

Luikart, G., P. R. England, D. Tallmon, S. Jordan, and P. Taberlet, 2003 The power and promise of population genomics: from genotyping to genome typing. Nat. Rev. Genet. 4: 981–994.

McCulloch, R., 1989 Local model influence. J. Am. Stat. Assoc. 84: 473–478.

Nei, M., and T. Maryuyama, 1975 Lewontin–Krakauer test for neutral genes. Genetics 80: 395.

Nicholson, G., A. V. Smith, F. Jónsson, Ó. Gústafsson, K. Stefánsson *et al.*, 2002 Assessing population differentiation and isolation from single-nucleotide polymorphism data. J. R. Stat. Soc. Series B Stat. Methodol. 64: 695–715.

Nielsen, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity 86: 641–647.

Nielsen, R., 2005 Disclosure of variation. Nature 434: 288–289.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton *et al.*, 2005a A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3: e170.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005b Genomic scans for selective sweeps using SNP data. Genome Res. 15: 1566–1575.

Nielsen, R., M. J. Hubisz, I. Hellmann, D. Torgerson, A. M. Andrés *et al.*, 2009 Darwinian and demographic forces affecting human protein coding genes. Genome Res. 19: 838–849.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics 146: 1501–1514.

Ntzoufras, I., 2009 *Bayesian Modeling Using WinBugs*. Wiley, Hoboken, NJ.

Payseur, B. A., A. D. Cutter, and M. W. Nachman, 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol. Biol. Evol. 19: 1143–1153.

Peng, F., and D. K. Dey, 1995 Bayesian analysis of outlier problems using divergence measures. Can. J. Stat. 23: 199–213.

Pérez-Cruz, F., 2008 Kullback–Leibler divergence estimation of continuous distributions, pp. 1666–1670 in IEEE International Symposium on Information Theory (ISIT), Toronto, Ontario, Canada.

Pérez-Figueroa, A., M. J. García-Pereira, M. Saura, E. Rolán-Alvarez, and A. Caballero, 2010 Comparing three different methods to detect selective loci using dominant markers. J. Evol. Biol. 23: 2267–2276.

Petry, D., 1983 The effect on neutral gene flow of selection at a linked locus. Theor. Popul. Biol. 23: 300–313.

Pickrell, J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8: e1002967.

Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li *et al.*, 2009 Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 19: 826–837.

Plummer, M., N. Best, K. Cowles, and K. Vines, 2006 Coda: output analysis and diagnostics for MCMC. R News 6: 7–11.

Przeworsky, M., G. Coop, and J. Wall, 2005 The signature of positive selection on standing variation. Evolution 59: 2312–2323.

R Core Team, 2013 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riebler, A., L. Held, and W. Stephan, 2008 Bayesian variable selection for detecting adaptive genomic differences among populations. Genetics 178: 1817–1829.

Robertson, A., 1975 Remarks on the Lewontin–Krakauer test. Genetics 80: 396.

Romero, I. G., C. B. Mallick, A. Liebert, F. Crivellaro, G. Chaubey *et al.*, 2012 Herders of indian and european cattle share their predominant allele for lactase persistence. Mol. Biol. Evol. 29: 249–260.

Ross, K. G., D. D. Shoemaker, M. J. B. Krieger, J. DeHeer, and L. Keller, 1999 Assessing genetic structure with multiple classes of molecular markers: a case study involving the introduced fire ant *Solenopsis invicta*. Mol. Biol. Evol. 16: 525–543.

Rousset, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics 142: 1357–1362.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.*, 2006 Positive natural selection in the human lineage. Science 312: 1614–1620.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.

Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78: 629–644.

Tang, K., K. R. Thornton, and M. Stoneking, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. 5: e171.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17: 520–526.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt *et al.*, 2007 Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. 39: 31–40.

Vitalis, R., P. Boursot, and K. Dawson, 2001 Interpretation of variation across marker loci as evidence of selection. Genetics 158: 1811–1823.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. 4: e72.

Wakeley, J., 1999 Nonequilibrium migration in human history. Genetics 153: 1863–1871.

Wakeley, J., 2004 Metapopulation models for historical inference. Mol. Ecol. 13: 865–875.

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.

Williamson, S. H., R. Hernandez, A. Fledel Alon, L. Zhu, R. Nielsen *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA 102: 7882–7887.

Wright, S., 1931 Evolution in mendelian populations. Genetics 16: 97–159.

Wright, S., 1949 Adaptation and selection, pp. 365–389 in *Genetics, Paleontology, and Evolution*, edited by G. L. Jepson, G. G. Simpson, and E. Mayr. University Press, Princeton, NJ.

Wright, S., 1969 *Evolution and the Genetics of Populations: Vol. II. The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

Communicating editor: W. Stephan

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.152991/-/DC1

Detecting and Measuring Selection from Gene Frequency Data

Renaud Vitalis, Mathieu Gautier, Kevin J. Dawson, and Mark A. Beaumont

Copyright © 2014 by the Genetics Society of America
DOI: 10.1534/genetics.113.152991

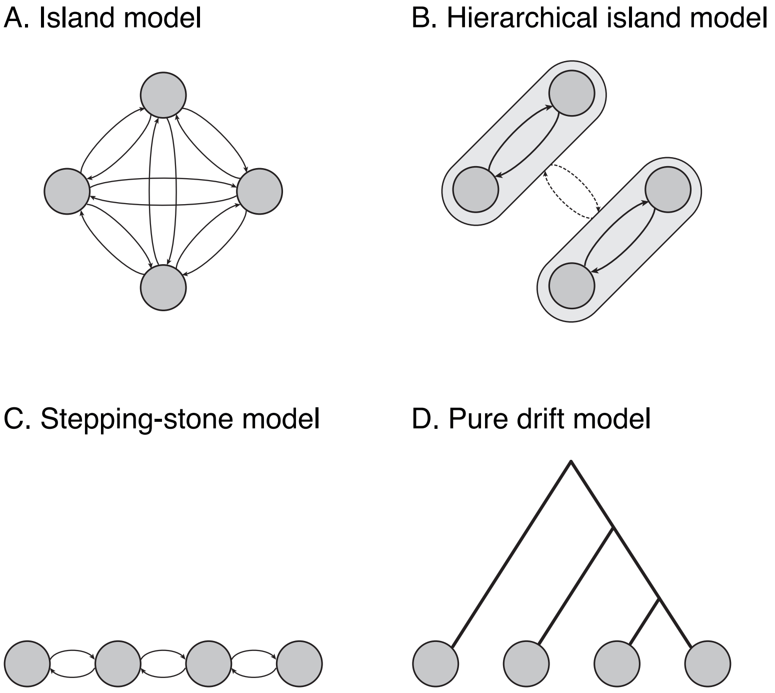

Figure S1 (A) Schematic representation of an island model. The actual data were simulated with n_d = 100 demes, each made of N = 250 diploid individuals (500 genes). Fifty diploid individuals (100 genes) were sampled per deme, in 9 demes. The migration rate (m = 0.003, plain arrows) was fixed to achieve the desired value of F_{ST} = 0.24, using equation 6 in Rousset (1996). (B) Schematic representation of a hierarchical island model. The actual data were simulated with 10 groups of 10 demes, each made of N = 250 diploid individuals (500 genes). Fifty diploid individuals (100 genes) were sampled per deme, in 3 groups of 3 demes. The migration rate within (m = 0.017, plain arrows) and among groups (m = 0.0003, dashed arrows) were fixed to achieve the desired values of F_{SC} = 0.05, F_{CT} = 0.05 and F_{ST} = 0.24, using equations A8–A10 in Excoffier *et al.* (2009). (C) Schematic representation of a stepping-stone model. The actual data were simulated with n_d = 100 demes, each made of N = 250 diploid individuals (500 genes). Fifty diploid individuals (100 genes) were sampled per deme, in 9 demes.The migration rate was fixed (m = 0.028, plain arrows), by trial and error, to achieve the desired value of F_{ST} = 0.24. (D) Schematic representation of a pure drift model. The actual data were simulated with 9 demes, diverging sequentially as depicted. The sample characteristics (number of individuals, number of sampled demes) were the same as in (A–C), and the divergence time (24 generations) between any two successive splits was tuned in order to achieve an overall F_{ST} of ≈ 0.24. In (A–D) 10,000 neutral markers were simulated.

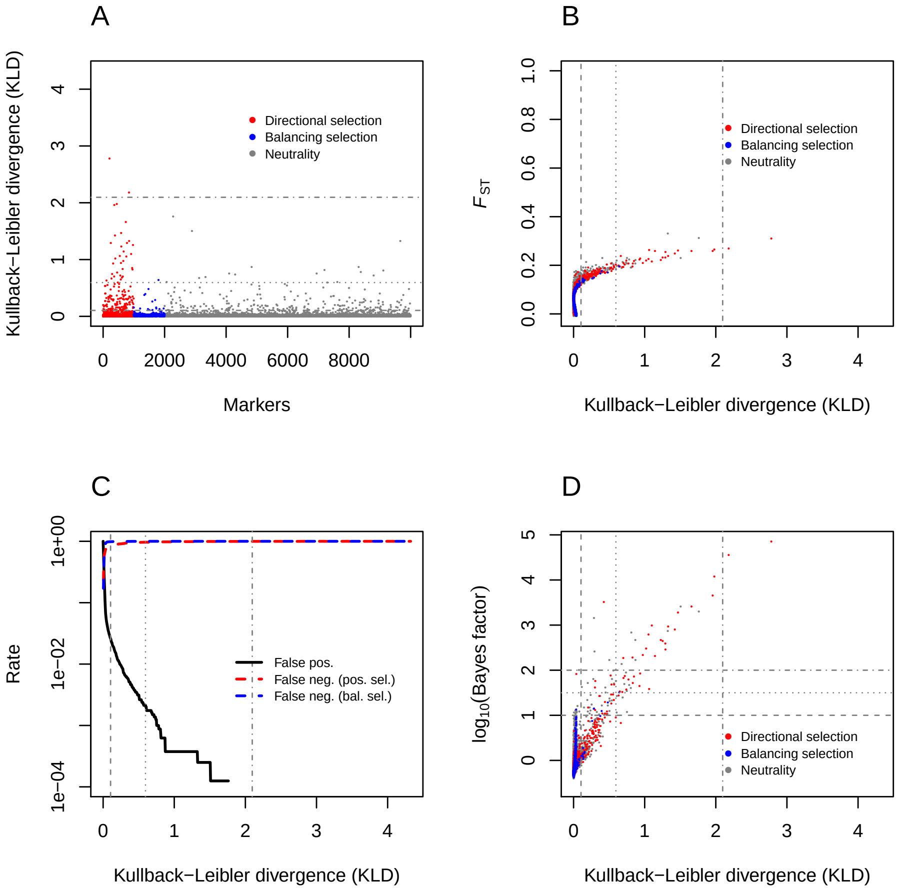

Figure S2 Analysis of the allele count data from dataset 1. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Relationship between the Bayes factor $\log_{10}$(BF) from the BAYESCAN analysis of dataset 1 and the KLD. The horizontal lines in (A) and the vertical lines in (B–D) indicate the KLD thresholds corresponding to the 95%-, the 99%- and the 99.9%-quantile of the of the KLD distribution from the pod analysis of dataset 1. In (D), the horizontal lines indicate the $\log_{10}$(BF) = 1, $\log_{10}$(BF) = 1.5 and $\log_{10}$(BF) = 2 thresholds, which correspond to ''strong'', ''very strong'' and ''decisive'' support, respectively, following Jeffreys' (1961) scale of evidence.

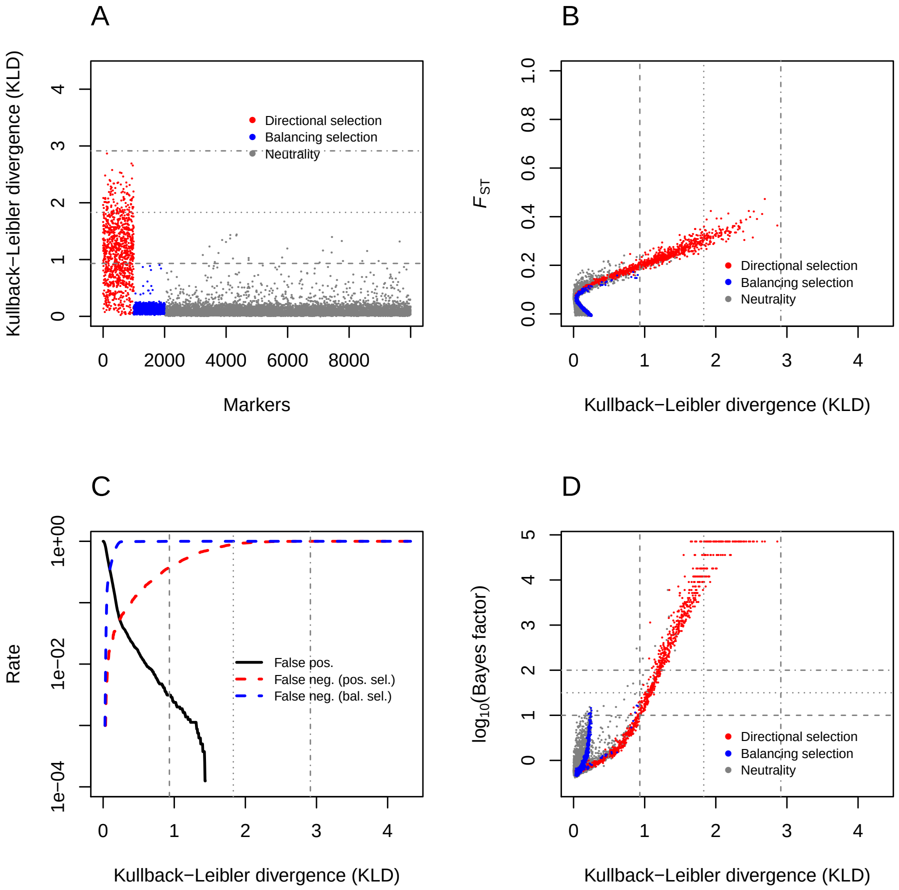

Figure S3 Analysis of the allele count data from dataset 2. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Relationship between the Bayes factor $\log_{10}$(BF) from the BAYESCAN analysis of dataset 2 and the KLD. The horizontal lines in (A) and the vertical lines in (B–D) indicate the KLD thresholds corresponding to the 95%-, the 99%- and the 99.9%-quantile of the of the KLD distribution from the pod analysis of dataset 2. In (D), the horizontal lines indicate the $\log_{10}$(BF) = 1, $\log_{10}$(BF) = 1.5 and $\log_{10}$(BF) = 2 thresholds, which correspond to ''strong'', ''very strong'' and ''decisive'' support, respectively, following Jeffreys' (1961) scale of evidence.

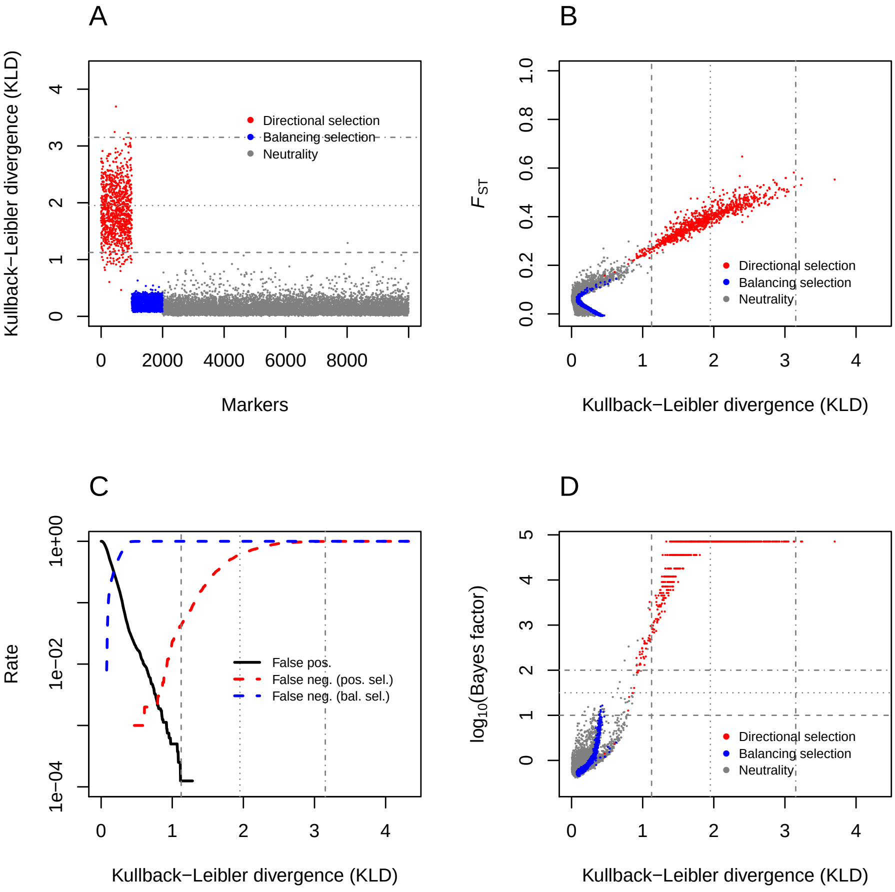

Figure S4 Analysis of the allele count data from dataset 3. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Relationship between the Bayes factor $\log_{10}$(BF) from the BayeScan analysis of dataset 3 and the KLD. The horizontal lines in (A) and the vertical lines in (B–D) indicate the KLD thresholds corresponding to the 95%-, the 99%- and the 99.9%-quantile of the of the KLD distribution from the pod analysis of dataset 3. In (D), the horizontal lines indicate the $\log_{10}$(BF) = 1, $\log_{10}$(BF) = 1.5 and $\log_{10}$(BF) = 2 thresholds, which correspond to ''strong'', ''very strong'' and ''decisive'' support, respectively, following Jeffreys' (1961) scale of evidence.

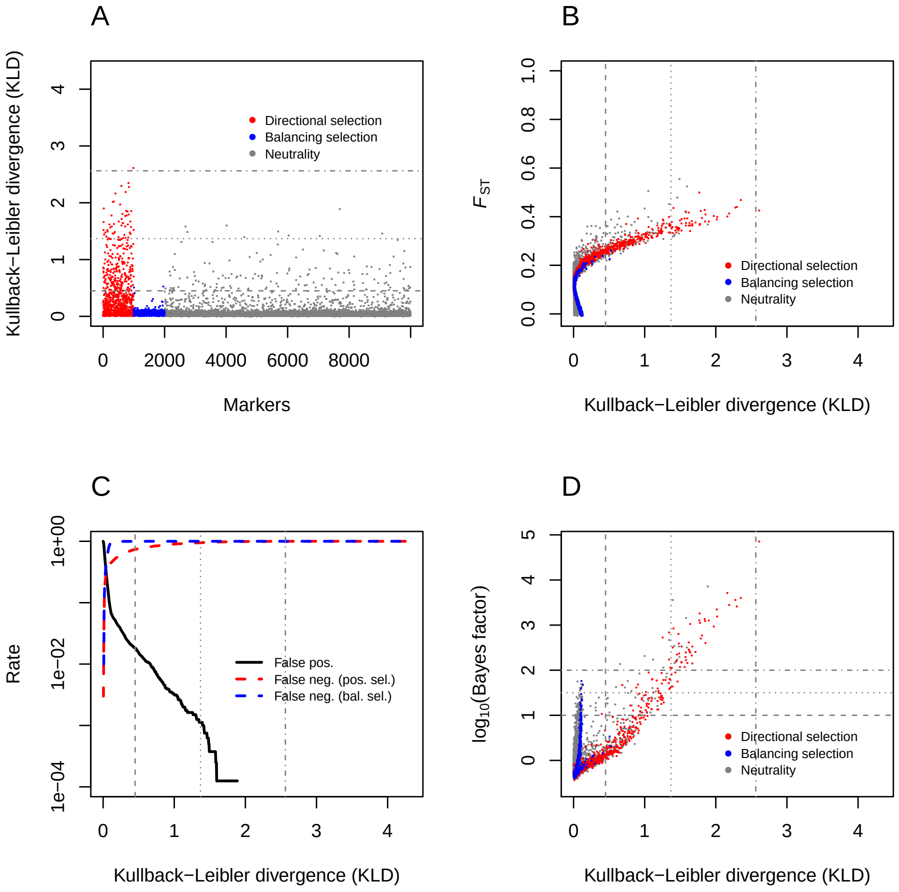

Figure S5 Analysis of the allele count data from dataset 4. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Relationship between the Bayes factor $\log_{10}(\mathrm{BF})$ from the BAYESCAN analysis of dataset 4 and the KLD. The horizontal lines in (A) and the vertical lines in (B–D) indicate the KLD thresholds corresponding to the 95%-, the 99%- and the 99.9%-quantile of the of the KLD distribution from the pod analysis of dataset 4. In (D), the horizontal lines indicate the $\log_{10}(\mathrm{BF}) = 1$, $\log_{10}(\mathrm{BF}) = 1.5$ and $\log_{10}(\mathrm{BF}) = 2$ thresholds, which correspond to ''strong'', ''very strong'' and ''decisive'' support, respectively, following Jeffreys' (1961) scale of evidence.

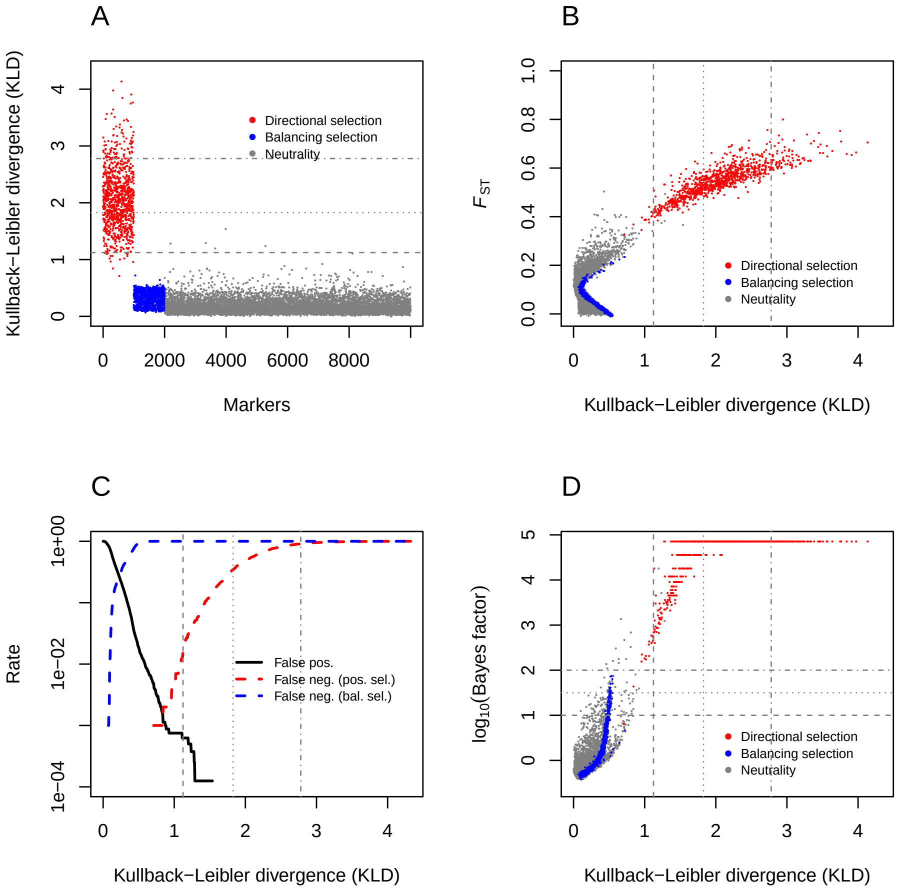

Figure S6 Analysis of the allele count data from dataset 6. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Relationship between the Bayes factor $\log_{10}$(BF) from the BAYESCAN analysis of dataset 6 and the KLD. The horizontal lines in (A) and the vertical lines in (B–D) indicate the KLD thresholds corresponding to the 95%-, the 99%- and the 99.9%-quantile of the of the KLD distribution from the pod analysis of dataset 6. In (D), the horizontal lines indicate the $\log_{10}$(BF) = 1, $\log_{10}$(BF) = 1.5 and $\log_{10}$(BF) = 2 thresholds, which correspond to "strong", "very strong" and "decisive" support, respectively, following Jeffreys' (1961) scale of evidence.

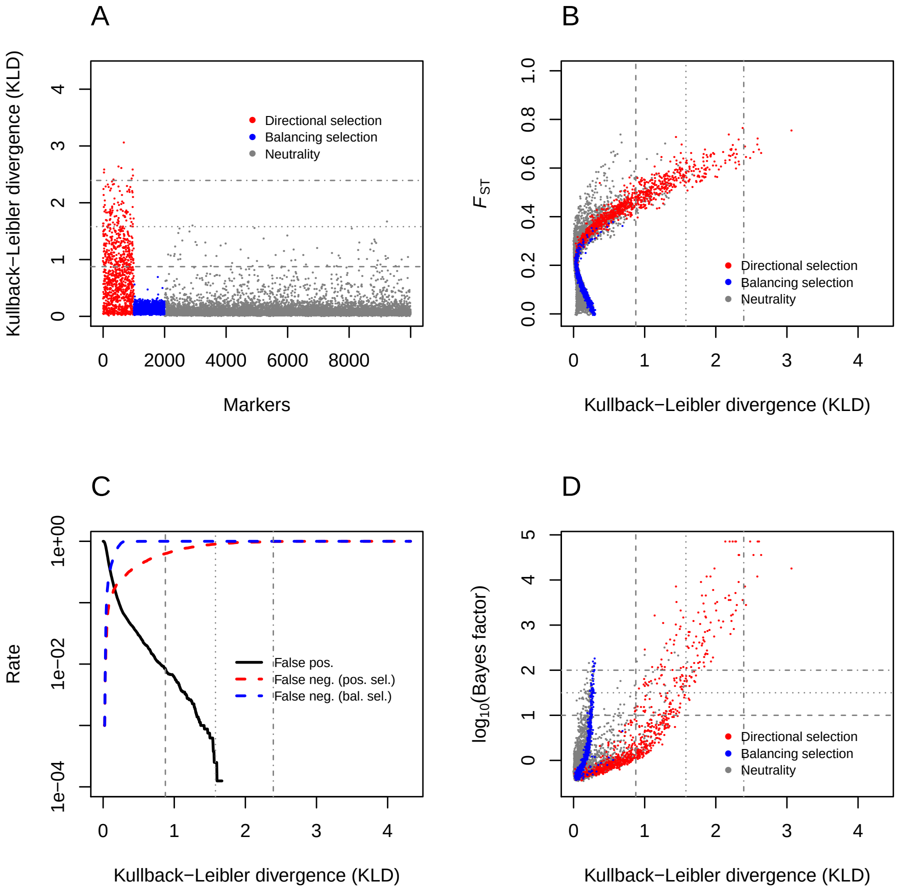

Figure S7 Analysis of the allele count data from dataset 7. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Relationship between the Bayes factor $\log_{10}$(BF) from the BayeScan analysis of dataset 7 and the KLD. The horizontal lines in (A) and the vertical lines in (B–D) indicate the KLD thresholds corresponding to the 95%-, the 99%- and the 99.9%-quantile of the of the KLD distribution from the pod analysis of dataset 7. In (D), the horizontal lines indicate the $\log_{10}$(BF) = 1, $\log_{10}$(BF) = 1.5 and $\log_{10}$(BF) = 2 thresholds, which correspond to ‘‘strong’’, ‘‘very strong’’ and ‘‘decisive’’ support, respectively, following Jeffreys’ (1961) scale of evidence.

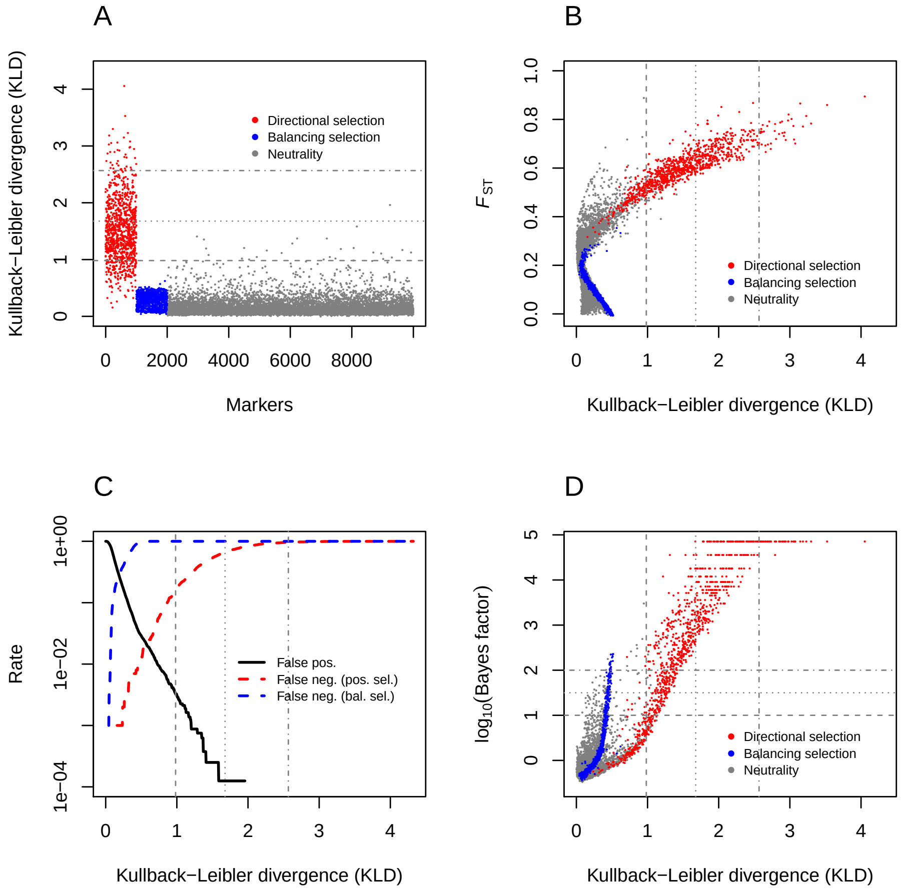

Figure S8 Analysis of the allele count data from dataset 8. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Relationship between the Bayes factor $\log_{10}(\text{BF})$ from the BAYESCAN analysis of dataset 8 and the KLD. The horizontal lines in (A) and the vertical lines in (B–D) indicate the KLD thresholds corresponding to the 95%-, the 99%- and the 99.9%-quantile of the of the KLD distribution from the pod analysis of dataset 8. In (D), the horizontal lines indicate the $\log_{10}(\text{BF}) = 1$, $\log_{10}(\text{BF}) = 1.5$ and $\log_{10}(\text{BF}) = 2$ thresholds, which correspond to ‘‘strong’’, ‘‘very strong’’ and ‘‘decisive’’ support, respectively, following Jeffreys’ (1961) scale of evidence.

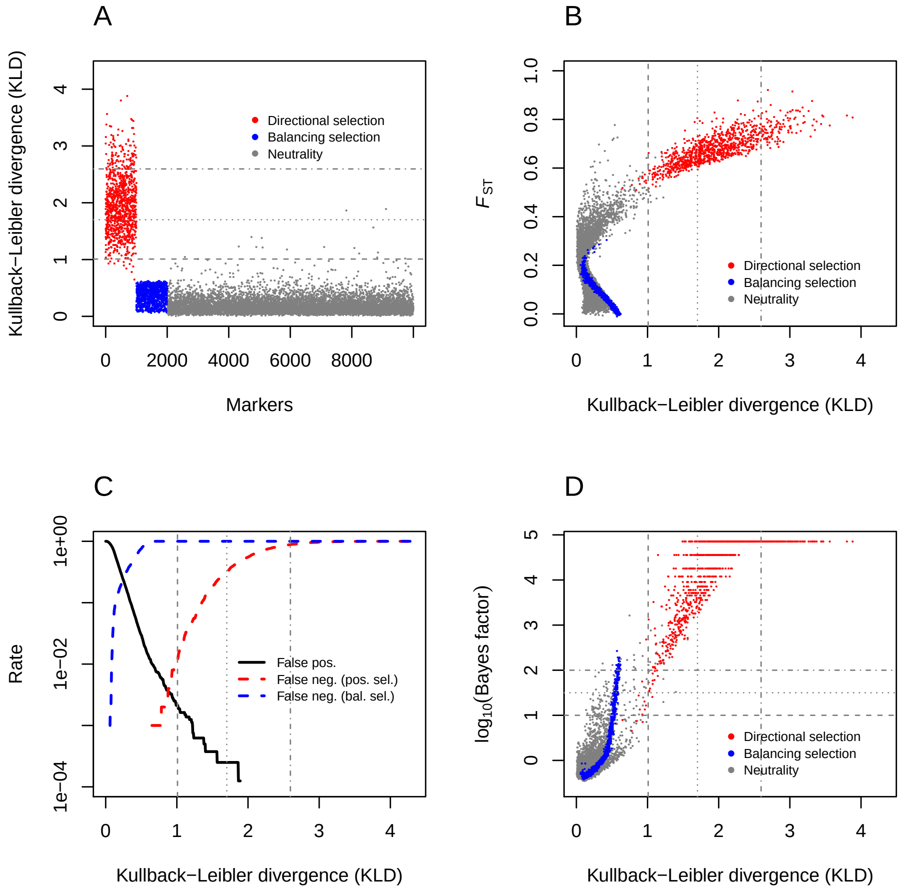

Figure S9 Analysis of the allele count data from dataset 9. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Relationship between the Bayes factor $\log_{10}$(BF) from the BayeScan analysis of dataset 9 and the KLD. The horizontal lines in (A) and the vertical lines in (B–D) indicate the KLD thresholds corresponding to the 95%-, the 99%- and the 99.9%-quantile of the of the KLD distribution from the pod analysis of dataset 9. In (D), the horizontal lines indicate the $\log_{10}$(BF) = 1, $\log_{10}$(BF) = 1.5 and $\log_{10}$(BF) = 2 thresholds, which correspond to ‘‘strong’’, ‘‘very strong’’ and ‘‘decisive’’ support, respectively, following Jeffreys’ (1961) scale of evidence.

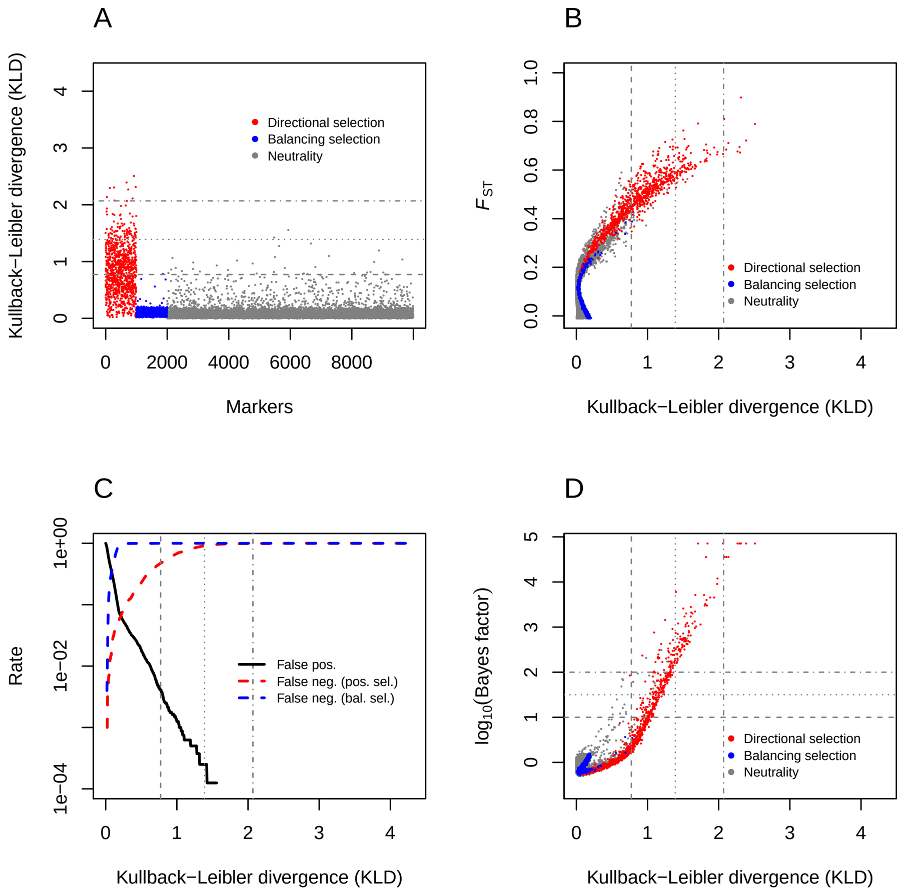

Figure S10 Analysis of the allele count data from dataset 10. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Relationship between the Bayes factor $\log_{10}$(BF) from the BayeScan analysis of dataset 10 and the KLD. The horizontal lines in (A) and the vertical lines in (B–D) indicate the KLD thresholds corresponding to the 95%-, the 99%- and the 99.9%-quantile of the of the KLD distribution from the pod analysis of dataset 10. In (D), the horizontal lines indicate the $\log_{10}$(BF) = 1, $\log_{10}$(BF) = 1.5 and $\log_{10}$(BF) = 2 thresholds, which correspond to ''strong'', ''very strong'' and ''decisive'' support, respectively, following Jeffreys' (1961) scale of evidence.

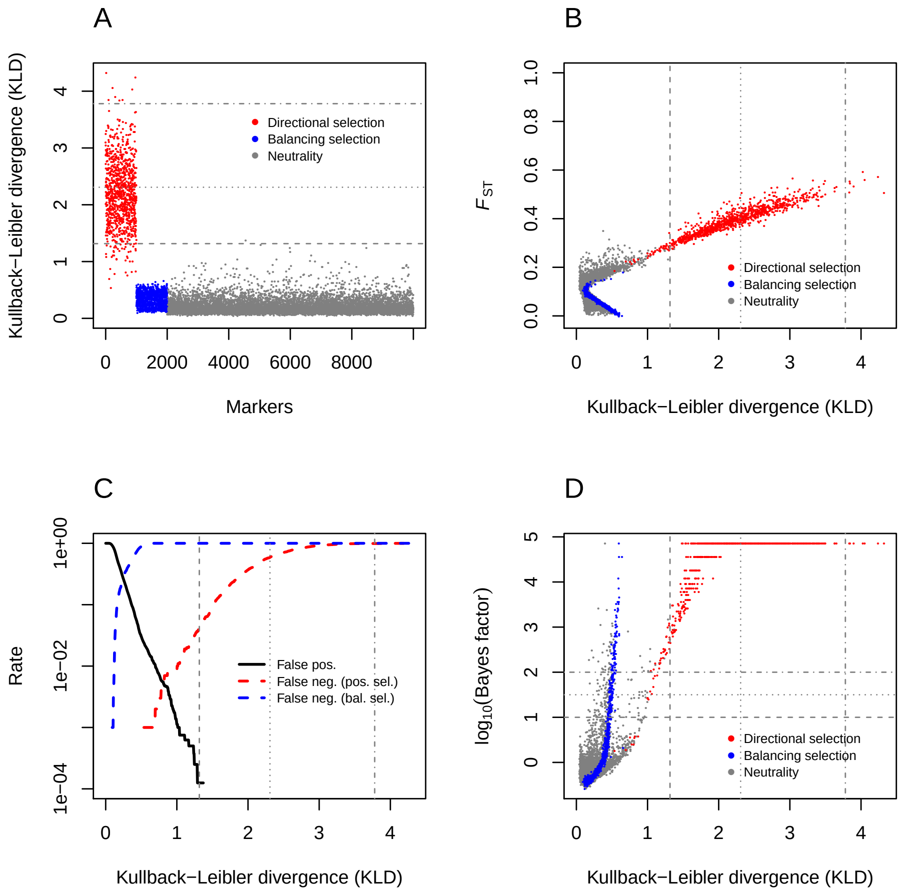

Figure S11 Analysis of the allele count data from dataset 11. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Relationship between the Bayes factor $\log_{10}$(BF) from the BayeScan analysis of dataset 11 and the KLD. The horizontal lines in (A) and the vertical lines in (B–D) indicate the KLD thresholds corresponding to the 95%-, the 99%- and the 99.9%-quantile of the of the KLD distribution from the pod analysis of dataset 11. In (D), the horizontal lines indicate the $\log_{10}$(BF) = 1, $\log_{10}$(BF) = 1.5 and $\log_{10}$(BF) = 2 thresholds, which correspond to ‘‘strong’’, ‘‘very strong’’ and ‘‘decisive’’ support, respectively, following Jeffreys’ (1961) scale of evidence.

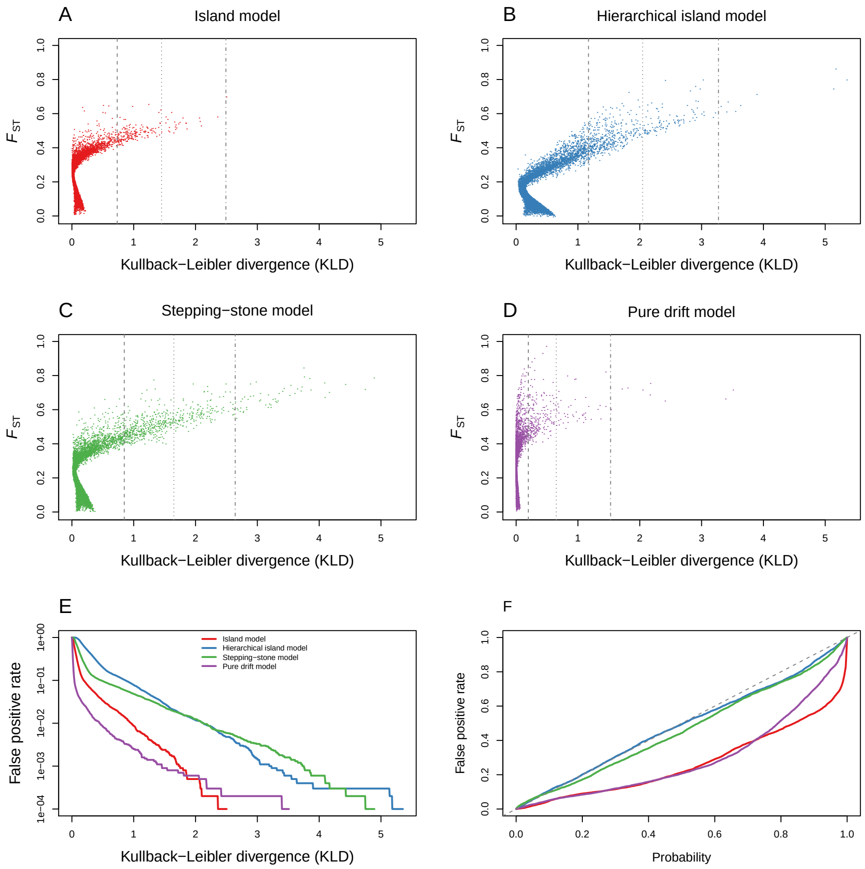

Figure S12 (A–D) SelEstim analysis of the datasets from Figure S1. (E) False positive rate (neutral loci detected as outliers) as a function of the Kullback–Leibler divergence (KLD) threshold, for the datasets analyzed in (A–D). (F) False positive rate, as a function of the quantile probability. For each dataset analysis, pseudo-observed data (pod) are generated from the joint posterior distribution of the model parameters, using a rejection-sampling algorithm (see File S2). The pod is then analyzed, using the same MCMC parameters (number and length of pilot runs, burn-in, chain length, etc.) as for the analysis of the original data. Each quantile probability defines a KLD threshold, which is used for model choice between selection and neutrality.

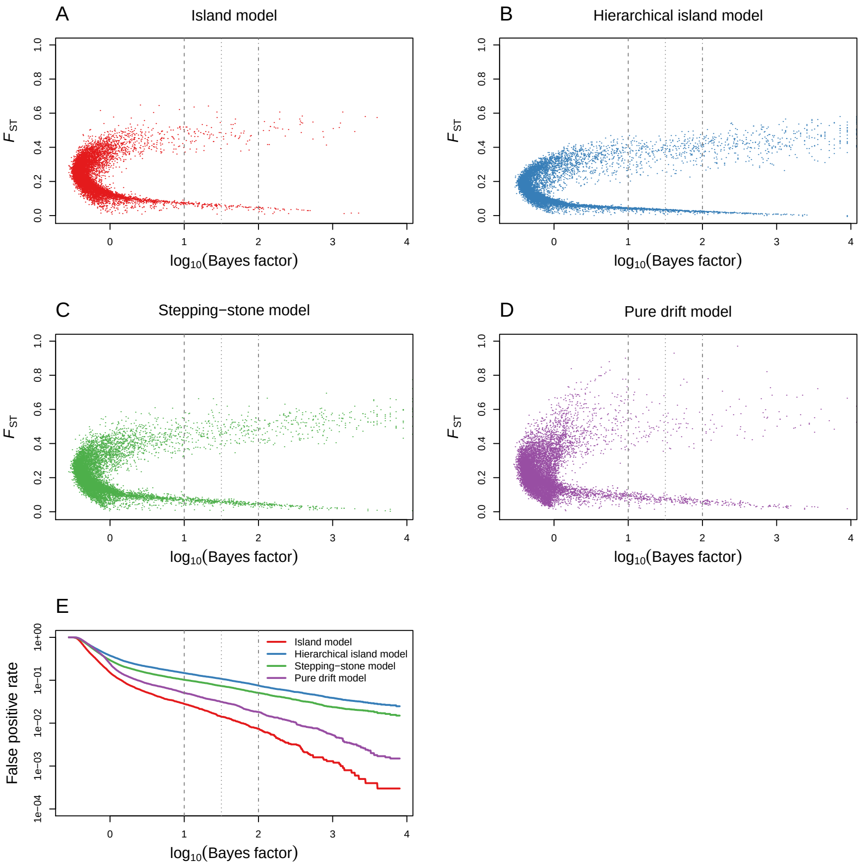

Figure S13 (A–D) BAYESCAN analyses of the datasets from Figure S1, using prior odds of 10 for the neutral model. (E) False positive rate as a function of the $\log_{10}$(BF) threshold. Vertical lines indicate the $\log_{10}$(BF) = 1, $\log_{10}$(BF) = 1.5 and $\log_{10}$(BF) = 2 thresholds, which correspond to ''strong'', ''very strong'' and ''decisive'' support, respectively, following Jeffreys' (1961) scale of evidence.

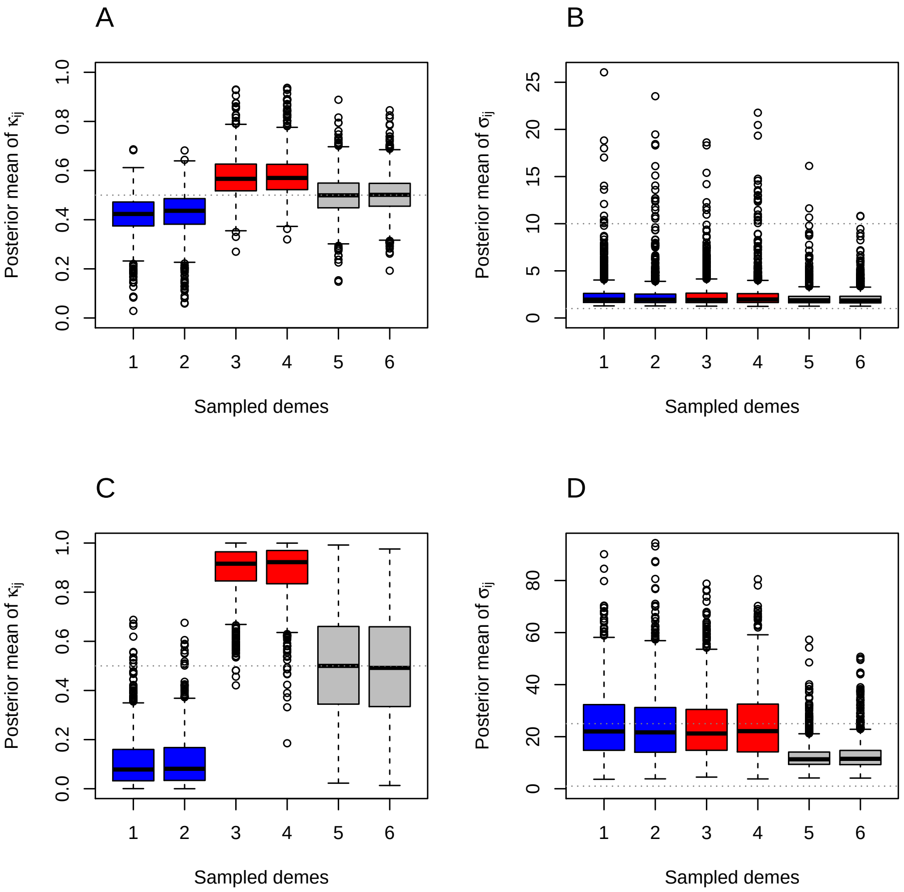

Figure S14 Analysis of the allele count data from datasets 1 and 2. (A) Boxplot representation of the posterior means of the parameters κ_{ij} (that indicate which allele is selected for) for the 1,000 positively selected loci in "blue" demes (1–2), "red" demes (3–4) and "uncolored" demes (5–6) in dataset 1. (B) Boxplot representation of the posterior means of the selection coefficients σ_{ij} for positively selected loci in dataset 1. For "blue" demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the "blue" allele being selected for (κ_{ij} = 0). For "red" demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the "red" allele being selected for (κ_{ij} = 1). The horizontal dotted lines indicate the true value of $\sigma_{ij} \equiv 2Ns_{ij}$ (top) and the prior mean σ_{ij} = 1 (bottom). For "uncolored" demes, the posterior means of the selection coefficients σ_{ij} are unconditional. (C) Idem as (A) for dataset 2. (D) Idem as (B) for dataset 2.

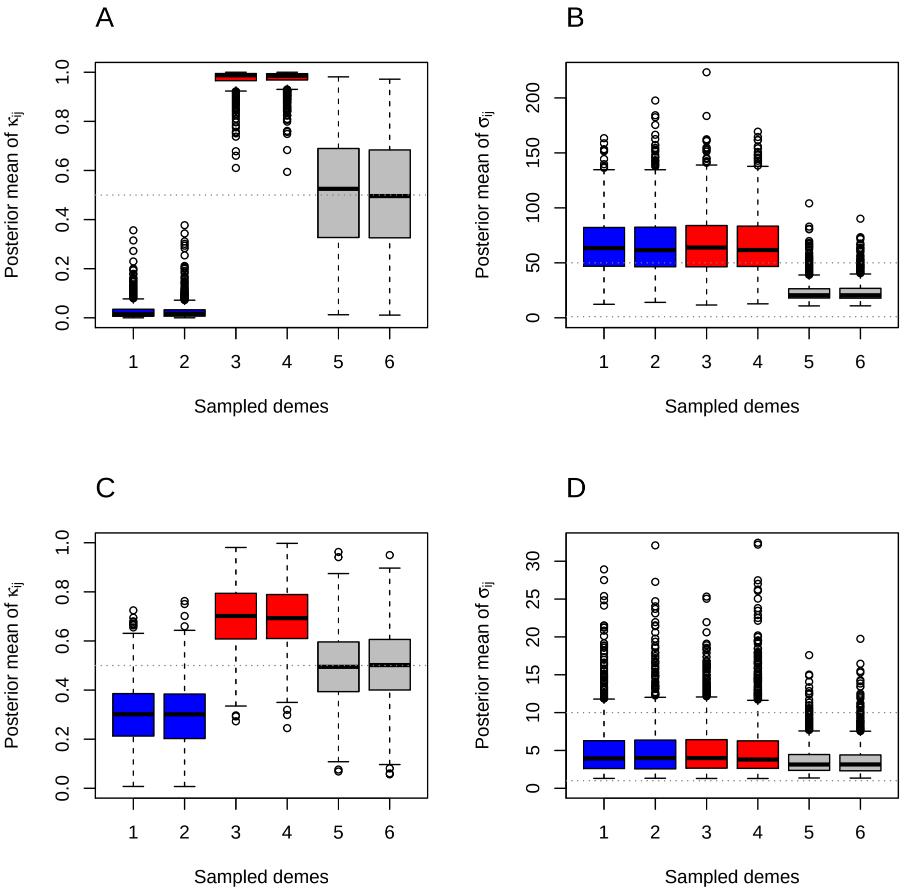

Figure S15 Analysis of the allele count data from datasets 3 and 4. (A) Boxplot representation of the posterior means of the parameters κ_{ij} (that indicate which allele is selected for) for the 1,000 positively selected loci in "blue" demes (1–2), "red" demes (3–4) and "uncolored" demes (5–6) in dataset 3. (B) Boxplot representation of the posterior means of the selection coefficients σ_{ij} for positively selected loci in dataset 3. For "blue" demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the "blue" allele being selected for (κ_{ij} = 0). For "red" demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the "red" allele being selected for (κ_{ij} = 1). The horizontal dotted lines indicate the true value of $\sigma_{ij} \equiv 2Ns_{ij}$ (top) and the prior mean σ_{ij} = 1 (bottom). For "uncolored" demes, the posterior means of the selection coefficients σ_{ij} are unconditional. (C) Idem as (A) for dataset 4. (D) Idem as (B) for dataset 4.

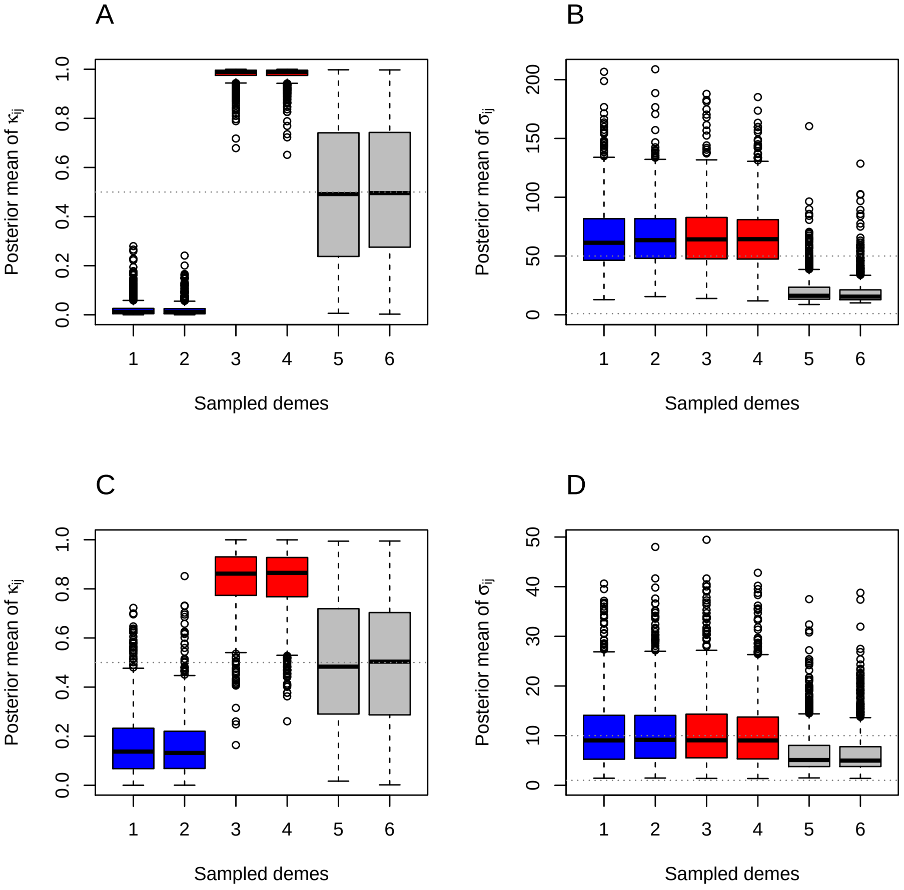

Figure S16 Analysis of the allele count data from datasets 6 and 7. (A) Boxplot representation of the posterior means of the parameters κ_{ij} (that indicate which allele is selected for) for the 1,000 positively selected loci in ''blue'' demes (1–2), ''red'' demes (3–4) and ''uncolored'' demes (5–6) in dataset 6. (B) Boxplot representation of the posterior means of the selection coefficients σ_{ij} for positively selected loci in dataset 6. For ''blue'' demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the ''blue'' allele being selected for (κ_{ij} = 0). For ''red'' demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the ''red'' allele being selected for (κ_{ij} = 1). The horizontal dotted lines indicate the true value of $\sigma_{ij} \equiv 2Ns_{ij}$ (top) and the prior mean σ_{ij} = 1 (bottom). For ''uncolored'' demes, the posterior means of the selection coefficients σ_{ij} are unconditional. (C) Idem as (A) for dataset 7. (D) Idem as (B) for dataset 7.

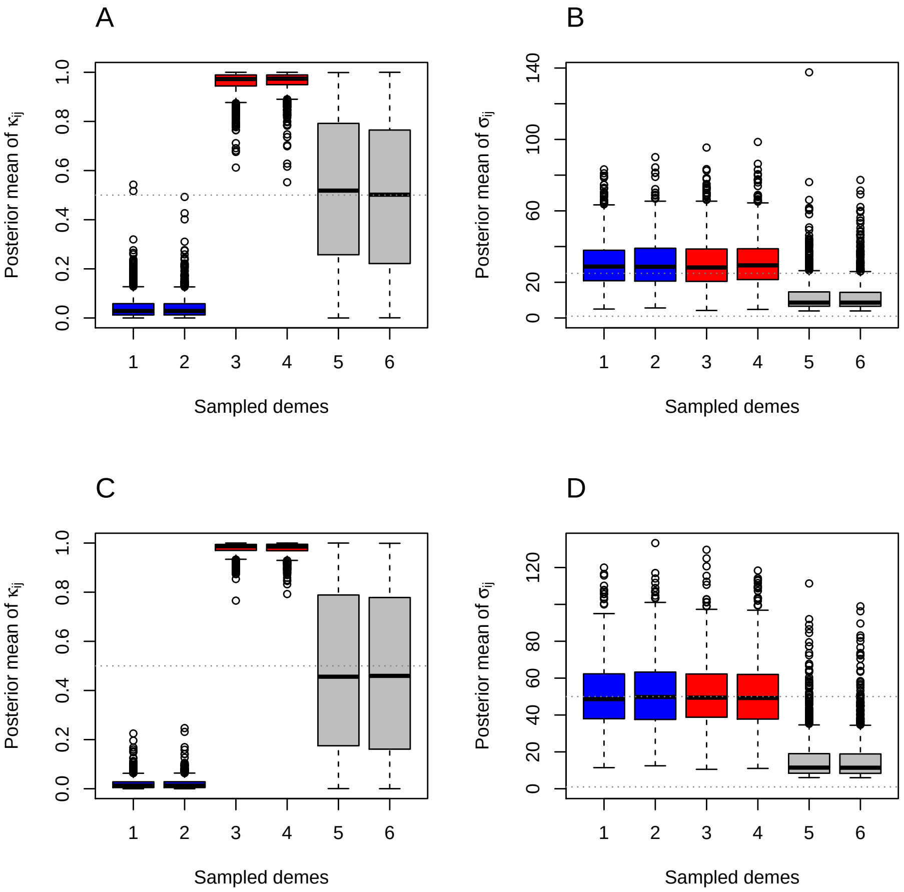

Figure S17 Analysis of the allele count data from datasets 8 and 9. (A) Boxplot representation of the posterior means of the parameters κ_{ij} (that indicate which allele is selected for) for the 1,000 positively selected loci in ‘‘blue’’ demes (1–2), ‘‘red’’ demes (3–4) and ‘‘uncolored’’ demes (5–6) in dataset 8. (B) Boxplot representation of the posterior means of the selection coefficients σ_{ij} for positively selected loci in dataset 8. For ‘‘blue’’ demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the ‘‘blue’’ allele being selected for ($\kappa_{ij} = 0$). For ‘‘red’’ demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the ‘‘red’’ allele being selected for ($\kappa_{ij} = 1$). The horizontal dotted lines indicate the true value of $\sigma_{ij} \equiv 2Ns_{ij}$ (top) and the prior mean $\sigma_{ij} = 1$ (bottom). For ‘‘uncolored’’ demes, the posterior means of the selection coefficients σ_{ij} are unconditional. (C) Idem as (A) for dataset 9. (D) Idem as (B) for dataset 9.

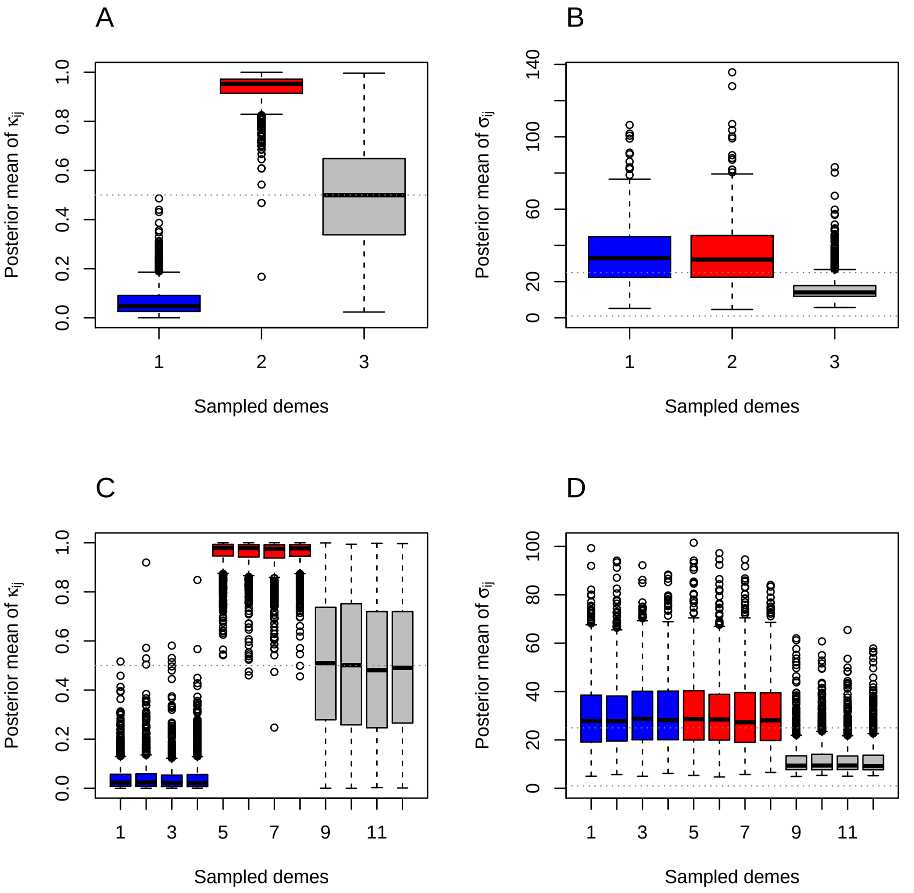

Figure S18 Analysis of the allele count data from datasets 10 and 11. (A) Boxplot representation of the posterior means of the parameters κ_{ij} (that indicate which allele is selected for) for the 1,000 positively selected loci in "blue" demes (1–2), "red" demes (3–4) and "uncolored" demes (5–6) in dataset 10. (B) Boxplot representation of the posterior means of the selection coefficients σ_{ij} for positively selected loci in dataset 10. For "blue" demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the "blue" allele being selected for (κ_{ij} = 0). For "red" demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the "red" allele being selected for (κ_{ij} = 1). The horizontal dotted lines indicate the true value of $\sigma_{ij} \equiv 2Ns_{ij}$ (top) and the prior mean σ_{ij} = 1 (bottom). For "uncolored" demes, the posterior means of the selection coefficients σ_{ij} are unconditional. (C) Idem as (A) for dataset 11. (D) Idem as (B) for dataset 11.

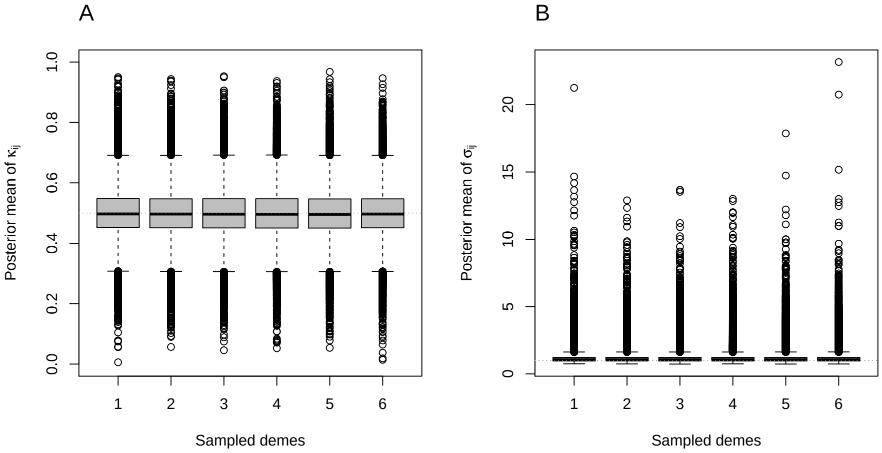

Figure S19 Analysis of the allele count data from a simulation of 50,000 neutral markers. The simulation was performed according to an island model with n_d = 50 ‘‘uncolored’’ demes, each made of N = 250 diploid individuals (500 genes). Samples were collected in six demes (50 individuals per deme). The migration rate was chosen to achieve the expected value of F_{ST} = 0.15, using equation 6 in Rousset (1996). The realized value was F_{ST} = 0.153 (multilocus estimate). (A) Boxplot representation of the posterior means of the parameters κ_{ij} (that indicate which allele is selected for) for the 50,000 neutral markers in ‘‘uncolored’’ demes (1–6). (B) Boxplot representation of the posterior means of the selection coefficients σ_{ij} for the 50,000 neutral markers (unconditional on κ_{ij}). The horizontal dotted line indicates the prior mean σ_{ij} = 1

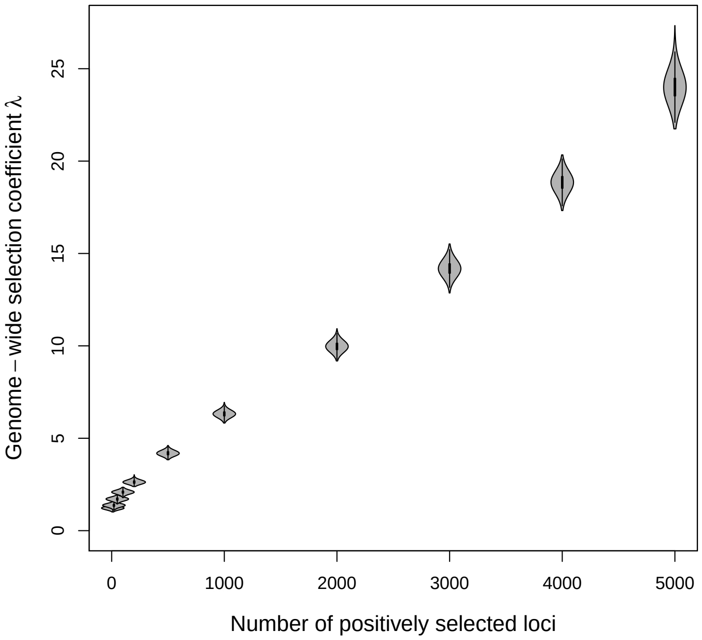

Figure S20 Posterior distributions (violin plot representation) of the genome-wide coefficient of selection λ as a function of the number of positively selected loci. We used the same parameter as for dataset 5 (see Table 1), but varying the proportion of selected loci from 10 to 5,000 out of 10,000 markers (hence, from 0.1% to 50%).

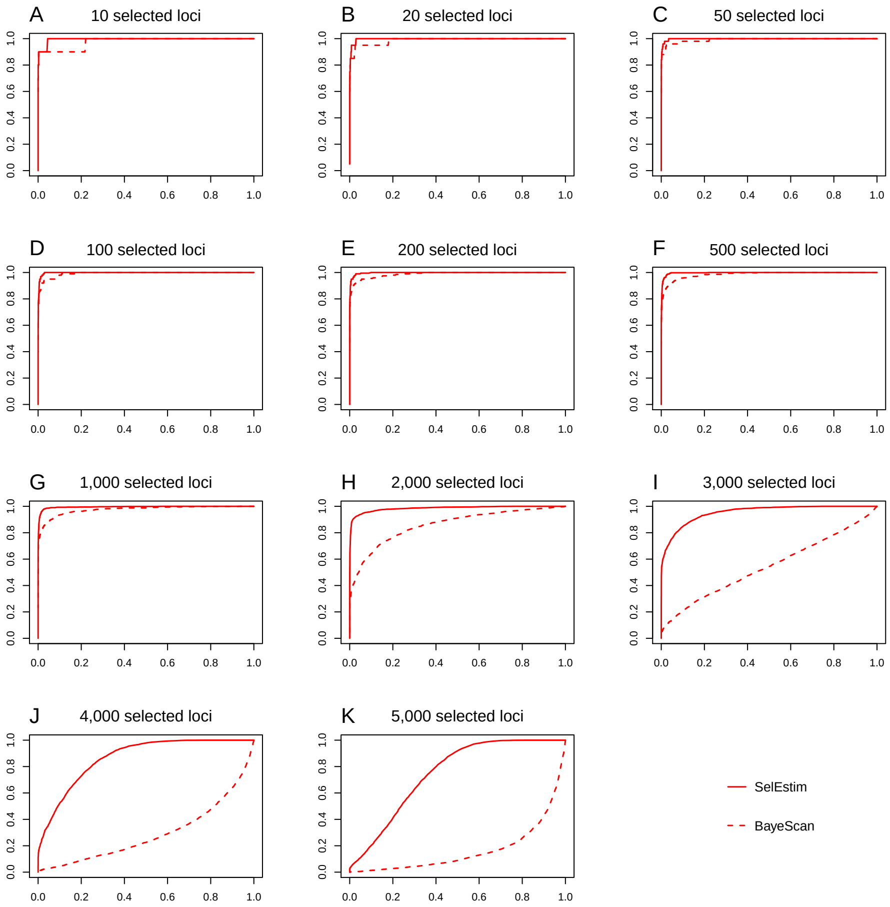

Figure S21 Receiver operating characteristic (ROC) analysis for the same datasets as in Figure S17 (from left to right, top to bottom). In the ROC analysis, the proportion of false positives and true positives is computed for each possible value of the threshold that is used to classify a locus under selection. For SelEstim, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and its centering distribution, while in the case of BayeScan it was the Bayes factor.

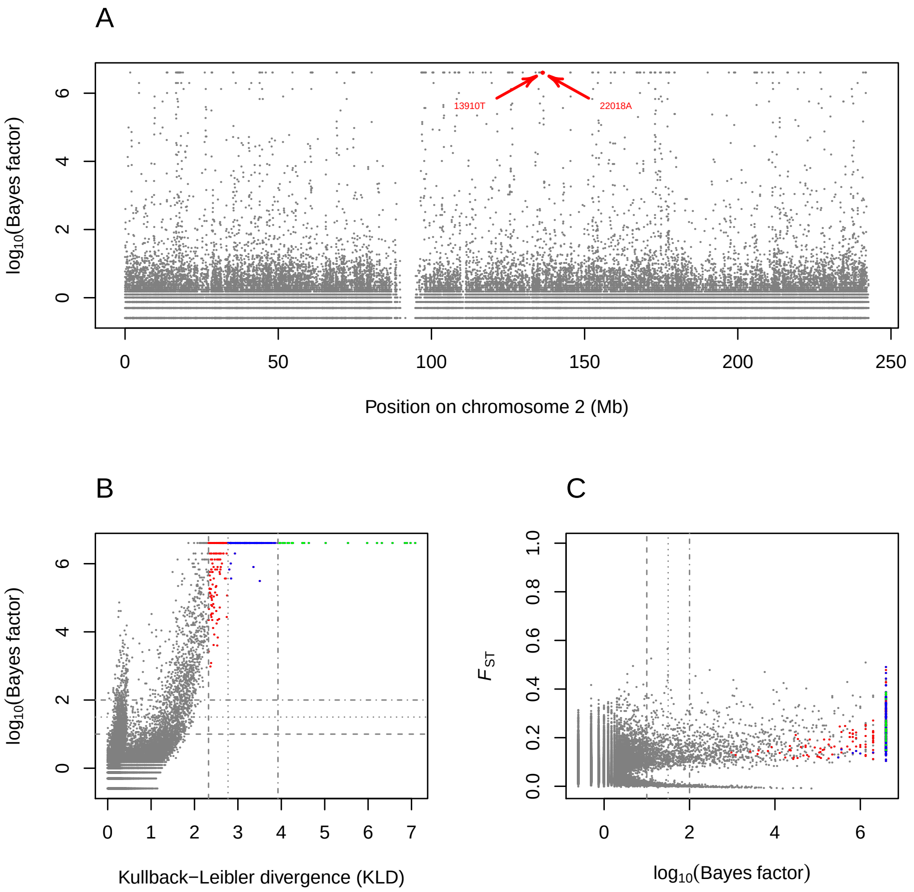

Figure S22 (A) BAYESCAN Bayes factor for the CEPH dataset analyses, along chromosome 2. The alleles -13910T and -22018A associated with lactase persistence are indicated in red. (B) Joint distribution of BAYESCAN Bayes factor and the Kullback–Leibler divergence (KLD) measure for all loci in the dataset. Markers in green have KLD ≥ 3.924, which corresponds to the 99.9%-quantile of the of the KLD distribution from the pod analysis; markers in blue have KLD ≥ 2.772, which corresponds to the 99.5%-quantile of the of the KLD distribution from the pod analysis; markers in red have KLD ≥ 2.324, which corresponds to the 99%-quantile of the of the KLD distribution from the pod analysis. (C) Joint distribution F_{ST} and BAYESCAN Bayes factor for all loci in the dataset.

Table S1 False positive rates using two calibration methods

	False positive rate (KLD)					
	Using McCulloch’s (1989) calibration			Using pseudo-observed data		
Dataset	α = 5%	α = 1%	α = 0.1%	α = 5%	α = 1%	α = 0.1%
12	0.002 (1.164)	0.000 (1.959)	0.000 (3.108)	2.764 (0.011)	0.374 (0.045)	0.026 (0.211)
13	0.000 (1.164)	0.000 (1.959)	0.000 (3.108)	1.226 (0.016)	0.076 (0.076)	0.004 (0.349)
14	0.016 (1.164)	0.010 (1.959)	0.002 (3.108)	3.308 (0.035)	0.514 (0.174)	0.048 (0.801)
15	0.008 (1.164)	0.000 (1.959)	0.000 (3.108)	1.880 (0.091)	0.164 (0.434)	0.002 (1.520)
16	0.068 (1.164)	0.008 (1.959)	0.000 (3.108)	1.722 (0.247)	0.194 (0.853)	0.008 (2.019)
17	0.140 (1.164)	0.020 (1.959)	0.000 (3.108)	1.712 (0.374)	0.186 (1.047)	0.022 (1.942)
18	0.182 (1.164)	0.010 (1.959)	0.000 (3.108)	1.478 (0.521)	0.178 (1.179)	0.010 (1.948)

SelEstim analyses of datasets 12–18. Left-hand side: proportion (%) of markers that were classified as outliers, using the threshold KLD = 1.164, 1.959 and 3.108, which equal the KLD between two Bernoulli distributions corresponding to flipping a fair coin and a biased coin that gives a head with probability 0.05, 0.01 and 0.001, respectively. Right-hand side: proportion (%) of markers that were classified as outliers, using the calibration based on pseudo-observed data (pod). For each dataset and each analysis, a rejection sampling algorithm (see File S2) is used to generate a pod from the joint posterior distribution of the model parameters. The quantiles of the KLD distribution from the pod analysis are then used to calibrate the KLD: the (1 - α)%-quantile of the KLD distribution from the pod analysis provides a α%-threshold KLD value, which is then used for model choice between selection and neutrality.

Table S2 False positive rates using two calibration methods

	False positive rate (KLD)					
	Using McCulloch's (1989) calibration			Using pseudo-observed data		
Dataset	α = 5%	α = 1%	α = 0.1%	α = 5%	α = 1%	α = 0.1%
Island model	0.560 (1.164)	0.050 (1.959)	0.000 (3.108)	1.960 (0.734)	0.250 (1.450)	0.010 (2.491)
Hierarchy	5.680 (1.164)	1.260 (1.959)	0.110 (3.108)	5.580 (1.171)	1.140 (2.047)	0.090 (3.274)
IBD	1.900 (1.164)	0.920 (1.959)	0.330 (3.108)	7.830 (0.121)	3.140 (0.655)	0.680 (2.346)
Pure drift	0.160 (1.164)	0.060 (1.959)	0.020 (3.108)	2.680 (0.197)	0.540 (0.649)	0.090 (1.528)

SelEstim analyses of datasets from Figure S19. Left-hand side: proportion (%) of markers that were classified as outliers, using the threshold KLD = 1.164, 1.959 and 3.108, which equal the KLD between two Bernoulli distributions corresponding to flipping a fair coin and a biased coin that gives a head with probability 0.05, 0.01 and 0.001, respectively. Right-hand side: proportion (%) of markers that were classified as outliers, using the calibration based on pseudo-observed data (pod). For each dataset and each analysis, a rejection sampling algorithm (see File S2) is used to generate a pod from the joint posterior distribution of the model parameters. The quantiles of the KLD distribution from the pod analysis are then used to calibrate the KLD: the (1 - α)%-quantile of the KLD distribution from the pod analysis provides a α%-threshold KLD value, which is then used for model choice between selection and neutrality.

File S1

Details on the componentwise Markov chain Monte Carlo algorithm

Here we provide the computational details for the componentwise Markov chain Monte Carlo updates. Our aim is to sample from the joint posterior distribution of $f(\mathbf{M}, \pi, \kappa, \sigma, \delta, \lambda|\mathbf{n})$, which is specified by equation (4) and by the directed acyclic graph (DAG) in Figure 1. To do so, we use a combination of the Metropolis–Hastings algorithm and the Gibbs sampler for generating observations from $f(\mathbf{M}, \pi, \kappa, \sigma, \delta, \lambda|\mathbf{n})$ using outputs from a Markov chain (see, e.g., Gelman *et al.* 2004).

Each Markov chain is initialized with random values of the parameters drawn from their prior densities, except for the parameters p_{ij}, for which the observed frequencies are used, and the parameters π_js, for which the Laplace values are calculated from the dataset frequencies. The updating sequence is as follows: (*i*) all Ln_d parameters p_{ij}; (*ii*) all n_d parameters M_i; (*iii*) all L parameters π_j; (*iv*) the hyperparameter λ; (*v*) all L hyperparameters δ_j; (*vi*) all Ln_d parameters σ_{ij}; (*vii*) all Ln_d parameters κ_{ij}. Since the full posterior distribution of the model can be decomposed as a product over loci and over populations (see equation 4), each update only requires the re-computation of the relevant terms of the distribution $f(\mathbf{M}, \pi, \kappa, \sigma, \delta, \lambda|\mathbf{n})$. This improves the computational efficiency of the algorithm considerably.

The confluent hypergeometric, or Kummer's, functions ${}_1F_1(a; b; z)$ (see, e.g., Abramowitz and Stegun 1965, p. 504) were computed following a procedure proposed by Pearson (Pearson 2009), which is based on the power series definition of the function:

$$ {}_1F_1(a; b; z) = \sum_{j=0}^{\infty} \underbrace{\frac{(a)_j}{(b)_j}\frac{z^j}{j!}}_{A_j}, \tag{S1.1}$$

where, for some parameter p, the Pochhammer symbol $(p)_j$ is defined as:

$$ (p)_0 = 1, \quad (p)_j = p(p+1)\ldots(p+j-1), \quad \text{for } j = 1, 2, \ldots\ . \tag{S1.2}$$

The computation of the terms of the power series in equation (S1.1) can then be carried out using the following procedure:

$$\begin{aligned}
&A_0 = S_0 = 1, \\
&A_{j+1} = A_j \frac{a+j}{b+j} \frac{z}{j+1}, \\
&S_{j+1} = S_j + A_{j+1}, \quad \text{for } j = 1, 2, \ldots
\end{aligned} \tag{S1.3}$$

where A_j represents the $(j+1)$th term of the power series in equation (S1.1), and S_j represents the sum of the first $(j+1)$ terms. The computation was stopped when both $|A_N|/|S_{N-1}| < 10^{-12}$ and $|A_{N+1}|/|S_N| < 10^{-12}$. This criterion is equivalent to truncating the series in equation (S1.1), and requires that two consecutive terms to be small compared to the sum already computed.

Updating p_{ij}**:** The parameters p_{ij} are updated iteratively in each deme, one locus at a time. In the ith deme, at locus j, one allele is chosen at random from a Bernoulli trial with probability 0.5. The new allele frequency p'_{ij} is chosen as a random variable drawn from a uniform distribution around the current value p_{ij}:

$$p'_{ij} \sim U\left(p_{ij} - \Delta_p, p_{ij} + \Delta_p\right). \tag{S1.4}$$

The size of the interval Δ_p is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40 (see, e.g., Gilks *et al.* 1996). Since p_{ij} is a frequency comprised between 0 and 1, if p'_{ij} is outside the interval $[0, 1]$, the excess is reflected back into the interval; that is, if $p'_{ij} < 0$ then p'_{ij} is reset to its absolute value $|p'_{ij}|$, and if $p'_{ij} > 1$ then p'_{ij} is reset to $2 - p'_{ij}$. This proposal is symmetric (Yang 2005). The updated allele frequency p'_{ij} is therefore accepted according to the appropriate Metropolis probability, which reads:

$$1 \wedge \frac{\mathcal{L}(p'_{ij}; \mathbf{n}_{ij})\psi(p'_{ij}; M_i, \pi_j, \kappa_{ij}, \sigma_{ij})}{\mathcal{L}(p_{ij}; \mathbf{n}_{ij})\psi(p_{ij}; M_i, \pi_j, \kappa_{ij}, \sigma_{ij})}. \tag{S1.5}$$

Equation (S1.5) can be rewritten as

$$1\wedge\exp\left[\sigma_{ij}\left(\tilde{p}'_{ij}-\tilde{p}_{ij}\right)\right]\frac{p'^{\,x_{ij}+M_i\pi_j-1}_{ij}(1-p'_{ij})^{(n_{ij}-x_{ij})M_i+(1-\pi_j)-1}}{p^{x_{ij}+M_i\pi_j-1}_{ij}(1-p_{ij})^{(n_{ij}-x_{ij})M_i+(1-\pi_j)-1}}, \tag{S1.6}$$

where $\tilde{p}'_{ij}\equiv\kappa_{ij}(1-p'_{ij})+(1-\kappa_{ij})p'_{ij}$.

Updating M_i: The parameters M_i are updated iteratively, one deme at a time. The proposed value M'_i is drawn from a lognormal distribution with median equal to the current value M_i, i.e.:

$$q(M_i\to M'_i)=\frac{1}{M'_i\nu_M\sqrt{2\pi}}\exp\left(\frac{-\ln(M'_i/M_i)^2}{2\nu_M^2}\right), \tag{S1.7}$$

where ν_M is the standard deviation on the log scale. The standard deviation ν_M is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. Because the lognormal jumping rule is asymmetric, a Metropolis–Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities (which is sometimes referred to as the ''Hastings term'': see, e.g., Gelman *et al.* 2004, p. 291). This means that when some moves are more likely to happen (because of the asymmetry of the proposal distribution), their probability of acceptance is decreased proportionately. Here, the ratio $q(M'_i\to M_i)/q(M_i\to M'_i)$ reduces to M'_i/M_i. In order to avoid computational problems with excessively small or large M_i values, all moves falling outside the interval $[0.0011,000]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value M'_i is accepted according to the appropriate Metropolis–Hastings probability, which is:

$$1\wedge\frac{\left[\prod_{j=1}^{L}\psi(p_{ij};M'_i,\pi_j,\kappa_{ij},\sigma_{ij})\right]f(M'_i)q(M'_i\to M_i)}{\left[\prod_{j=1}^{L}\psi(p_{ij};M_i,\pi_j,\kappa_{ij},\sigma_{ij})\right]f(M_i)q(M_i\to M'_i)}. \tag{S1.8}$$

Equation (S1.8) can be rewritten as

$$1\wedge\left[\frac{\Gamma(M_i)}{\Gamma(M'_i)}\right]^L \frac{\prod_{j=1}^{L}\Gamma(M_i\pi_j)\Gamma(M_i(1-\pi_j))_1F_1(M_i\tilde{\pi}_{ij};M_i;\sigma_{ij})p_{ij}^{M'_i\pi_j}(1-p_{ij})^{M'_i(1-\pi_j)}}{\prod_{j=1}^{L}\Gamma(M'_i\pi_j)\Gamma(M'_i(1-\pi_j))_1F_1(M'_i\tilde{\pi}_{ij};M'_i;\sigma_{ij})p_{ij}^{M_i\pi_j}(1-p_{ij})^{M_i(1-\pi_j)}} \qquad \text{(S1.9)}$$

Updating π_j: The parameters π_j are updated iteratively, one locus at a time. In the ith deme, at locus j, one allele is chosen at random from a Bernoulli trial with probability 0.5. The proposed allele frequency π'_j is chosen as a random variable drawn from a uniform distribution around the current value π_j:

$$\pi'_j \sim U\left(\pi_j - \Delta_\pi, \pi_j + \Delta_\pi\right). \qquad \text{(S1.10)}$$

The size of the interval Δ_π is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. Since π_j is a frequency comprised between 0 and 1, if π'_j is outside the interval $[0,1]$, the excess is reflected back into the interval; that is, if $\pi'_j < 0$ then π'_j is reset to its absolute value $|\pi'_j|$, and if $\pi'_j > 1$ then π'_j is reset to $2-\pi'_j$. This proposal is symmetric, and the updated allele frequency π'_j is therefore accepted according to the appropriate Metropolis probability, which reads:

$$1\wedge\frac{\left[\prod_{i=1}^{n_\mathrm{d}}\psi(p_{ij};M_i,\pi'_j,\kappa_{ij},\sigma_{ij})\right]f(\pi'_j)}{\left[\prod_{i=1}^{n_\mathrm{d}}\psi(p_{ij};M_i,\pi_j,\kappa_{ij},\sigma_{ij})\right]f(\pi_j)}. \qquad \text{(S1.11)}$$

Equation (S1.11) can be rewritten as

$$1\wedge\frac{\prod_{i=1}^{n_\mathrm{d}}\Gamma(M_i\pi_j)\Gamma(M_i(1-\pi_j))_1F_1(M_i\tilde{\pi}_{ij};M_i;\sigma_{ij})p_{ij}^{M_i\pi'_j}(1-p_{ij})^{M_i(1-\pi'_j)}}{\prod_{i=1}^{n_\mathrm{d}}\Gamma(M_i\pi'_j)\Gamma(M_i(1-\pi'_j))_1F_1(M_i\tilde{\pi}'_{ij};M_i;\sigma_{ij})p_{ij}^{M_i\pi_j}(1-p_{ij})^{M_i(1-\pi_j)}}, \qquad \text{(S1.12)}$$

where $\tilde{\pi}'_{ij} \equiv \kappa_{ij}(1-\pi'_j)+(1-\kappa_{ij})\pi'_j$.

Updating λ: The proposed value of the hyperparameter λ' is drawn from a lognormal distribution with median equal to the current value λ, i.e.:

$$q(\lambda \to \lambda') = \frac{1}{\lambda' \nu_\lambda \sqrt{2\pi}} \exp\left(\frac{-\ln(\lambda'/\lambda)^2}{2\nu_\lambda^2}\right), \tag{S1.13}$$

where ν_λ is the standard deviation on the log scale. The standard deviation ν_λ is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. Because the lognormal jumping rule is asymmetric, a Metropolis–Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities. This means that when some moves are more likely to happen (because of the asymmetry of the proposal distribution), their probability of acceptance is decreased proportionately. Here, the ratio $q(\lambda' \to \lambda)/q(\lambda \to \lambda')$ reduces to λ'/λ. In order to avoid computational problems with excessively small or large λ' values, all moves falling outside the interval $[0, 500]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value λ' is accepted according to the appropriate Metropolis–Hastings probability, which is:

$$1 \wedge \frac{\left[\prod_{j=1}^{L} f(\delta_j|\lambda')\right] f(\lambda'|\Lambda) q(\lambda' \to \lambda)}{\left[\prod_{j=1}^{L} f(\delta_j|\lambda)\right] f(\lambda|\Lambda) q(\lambda \to \lambda')}. \tag{S1.14}$$

Equation (S1.14) can be rewritten as

$$1 \wedge \left(\frac{\lambda}{\lambda'}\right)^{L-1} \exp\left[(\lambda' - \lambda)\left(\frac{\sum_{j=1}^{L} \delta_j}{\lambda\lambda'} - \frac{1}{\Lambda}\right)\right] \tag{S1.15}$$

Updating δ_j: The parameters δ_j are updated iteratively, one locus at a time. The proposed value of the hyperparameters δ'_j is drawn from a lognormal distribution with median equal to the current value δ_j, i.e.:

$$q(\delta_j \to \delta'_j) = \frac{1}{\delta'_j \nu_\delta \sqrt{2\pi}} \exp\left(\frac{-\ln(\delta'_j/\delta_j)^2}{2\nu_\delta^2}\right), \tag{S1.16}$$

where ν_δ is the standard deviation on the log scale. The standard deviation ν_δ is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. Because the lognormal jumping rule is asymmetric, a Metropolis–Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities. This means that when some moves are more likely to happen (because of the asymmetry of the proposal distribution), their probability of acceptance is decreased proportionately. Here, the ratio $q(\delta'_j \to \delta_j)/q(\delta_j \to \delta'_j)$ reduces to δ'_j/δ_j. In order to avoid computational problems with excessively small or large δ_j values, all moves falling outside the interval $[0, 500]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value δ'_j is accepted according to the appropriate Metropolis–Hastings probability, which is:

$$1 \wedge \frac{\left[\prod_{i=1}^{n_\mathrm{d}} f(\sigma_{ij}|\delta'_j)\right] f(\delta'_j|\lambda) q(\delta'_j \to \delta_j)}{\left[\prod_{i=1}^{n_\mathrm{d}} f(\sigma_{ij}|\delta_j)\right] f(\delta_j|\lambda) q(\delta_j \to \delta'_j)}. \tag{S1.17}$$

Equation (S1.17) can be rewritten as

$$1 \wedge \left(\frac{\delta_j}{\delta'_j}\right)^{n_\mathrm{d}-1} \exp\left[\left(\delta'_j - \delta_j\right)\left(\frac{\sum_{i=1}^{n_\mathrm{d}} \sigma_{ij}}{\delta_j \delta'_j} - \frac{1}{\lambda}\right)\right] \tag{S1.18}$$

Updating σ_{ij}: The parameters σ_{ij} are updated iteratively in each deme, one locus at a time. In the ith deme, at locus j, the proposed value of the parameters σ'_{ij} is drawn from a lognormal distribution with median equal to the current value σ_{ij}, i.e.:

$$q(\sigma_{ij} \to \sigma'_{ij}) = \frac{1}{\sigma'_{ij}\nu_\sigma\sqrt{2\pi}} \exp\left(\frac{-\ln(\sigma'_{ij}/\sigma_{ij})^2}{2\nu_\sigma^2}\right), \tag{S1.19}$$

where ν_σ is the standard deviation on the log scale. The standard deviation ν_σ is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. Because the lognormal jumping rule is asymmetric, a Metropolis–Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities. This means that when some moves are more likely to happen (because of the asymmetry of the proposal distribution), their probability of acceptance is decreased proportionately. Here, the ratio $q(\sigma'_{ij} \to \sigma_{ij})/q(\sigma_{ij} \to \sigma'_{ij})$ reduces to σ'_{ij}/σ_{ij}. In order to avoid computational problems with excessively small or large σ_{ij} values, all moves falling outside the interval $[0, 500]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value σ'_{ij} is accepted according to the appropriate Metropolis–Hastings probability, which is:

$$\frac{\psi(p_{ij}; M_i, \pi_j, \kappa_{ij}, \sigma'_{ij}) f(\sigma'_{ij}|\delta_j) q(\sigma'_{ij} \to \sigma_{ij})}{\psi(p_{ij}; M_i, \pi_j, \kappa_{ij}, \sigma_{ij}) f(\sigma_{ij}|\delta_j) q(\sigma_{ij} \to \sigma'_{ij})}. \tag{S1.20}$$

Equation (S1.20) can be rewritten as

$$\frac{\sigma'_{ij}}{\sigma_{ij}} \exp\left[(\sigma'_{ij} - \sigma_{ij}) \left(\tilde{p}_{ij} - \frac{1}{\delta_j}\right)\right] \frac{{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij})}{{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma'_{ij})}. \tag{S1.21}$$

Updating κ_{ij}: The parameters κ_{ij} are updated iteratively in each deme, one locus at a time. In the ith deme, at locus j, the variable κ_{ij}, which indicates which of the two alleles is selected for, is updated using Gibbs sampling based on the conditional posterior distribution:

$$f(\kappa_{ij}|\theta_{[-\kappa_{ij}]}) \propto \psi(p_{ij}; M_i, \pi_j, \kappa_{ij}, \sigma_{ij}) f(\kappa_{ij}), \tag{S1.22}$$

where $\theta_{[-\kappa_{ij}]}$ represents all the model parameters but κ_{ij}. Since κ_{ij} can only take two integer values (0 and 1), it can be shown that:

$$\Pr(\kappa_{ij} = 0|\theta_{[-\kappa_{ij}]}) \propto \frac{1}{2}\left[\frac{\exp\left[\sigma_{ij} p_{ij}\right]}{{}_1F_1(M_i\pi_j; M_i; \sigma_{ij})}\right], \tag{S1.23}$$

and

$$\Pr(\kappa_{ij} = 1|\theta_{[-\kappa_{ij}]}) \propto \frac{1}{2}\left[\frac{\exp\left[\sigma_{ij}(1 - p_{ij})\right]}{{}_1F_1(M_i(1 - \pi_j); M_i; \sigma_{ij})}\right]. \tag{S1.24}$$

Therefore, the conditional posterior distribution of $\left(\kappa_{ij}|\theta_{[-\kappa_{ij}]}\right)$ from equation (S1.22) can be rewritten as

$$\left(\kappa_{ij}|\theta_{[-\kappa_{ij}]}\right) \sim \mathrm{Bernoulli}\left(\rho\right), \tag{S1.25}$$

where

$$\begin{aligned}\rho &\equiv \frac{\Pr(\kappa_{ij}=0|\theta_{[-\kappa_{ij}]})}{\Pr(\kappa_{ij}=0|\theta_{[-\kappa_{ij}]})+\Pr(\kappa_{ij}=1|\theta_{[-\kappa_{ij}]})}\\ &= \left[1+\frac{{}_1F_1(M_i\pi_{ij};M_i;\sigma_{ij})}{{}_1F_1(M_i(1-\pi_{ij});M_i;\sigma_{ij})}\exp\left[\sigma_{ij}(1-2p_{ij})\right]\right]^{-1}.\end{aligned} \tag{S1.26}$$

Literature Cited

Abramowitz, M., and I. A. Stegun, 1965 *Handbook of Mathematical Functions*. Dover Publication, Inc., New York.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2004 *Bayesian Data Analysis*. Chapman & Hall, New York, 2nd edition.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, 1996 *Markov Chain Monte Carlo in Practice*. Chapman & Hall, New York, 2nd edition.

Pearson, J., 2009 *Computation of Hypergeometric Functions*. Ph.D. thesis, University of Oxford.

Yang, Z., 2005 Bayesian inference in molecular phylogenetics. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*. Oxford University Press, Oxford, 63–90.

File S2

Details on the algorithm to sample from the inference model

In order to provide a decision criterion for discriminating between neutral and selected markers, we calibrate the Kullback–Leibler divergence (KLD) using simulations from a predictive distribution based on the observed data set. To that end, we generate pseudo-observed data as follows.

We set the hyperparameters M_i, π_j and λ to their respective posterior means $\bar{M}_i$, $\bar{\pi}_j$ and $\bar{\lambda}$, as estimated from the MCMC. Then we draw δ_j from an exponential distribution $\sim \exp(\bar{\lambda}^{-1})$ and we draw σ_{ij} from an exponential distribution $\sim \exp\left(\delta_j^{-1}\right)$. Last, the parameter κ_{ij} is drawn from a Bernoulli distribution (with parameter the posterior mean $\bar{\kappa}_{ij}$).

We aim at sampling the allele frequency p_{ij} from the distribution with density $f(p_{ij})$ defined by equations 2 and 3 in the main text. Because the cumulative distribution function of the distribution with density $f(p_{ij})$ is not tractable, we use a rejection-sampling algorithm. To that end, we define an instrumental distribution $g(p_{ij}) \sim \mathrm{Beta}(M_i\pi_j, M_i(1-\pi_j))$, with density:

$$g(p_{ij}) = \frac{\Gamma(M_i)}{\Gamma(M_i\pi_j)\Gamma(M_i(1-\pi_j))} p_{ij}^{M_i\pi_j-1}(1-p_{ij})^{M_i(1-\pi_j)-1} \qquad \text{(S2.1)}$$

We further need to define a constant u, such that $f(p_{ij}) \leq [ug(p_{ij})]$ over the support $[0,1]$. Noting that:

$$\frac{f(p_{ij})}{g(p_{ij})} = \frac{\exp(\sigma_{ij}\tilde{p}_{ij})}{{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij})} \qquad \text{(S2.2)}$$

then, if we define $u \equiv \exp(\sigma_{ij})/{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij})$ we get:

$$\frac{f(p_{ij})}{ug(p_{ij})} = \exp(\sigma_{ij}(\tilde{p}_{ij}-1)) \qquad \text{(S2.3)}$$

Since $0 \leq \tilde{p}_{ij} \leq 1$ and $\sigma_{ij} \geq 0$, by definition, we have $\exp(\sigma_{ij}(\tilde{p}_{ij}-1)) \leq 1$ and therefore $f(p_{ij}) \leq [ug(p_{ij})]$. A straightforward algorithm to sample from the distribution with density $f(p_{ij})$ is then:

(1) Sample x from a beta distribution $\mathrm{Beta}(M_i\pi_j, M_i(1-\pi_j))$ and y from $\mathcal{U}(0,1)$ (the uniform distribution over the unit interval).

(2) Check whether or not $y < f(x)/[ug(x)]$ or equivalently (see equation S2.3) if $\log(y) < \sigma_{ij}(\tilde{p}_{ij} - 1)$:

- If this holds, accept x and set $\tilde{p}_{ij} = x$;
- if not, reject the value of x and repeat the sampling step (1).

(3) Compute $p_{ij} = \tilde{p}_{ij}(1-\kappa_{ij}) + (1-\tilde{p}_{ij})\kappa_{ij}$.

Finally, we draw the allele counts $\mathbf{n}_{ij}$ in the ith deme at the jth locus by a random draw from the binomial distribution $\sim \mathcal{B}(n_{ij}, p_{ij})$. We repeat this procedure for each locus j in each deme i.

This algorithm is computationally efficient, since it avoids computing ${}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij})$ (see equations 2 and 3 in the main text). However, the efficiency of the algorithm may be very low for large values of σ_{ij}. This is so because the expected number of iterations required until an x is successfully generated is exactly the bounding constant $u \equiv \exp(\sigma_{ij})/{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij})$. Therefore, to avoid the algorithm getting stuck in very long loops, we adopt an alternative strategy whenever $u > 10^4$: in such case, we draw x from a beta distribution $\mathrm{Beta}(\alpha, \beta)$ with the same first two moments as the target distribution (equations 2 and 3 in the main text). Little algebra shows that: $\alpha = m_1(m_2 - m_1)/(m_1^2 - m_2)$ and $\beta = \alpha(1/m_1 - 1)$, where

$$m_1 = \tilde{\pi}_{ij}\left(\frac{{}_1F_1(M_i\tilde{\pi}_{ij}+1; M_i+1; \sigma_{ij})}{{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij})}\right) \tag{S2.4}$$

and

$$m_2 = \tilde{\pi}_{ij}\left(\frac{M_i\tilde{\pi}_{ij}+1}{M_i+1}\right)\left(\frac{{}_1F_1(M_i\tilde{\pi}_{ij}+2; M_i+2; \sigma_{ij})}{{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij})}\right) \tag{S2.5}$$