Identifying Loci Under Selection Against Gene Flow in Isolation-with-Migration Models

Vitor C. Sousa,*,[1,2] Miguel Carneiro,[†] Nuno Ferrand,[†] and Jody Hey*,[1]

*Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, and [†]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, 4099-002 Porto, Portugal

ABSTRACT When divergence occurs in the presence of gene flow, there can arise an interesting dynamic in which selection against gene flow, at sites associated with population-specific adaptations or genetic incompatibilities, can cause net gene flow to vary across the genome. Loci linked to sites under selection may experience reduced gene flow and may experience genetic bottlenecks by the action of nearby selective sweeps. Data from histories such as these may be poorly fitted by conventional neutral model approaches to demographic inference, which treat all loci as equally subject to forces of genetic drift and gene flow. To allow for demographic inference in the face of such histories, as well as the identification of loci affected by selection, we developed an isolation-with-migration model that explicitly provides for variation among genomic regions in migration rates and/or rates of genetic drift. The method allows for loci to fall into any of multiple groups, each characterized by a different set of parameters, thus relaxing the assumption that all loci share the same demography. By grouping loci, the method can be applied to data with multiple loci and still have tractable dimensionality and statistical power. We studied the performance of the method using simulated data, and we applied the method to study the divergence of two subspecies of European rabbits (*Oryctolagus cuniculus*).

UNDERSTANDING speciation requires that we determine the role played by natural selection, as well as the roles of gene exchange and other demographic factors (Dobzhansky 1951; Maynard Smith 1966; Bush 1975; Endler 1977; Templeton 1981; Arnold 1997; Barton 2001). The genetic patterns of present-day populations potentially harbor much information about these processes, and in recent years investigators have developed sophisticated methods for quantifying different kinds of factors, including levels of gene flow between populations (Beerli and Felsenstein 1999; Nielsen and Wakeley 2001), admixture proportions (Chikhi *et al.* 2001), times of population separation (Nielsen and Wakeley 2001), and rates of population size change (Beaumont 1999; Hey 2005). This progress has been possible mainly through the development of full-likelihood model-based statistical methods that draw upon the coalescent theory of gene genealogies. However, quantifying the effects of selection in the divergence of populations has remained a challenging problem. Diffusion-theory–based methods can incorporate selection (Gutenkunst *et al.* 2009); however coalescent-based approaches to modeling selection do not readily admit to the kinds of flexibility required of divergence modeling (Hudson and Kaplan 1988; Neuhauser and Krone 1997; Wakeley 2008). As a result most of the model-based methods developed in recent years for studying divergence ignore the effects of selection and rely upon an assumption that neutral mutations and demography have been the sole determinants of patterns in the data (*e.g.*, Kuhner *et al.* 1998; Beerli and Felsenstein 2001; Hey and Nielsen 2007). Alternative approaches have focused on statistics that capture signatures of selection, such as elevated differentiation [*e.g.*, F_{ST}-based (Beaumont 2005)] or reduced genetic diversity and extensive linkage disequilibrium (LD), as detected by the extended haplotype homozygosity (EHH) (Sabeti *et al.* 2002), integrated haplotype score (iHS) (Voight *et al.* 2006), cross population EHH (XP-EHH) (Sabeti *et al.* 2007), and related statistics (Tang *et al.* 2007). Rather than obtaining likelihoods for the data as a function of model parameters, methods such as these can be seen as nonparametric genome scans in which a given statistic is computed for each locus or window on the

doi: 10.1534/genetics.113.149211
Manuscript received January 3, 2013; accepted for publication February 14, 2013
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.149211/-/DC1.
Data and scripts to perform the simulation study are deposited at the Dryad Repository at http://dx.doi.org/10.5061/dryad.vr7bb.
[1]Corresponding authors: Department of Genetics, Rutgers, The State University of New Jersey, 604 Allison Rd., Piscataway, NJ 08854. E-mail: hey@biology.rutgers.edu; and vitor.sousa@iee.unibe.ch
[2]Present address: Institute of Ecology and Evolution, University of Bern, Baltzerstr. 6, 3012 Bern, Switzerland.

genome, followed by the identification of genomic segments exhibiting outlier values. These statistics are thus useful to detect regions under selection, but they do not provide direct estimates for relevant parameters and are prone to several confounding factors. For instance, extreme outlier F_{ST} values in X-linked loci could be interpreted as evidence for selection, although they could simply result from the smaller effective size of the X chromosome. Model-based, likelihood approaches offer the key advantages, over genome scans based on summary statistics, that they can include diverse processes affecting the divergence of populations (including inheritance differences of autosomes and X chromosomes), and they provide for likelihood-ratio tests of alternative models.

It is possible to study selection using neutral coalescent models, albeit indirectly, through the imprint that selection can have on gene genealogies at linked neutral sites, including effects on the levels of gene flow (Petry 1983; Barton and Bengtsson 1986; Charlesworth *et al.* 1997; Rieseberg 2001; Fusco and Uyenoyama 2011) and on effective population sizes (Galtier *et al.* 2000; Charlesworth 2009; Gossmann *et al.* 2011). The principle is that when two or more populations diverge in the presence of gene flow, some genes may become differentiated due to natural selection acting differently in the two populations, whereas other genes may move freely between populations. Natural selection can be a strong barrier to gene flow, but if hybrids and backcross hybrids are not inviable or sterile, then selection has its greatest effects near those genomic regions associated with population-specific adaptations (Orr 1996; Barton 2001; Wu 2001; Pinho and Hey 2010). Similar barriers to gene flow may occur upon secondary contact of populations that had previously fixed different alleles for loci involved in genetic incompatibilities (Rieseberg 2001; Navarro and Barton 2003a; Nachman and Payseur 2012). Both scenarios lead to variation in the amount of gene flow along the genome, as regions under selection show signatures of reduced migration (Charlesworth *et al.* 1997; Navarro and Barton 2003b; Butlin 2005; Pinho and Hey 2010). In recent years the list of cases of apparent divergence in the face of gene flow has steadily grown (*e.g.*, Won *et al.* 2005; Geraldes *et al.* 2006; Hey 2006; Kronforst *et al.* 2006; Teeter *et al.* 2008; Nadachowska and Babik 2009; Carneiro *et al.* 2009, 2010; Smadja and Butlin 2011). Also, events such as selective sweeps can be seen as a population bottleneck specific to some loci (Galtier *et al.* 2000), and hence selection can lead to variation in the effective sizes along the genome (Gossmann *et al.* 2011).

Isolation-with-migration (IM) models have come to play an important role in the study of population divergence because they have the potential to capture key components of demographic history, including effective population sizes and gene flow between populations, as well as the time of separation from ancestral populations (Nielsen and Wakeley 2001; Hey and Machado 2003; Hey and Nielsen 2004). Most applications assume that all sampled loci have shared in the same demographic process. However, one way to allow for the study of demographic history, when loci have experienced varying rates of gene flow because of natural selection, is to allow each locus to have its own gene flow parameters (Won *et al.* 2005). This approach has also been considered in a hierarchical modeling framework, using approximate Bayesian computation (ABC) (Bazin *et al.* 2010). In principle this approach allows not only for the fitting of a neutral demographic model to data with a history of gene flow shaped by natural selection, but also the identification of those loci that are near genes that have experienced reduced gene flow due to natural selection (*i.e.*, those loci with low values for gene flow estimates). This method has been used, for example, in the study of divergence among *Heliconius* butterflies (Bull *et al.* 2006) and European rabbits (Geraldes *et al.* 2006). In those studies, it was possible to detect genes with evidence of no gene flow, despite moderate levels of migration in other genes, suggesting that selection had an important role in divergence. However, a drawback of this approach is the high dimensionality of the parameter space as the number of loci increases, which leads to wide confidence intervals for some individual parameters as well as limited options for testing of models.

In effect, most approaches to modeling divergence lie at the ends of two extremes, either with all loci assumed to share the same parameters or with each locus characterized by its own parameters. A different approach is to allow each locus to fall into one of a small number of groups of loci, with each group associated with a set of migration rates and/or effective population sizes. With a small number of groups of loci such an approach would offer the benefit of having a parameter set that is tractable in size. However, with multiple groups of loci the question arises, Which loci fall into which group? The assignment of loci to groups could be fixed by the investigator; however, a more general approach would allow both the estimation of which loci fall into which groups as well as the estimation of the parameters associated with each group. Here we develop this idea and describe its implementation and testing. The performance of the method was examined using simulated data, and we applied the method to the study of the divergence of two subspecies of European rabbits (*Oryctolagus cuniculus*).

Model

We focus on the basic two-population isolation-with-migration model with six demographic parameters: three effective population sizes for the three populations (for populations 1, 2, and ancestral), two migration rates (one for each direction), and a time at which the ancestral population separated into the two descendant populations. We distinguish the splitting time, t, from those parameters that provide for the rates of specific types of events in the coalescent process (*i.e.*, migration and population size parameters that we refer to collectively as Φ). Parameters are scaled by the mean mutation rate across loci μ, and hence the effective sizes are given by $4N_i\mu$, the migration rates by $m_{i\rightarrow j} = M_{i\rightarrow j}/\mu$, and the time of split by $t = T\mu$, where N_i is the effective size of

the ith population, $M_{i\to j}$ is the migration rate per generation between population i and j, and T is the time of split (in generations) (Hey and Nielsen 2004). In the case of multiple loci, each locus l will also have a mutation rate scalar u_l and an inheritance scalar h_l, hence modeling explicitly variation in the mutation rates and modes of inheritance across loci (Hey and Nielsen 2004). No recombination within loci and free recombination among loci are assumed.

Theoretical studies have shown that the effects of selection on linked neutral sites are well approximated by a purely neutral process with a reduction in the migration rate, proportional to the barrier to gene flow caused by selection (Petry 1983; Barton and Bengtsson 1986; Charlesworth *et al.* 1997; Navarro and Barton 2003a; Fusco and Uyenoyama 2011). Similarly, neutral loci linked to regions of the genome under directional or background selection suffer reductions in their effective sizes proportional to the selective strength (Charlesworth *et al.* 1993; Galtier *et al.* 2000; Charlesworth 2009; Gossmann *et al.* 2011). Different modes of selection can thus be modeled by altered demographic parameters. For instance, selection against gene flow, resulting from either local adaptation or genetic incompatibilities in the hybrids, would be reflected as a reduction in the migration rates. Adaptive introgression, on the other hand, would lead to increased migration rates. Likewise, genomic regions undergoing repeated selective sweeps would be seen as having a reduced effective size. Therefore, we assume that the effects of selection on linked sites can be described in terms of altered migration rates and/or effective population sizes. We consider a model where loci are classified into groups with each group having its own set of migration rate and/or effective population size parameters, thus relaxing the assumption that all loci share the same demography. In this general framework the only one of the six demographic parameters that remains shared by all loci is t.

In principle the number of groups of loci could be treated as an unknown; however, we focus on the case where the maximum number of groups K is set by the investigator and specifically on the simplest case where loci can be classified into two groups ($K = 2$) representing (1) loci with histories affected by linkage to genes under selection and (2) loci not affected by selection. It is important to appreciate that the identification of a group as having loci affected by selection depends entirely upon how the investigator interprets the parameter estimates of the different groups of loci. Here we focus on the case of selection against gene flow, and hence the group of loci with reduced migration rate estimates corresponds to loci potentially linked to sites under selection. The assignment of loci to groups is represented by an assignment vector a, where a_l is the group to which locus l belongs, $l = (1, \ldots, L)$. For instance, in a case with four loci and two groups, $a = (1, 1, 2, 2)$ indicates that loci 1 and 2 belong to group 1 and loci 2 and 3 belong to group 2. With more than one group of loci the set of migration and effective population size parameters, Φ, will include additional terms. In the above example, instead of one set of effective sizes (three parameters) and one pair of migration rates (two parameters), the model includes one set for each group, that is, two sets of effective sizes (six parameters) and two sets of migration rates (four parameters).

Given genetic data from L independent loci, sampled from each of two closely related populations or species, the goal is to obtain an estimate of the vector of locus assignments, $\hat{a}$, as well as the demographic parameters of the IM model, $\hat{\Phi}$ and $\hat{t}$. To connect the data to these unknowns we consider for locus l a genealogy, G_l, and for all loci the set of genealogies, $G = (G_l, \ldots, G_L)$, that describe the historical coancestry of the sampled sequences, including the tree topologies, as well as the times of coalescent and migration events (Hey and Nielsen 2004). As conceptualized by Felsenstein (1988) and now common practice in population genetics inference (*e.g.*, Kuhner *et al.* 1998; Beaumont 1999; Beerli and Felsenstein 1999; Hey and Nielsen 2004; Kuhner 2006), we consider the range of possible genealogies by approximating an integration over the genealogical space. Following the approach developed by Hey and Nielsen (2007) this integration provides for the posterior probability of the parameters of interest,

$$\pi(\Phi, t, a|X) = \int \pi(\Phi|G, t, a)\pi(G, t, a|X)dG, \qquad (1)$$

where $\pi(\Phi|G, t, a)$ is the conditional probability of the parameters given the genealogies, the splitting time, and the assignment, and $\pi(G, t, a|X)$ is the probability of genealogies, splitting time, and assignment given the data. Although this integral is not analytically tractable except for the very small sample sizes, as noted by Hey and Nielsen (2007), Equation 1 suggests a two-step Monte Carlo integration approximation. This works by first sampling genealogies, times of split, and assignment vectors from $\pi(G, t, a|X)$, which are then used to approximate the posterior of the demographic parameters $\pi(\Phi|X)$ in a second step (Hey and Nielsen 2007). Although this approach does not provide an estimate of the joint posterior $\pi(\Phi, t, a|X)$, it does provide estimates of the marginal posterior for a and t (first step), as well as the marginal posterior for Φ, which includes all of the rates parameters for genetic drift and gene flow (second step).

In the first step, a Markov chain Monte Carlo (MCMC) simulation is used to collect samples of $\{G, t, a\}$ from the posterior $\pi(G, t, a|X) \propto f(X|G)\pi(G|t, a)\pi(t)\pi(a)$, where $f(X|G)$ is the likelihood of the data given the genealogies, $\pi(G|t, a)$ is the prior probability of the genealogies conditional on the times of split and assignment, $\pi(t)$ is the prior of the times of split, and $\pi(a)$ is the prior of the assignment vector. The likelihood $f(X|G)$ is computed using conventional methods, such as by mapping mutations onto G in the case of the infinite-sites mutation model or by parameterizing the mutation process under a finite-sites model and

using the pruning algorithm (Felsenstein 1981a). The prior probability $\pi(G|t, a)$ is obtained by integrating over Φ (Hey and Nielsen 2007),

$$\pi(G|t,a) = \int \pi(G|\Phi,t,a)\pi(\Phi)d\Phi, \tag{2}$$

where $\pi(\Phi)$ is the prior distribution for the migration rates and effective population sizes, and $\pi(G|\Phi, t, a)$ is the probability of the genealogies conditional on the parameters and assignment. The calculation of this last term, $\pi(G|\Phi, t, a)$, is based on coalescent theory (Hey and Nielsen 2007; Hey 2010; Sousa *et al.* 2011) and is actually a fairly tractable function of quantities determined from G, including for each rate component of Φ (1) a count of the number of events across G that the rate pertains to and (2) a sum of the total rate for that parameter across G (see, *e.g.*, the appendix to Hey 2010). So too is the solution to the integration in Equation 2 analytical and straightforward. The sample of $\{G, t, a\}$ values can be used directly to estimate the marginal posterior distributions for t and a. Thus, this first step approximates the marginal posterior $\pi(t, a|X)$, providing estimates for the times of split and assignment of loci into groups.

The second step consists of using the sample of $\{G, t, a\}$ values to estimate the marginal posterior for Φ. Applying Bayes' theorem, the conditional probability of the parameters given the genealogies can be simplified to $\pi(\Phi|G, t, a) = \pi(G|\Phi, t, a)\pi(\Phi)/\pi(G|t, a)$. Given a sample of n genealogies, times of split and assignment from the posterior, $(G^{(i)}, t^{(i)}, a^{(i)}) \sim \pi(G, t, a|X)$ $(i = 1, \ldots, n)$, we estimate the marginal posterior distribution of the drift and migration parameters as

$$\pi(\Phi|X) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(G^{(i)}|\Phi, t^{(i)}, a^{(i)})\pi(\Phi)}{\pi(G^{(i)}|t^{(i)}, a^{(i)})}. \tag{3}$$

Note that the marginal posterior $\pi(\Phi|X)$ is not conditioned on particular values for t or a, but is in effect estimated by integrating over these other parameters. In sum, given that the joint posterior $\pi(\Phi, t, a|X)$ can be expressed by Equation 1, we can apply the above two-step procedure to obtain the marginal posterior distributions $\pi(t, a|X)$ (first step) and $\pi(\Phi|X)$ (second step) and hence estimate all the parameters of interest, including t, a, and Φ.

Inference Framework

The method described above is a general one for estimating marginal distributions for t, a, and Φ. Our aim is to map variation of demographic parameters along the genome by identifying loci sharing similar migration rates and/or effective sizes. For this reason we propose an inference scheme as follows: (1) estimation of the assignment $\hat{a}$ using the marginal posterior $\pi(a|X)$; (2) estimation of the demographic parameters for each group of loci based on the marginal posterior $\pi(\Phi|X)$; and finally (3) assessment, via likelihood-ratio tests, of the fit of alternative models to the data. We note that steps 2 and 3 can also be applied conditional on a fixed assignment, which can be useful when investigators aim to test for differences between groups given a known set of candidate loci.

We developed an MCMC simulator to generate samples from the joint posterior distribution $\pi(G, t, a|X) \propto f(X|G)\pi(G|t, a)\pi(t)\pi(a)$. In the most general case, at each iteration, we update the genealogy G to G', the times of split t to t', and the assignment a to a'. Following the Metropolis–Hastings (MH) criterion, these are accepted with probability

$$\min\left(1, \frac{f(X|G')}{f(X|G)}\frac{\pi(G'|t',a')}{\pi(G|t,a)}\frac{\pi(t')}{\pi(t)}\frac{\pi(a')}{\pi(a)} \times \frac{q[(G',t',a')\rightarrow(G,t,a)]}{q[(G,t,a)\rightarrow(G',t',a')]}\right), \tag{4}$$

where $q[(G, t, a) \rightarrow (G', t', a')]$ and $q[(G', t', a) \rightarrow (G, t, a)]$ are the proposal probabilities. The proposal for the assignment is described in detail below. For the genealogies, times of split, and mutation rate scalars, we used the same proposal schemes as in Hey and Nielsen (2007) and Hey (2010).

Assignment of loci

For generality we consider models that may have multiple groups of loci with respect to migration rates only, a_m, or with respect to only the rates of genetic drift (*i.e.*, effective population sizes), a_θ, or both. Similarly, in models with multiple sets of migration rates and effective population sizes, we can allow the assignment vectors for both types of parameters to be shared, $a = a_\theta = a_m$, or allow them to be independent of each other, $a = (a_\theta, a_m)$. Thus, one can investigate scenarios where only gene flow or effective population sizes vary among groups and scenarios where both effective sizes and migration rates differ among groups. Among the latter are models in which the assignment vector is the same for migration rates and effective population sizes as well as models in which they are free to vary from one another.

For the Markov chain simulation, the update of the assignment vector(s) proceeds by randomly picking one element of the vector and uniformly changing its label to a different group. If assignment is shared for migration and drift parameters, $a = a_\theta = a_m$, then there is only one vector to update; and if there are two independent vectors, $a = (a_\theta, a_m)$, then this update is applied to each vector separately. This update has proposal probability $q(a \rightarrow a') = L^{-1}(K - 1)^{-1}$, where K is either K_θ or K_m, depending on whether we are updating a_θ or a_m, and where K_θ and K_m are the numbers of groups for the effective sizes and migration rates, respectively. Given that when updating the assignment vectors the genealogies and times of split are not updated, the proposed assignment vector a' is accepted with probability

$$\min\left(1, \frac{\pi(G|t,a')}{\pi(G|t,a)}\frac{\pi(a')}{\pi(a)}\right), \quad (5)$$

where $\pi(a')/\pi(a)$ is the prior ratio.

We consider a uniform prior on the number of loci assigned to each group, based on the premise that it is equally likely to have any number of selected loci, from zero (no selected loci) to L. Denoting m_g as the number of loci in group g, and $\mathbf{m} = (m_1, \ldots, m_K)$ as the configuration with $m_1, \ldots, m_K$ loci belonging to groups 1 to K ($L = \sum_{i=1}^{K} m_i$), the prior of assignment can be written as $\pi(a, \mathbf{m}) = p(a|\mathbf{m})\, p(\mathbf{m})$, where $p(a|\mathbf{m})$ is the probability of an assignment vector conditional on the configuration $\mathbf{m}$, and $p(\mathbf{m})$ is its prior. We assumed that all $\mathbf{m}$ are equally likely, with probability $p(\mathbf{m}) = \binom{L+K-1}{L}^{-1}$. Conditional on $\mathbf{m}$, the probability of a givenassignment vector is $p(a|\mathbf{m}) = (K!\binom{L}{\mathbf{m}})^{-1}$. Therefore, the prior ratio simplifies to $\pi(a')/\pi(a) = \pi(a', \mathbf{m}')/\pi(a, \mathbf{m}) = (m_1'! \ldots m_K'!)/(m_1! \ldots m_K!)$.

Summarizing samples of a drawn from the posterior probability distribution: The Markov chain simulator provides a sample from the marginal posterior of assignment vectors $a^{(i)}$ ($i = 1, \ldots, n$) where n is the number of vectors sampled. However, summarizing this sample is complicated by the fact that the labels of groups are exchangeable. For example, despite having different labels, the two vectors $a^{(1)} = (1, 1, 2, 2)$ and $a^{(2)} = (2, 2, 1, 1)$ are equivalent as they both cluster loci 1 and 2 separately from loci 3 and 4. In this simple example each locus is assigned once to group g_1 and once to group g_2, resulting in a probability of $1/K$ (0.5) of belonging to each group. This clustering would thus be missed by looking at these probabilities. This equivalence, known as label switching (Lee *et al.* 2009), is a well-known property of MCMC Bayesian mixture models, such as are used in methods to assign individuals to populations (*e.g.*, Dawson and Belkhir 2001; Huelsenbeck and Andolfatto 2007). It results from the fact that if the MCMC chain is mixing properly, all the equivalent indexing assignment vectors will be sampled equally. To summarize the posterior sample in a way that accounts for the uncertainty of assignment of each locus, as well as for label switching, we considered three methods: (1) pairwise coassignment probabilities (*e.g.*, Dawson and Belkhir 2001; Onogi *et al.* 2011), (2) mean assignment based on partition distances (Huelsenbeck and Andolfatto 2007), and (3) marginal probabilities after relabeling [*e.g.*, pivotal reordering (Lee *et al.* 2009)]. With the first approach it is possible to identify, using Bayes factors, those pairs of loci for which there is strong evidence that both loci belong to the same group. However, coassignment probabilities, while not sensitive to label switching, do not help in assessing whether a locus belongs to a given group. For this we use the mean assignment, which summarizes the posterior sample, albeit without associated probabilities; and we assess the marginal posterior probability of assignment of each locus after relabeling with respect to a reference assignment. We explain these approaches in detail below.

Pairwise coassignment probabilities: For each pair of loci l and j ($l \neq j$), the posterior coassignment probability, $\pi(a_l = a_j|X)$, is given by the proportion of sampled vectors in which the two loci are classified into the same group, irrespective of the actual label (Dawson and Belkhir 2001). Following Huelsenbeck and Andolfatto (2007), we used the Bayes factor (BF) as a measure of the evidence provided by the data that the two loci belong to the same group,

$$\mathrm{BF}_{a_l=a_j} = \frac{\pi(a_l = a_j|X)/(1 - \pi(a_l = a_j|X))}{\pi(a_l = a_j)/(1 - \pi(a_l = a_j))}, \quad (6)$$

where $\pi(a_l = a_j|X)$ is the posterior, estimated as the number of assignment vectors $n_{l=j}$ in which the two loci were classified into the same group, and $\pi(a_l = a_j)$ is the prior. The prior probability of coassignment, $\pi(a_l = a_j)$ is found by summing over all possible configurations $\mathbf{m}$, $\pi(a_j = a_l) = \sum_{\mathbf{m}} p(a_j = a_l|\mathbf{m})\mathbf{p}(\mathbf{m})$, where $p(a_j = a_l|\mathbf{m}) = L^{-1}(L-1)^{-1}(m_1(m_1 - 1) + \ldots + m_K(m_K - 1))$. With two groups of loci $\pi_{K=2}(a_j = a_l) = 2/3$, and with three groups of loci $\pi_{K=3}(a_j = a_l) = 1/2$. Given that we are using a sample of n vectors to obtain the BF, we used the estimator of Grelaud *et al.* (2009), approximating $\pi(a_l = a_j|X) = (n_{l=j} + 1)/n$, and $(1 - \pi(a_l = a_j|X)) = (n - n_{l=j} + 1)/n$.

Mean assignment: Although it would not be meaningful to take an arithmetic average of sampled assignments, we can use the idea of "mean assignment", $\bar{a}$, which is the vector that minimizes the squared partition distance to all of the sampled vectors (Huelsenbeck and Andolfatto 2007; Choi and Hey 2011; Onogi *et al.* 2011). The partition distance is defined as the minimum number of elements that must be removed to make two assignment vectors identical, over all possible labeling of one of the vectors (Gusfield 2002). Because the minimum is taken over possible labelings, the mean assignment is not sensitive to label switching. The mean assignment computation was implemented by adapting the mean partition functions from Onogi *et al.* (2011) and Choi and Hey (2011).

Relabeling assignment vectors: Another solution to deal with label switching is to relabel each of the MCMC sampled vectors to minimize the partition distance from some reference vector (Lee *et al.* 2009). Here, we considered the mean assignment $\bar{a}$ to be the reference vector. For each sampled assignment vector, $a^{(i)}$ ($i = 1, \ldots, n$), we considered the $K!$ possible labelings and selected the one with the minimal partition distance to the mean assignment. For instance, with two groups of loci and a mean assignment $\bar{a} = (1, 0, 1)$, for the sampled vector $a = (0, 1, 0)$ the two possible labelings are $a^{(1)*} = (0, 1, 0)$ and $a^{(2)*} = (1, 0, 1)$, and the one with minimal distance is $a^{(2)*}$. After performing this operation on all sampled vectors we obtain a relabeled sample $a^{(i)*}$ ($i = 1, \ldots, n$). Then, the marginal posterior probability for locus l, $p(a_l|X)$, can be estimated by the proportion of assignments into each group. Note

that after relabeling we obtain a new set of configurations $\mathbf{m}^* = (m_1^*, \ldots, m_K^*)$ and that the configurations $\mathbf{m}^*$ are not affected by label switching, in the sense that the groups are defined by the number of elements assigned to each group; *i.e.*, $m_1^* \geq \ldots \geq m_K^*$, where m_g is the number of elements in group g. With two groups of loci g_1 is defined as the group with more elements (m_1^*).

For each locus l, the Bayes factor in favor of belonging to group g_s is

$$\mathrm{BF}_{a_l=g_s} = \frac{\pi(a_l = g_s|X)/(1 - \pi(a_l = g_s|X))}{\pi(a_l = g_s)/(1 - \pi(a_l = g_s))}, \qquad (7)$$

where $\pi(a_l = g_s|X)$ is the marginal posterior probability, and $\pi(a_l = g_s)$ is the prior probability. The prior probability is computed taking into account the relabeling operation. First, we need to consider the mean assignment $\bar{a}_{\pi(a)}$ of a random sample from the prior distribution $\pi(a, \mathbf{m})$, as it is used as the reference vector. Because we assumed a uniform prior on $\mathbf{m}$, we expect the vector with all loci classified into the same group to be the most frequent, implying that the prior mean assignment is a vector where all loci are classified into the same group, $\bar{a}_{\pi(a)} = (g_1, \ldots, g_1)$. Second, conditional on the prior mean assignment $\bar{a}_{\pi(a)}$, the prior probability that locus l belongs to group g_1 is obtained by summing over all possible configurations $\mathbf{m}^*$, $\pi(a_l = g_1) = \sum_{\mathbf{m}^*} p(a_l = g_1|\mathbf{m}^*)p(\mathbf{m}^*)$, where $p(a_l = g_1|\mathbf{m}^*) = m_{g_1}^*/\mathrm{L}$. The prior probability of $\pi(\mathbf{m}^*)$ is obtained by summing the probabilities of all the $\mathbf{m}$ that have the same configuration irrespective of the labels. With two groups of loci, the marginal prior probability is $\pi(a_l = g_1) = (3L + 4)/(4L + 4)$ if the number of loci L is even and $\pi(a_l = g_1) = (3L + 1)/4L$ if L is odd. Note that g_1 corresponds to the group with more elements in the estimated mean assignment $\bar{a}$. Identifying which group has been affected by selection depends upon the demographic estimates obtained. For selection against gene flow, the selected group is the one with lower migration rates. Thus, if g_1 corresponds to the group with lower migration rates, we replace $\pi(a_l = g_s)$ in Equation 7 by $\pi(a_l = g_1)$; otherwise, we replace it by $\pi(a_l = g_2) = 1 - \pi(a_l = g_1)$. Based on the Bayes factors of all loci, we inferred an assignment vector, denoted $\hat{a}_{\mathrm{BF}}$, that summarizes the sample of vectors from the MCMC. We classified a given locus into the selected group if its corresponding $\log_{10}(\mathrm{BF}) > 1.0$, which is considered to be strong evidence according to Jeffrey's scale (Kass and Raftery 1995). As for the Bayes factors of coassignment, we obtained the Bayes factors following the Grelaud *et al.* (2009) estimator.

It is noteworthy that it is possible for the mean assignment $\bar{a}$ and for the estimated assignment $\hat{a}_{\mathrm{BF}}$ to include only a subset of the number of groups, K, in the model. For example, it is possible that the mean assignment in a model with $K = 2$ places all loci into just one group. Thus in a general sense, $\bar{a}$ and $\hat{a}_{\mathrm{BF}}$ contain information about both the assignment of loci to groups and the number of groups supported by the data.

Estimating parameters for groups of loci

For estimation of the marginal posterior probability for the rate parameters, $\pi(\Phi|X)$, we again addressed label switching by relabeling assignment vectors and then replacing $a^{(i)}$ by $a^{*(i)}$ in Equation 3. For each group of loci we obtained a marginal posterior for the drift and migration rate parameters. Then, based on these distributions we identified either g_1 or g_2, as the group affected by selection, assuming that under our model of selection against gene flow, loci affected by selection show lower migration rates. We also investigated scenarios where loci with low migration rates also have lower effective sizes (see below).

Likelihood-ratio tests

An important question is whether a model with parameters specific for different groups of loci (the full, or alternative, model) can explain a data set better than a model for which all loci share the same demographic parameters (the nested, or null, model). This question can be addressed using a likelihood-ratio test (LRT) based on the difference between the maximum likelihoods under the full and nested models. A sufficiently large difference in likelihoods, and thus a rejection of the nested model, can be interpreted as a finding of real differences between the groups of loci and thus of the presence of a factor or factors (*e.g.*, linked loci under selection) that alter the demographic picture for a subset of loci.

We extended the LRT developed in Hey and Nielsen (2007), for IM models with different numbers of parameters, to the current problem in which there are multiple groups of loci each with respective sets of demographic parameters. Here we describe in some detail the implementation of the LRT because we noticed that the original report of the method (Hey and Nielsen 2007) is partly misleading for suggesting that the ratio of the maximum posterior probability densities converges to the ratio of maximum likelihoods. Given that uniform prior distributions are used, the posterior probability is proportional to the likelihood, and hence the set of parameters that maximizes the posterior also maximizes the likelihood. However, in general, it does not follow that the ratio of maximum likelihoods converges to the ratio of posterior distributions.

Consider the likelihoods $f_F(X|\hat{\Phi}_F)$ and $f_N(X|\hat{\Phi}_N)$, where $\hat{\Phi}_F$ and $\hat{\Phi}_N$ are parameters that maximize the likelihood for the full and nested models, respectively. Although these can be expressed with the Felsenstein equation (Felsenstein 1988), we cannot obtain these likelihoods directly. In the method of Hey and Nielsen (2007) both the parameter values that maximize the likelihood and the ratio of the likelihoods are approximated using an importance sampling approach in which the likelihood of each model is approximated using a sample of genealogies from a proposal distribution $q(G)$. In this case, the optimal distribution corresponds to the posterior of genealogies given the data and the maximum-likelihood parameters, $q(G) = \pi(G|X, \hat{\Phi})$, as it provides an exact estimate of $f(X|\hat{\Phi})$. However, although we

can obtain samples of genealogies from $\pi(G|X,\hat{\Phi})$, we cannot evaluate the importance weights analytically. Instead we use the posterior of genealogies sampled under the full model, $\pi_F(G|X)$, as our proposal distribution. This is similar to the optimum proposal in the sense that it is the posterior of genealogies given the data, with the advantage that it does not depend on the parameter values and that we can compute the importance weights. Hence, in principle, it can be used as a proposal under the full and various nested models with different parameter values. It is noteworthy that this proposal distribution depends on the prior distributions of the full model, and hence the specified priors affect the variance of the estimator for the likelihood. Given a set of n genealogies sampled from the proposal, $G^{(i)} \sim \pi_F(G|X)$ $(i = 1,\ldots,n)$, and noting that $\pi_F(G|X) = f(X|G)\pi_F(G)/f_F(X)$, the likelihood of the full model is approximated as

$$f_F(X|\Phi_F) \approx f_F(X)\frac{1}{n}\sum_{i=1}^{n}\frac{\pi_F(G^{(i)}|\Phi_F)}{\pi_F(G^{(i)})} \tag{8}$$

and the likelihood of the nested model as

$$f_N(X|\Phi_N) \approx f_F(X)\frac{1}{n}\sum_{i=1}^{n}\frac{\pi_N(G^{(i)}|\Phi_N)}{\pi_F(G^{(i)})}, \tag{9}$$

where $f_F(X)$ is the marginal likelihood of the data under the full model, $\pi_F(G)$ is the prior probability of genealogies under the full model, and $\pi_F(G|\Phi_F)$ and $\pi_F(G|\Phi_N)$ are the probabilities of the genealogies given the parameter values under the full and nested models, respectively. Although we can compute $\pi_F(G)$ (Equation 2) and $\pi_F(G^{(i)}|\Phi_F)$ and $\pi_N(G^{(i)}|\Phi_N)$ with coalescent theory, the marginal likelihood $f_F(X)$ is an unknown constant, and hence we can obtain only relative likelihoods. However, with the relative likelihoods we are still able to find the set of parameters that maximize the likelihoods and, furthermore, we are able to approximate the ratio of the maximum likelihoods of both models, as the $f_F(X)$ terms cancel out. Therefore, an estimate of the likelihood-ratio statistic $\hat{\Lambda} = \log(f_N(X|\hat{\Phi}_N)/f_F(X|\hat{\Phi}_F))$ can be obtained as

$$\hat{\Lambda} = \log\left(\sum_{i=1}^{n}\frac{\pi_N(G^{(i)}|\hat{\Phi}_N)}{\pi_F(G^{(i)})}\right) - \log\left(\sum_{i=1}^{n}\frac{\pi_F(G^{(i)}|\hat{\Phi}_F)}{\pi_F(G^{(i)})}\right). \tag{10}$$

In short, this particular case of importance sampling makes it possible to conduct LRTs for a variety of nested models, using only samples drawn from the posterior for the full model. Under certain regularity conditions and for unbounded parameters the test statistic -2Λ converges to a χ^2-distribution with d d.f., where d is the difference in the dimensionality (number of parameters) of the two models (Hey and Nielsen 2007). This is the case for the typical model comparisons in an IM model.

However, these conditions may not hold for models with groups of loci. The reason is that when treating the assignment a as a parameter, our model can be seen as a mixture. It is known that mixture models do not follow the conditions required for the likelihood ratios to converge to a χ^2-distribution (*e.g.*, Mendell *et al.* 1991; Lo *et al.* 2001; Hall and Stewart 2005; Garel 2007). This arises as a result of nonidentifiability when finding the maximum likelihood and defining the nested models, especially when the full model allows both the mixture proportions and the parameters to vary. In those cases the likelihood-ratio test statistic may not converge or may converge to complex distributions (Garel 2007; Azaïs *et al.* 2009). To minimize these issues we considered two alternatives in which the assignment is not included in the maximization of the likelihood: (1) compute the likelihood ratios conditional on a given fixed assignment of loci a_f, $f_N(X|\hat{\Phi}_N, a_f)/f_F(X|\hat{\Phi}_F, a_f)$; and (2) compute the likelihoods integrating over the assignment, *i.e.*, $f(X|\hat{\Phi}) = \int f(X|\hat{\Phi}, a)\pi(a)da$, hence looking at the likelihood ratios $f_N(X|\hat{\Phi}_N)/f_F(X|\hat{\Phi}_F)$.

Materials and Methods

Simulation study

The performance of the inference framework was evaluated using data sets simulated under three different classes of models: (1) a conventional neutral model with one group of loci, with all loci sharing migration and effective population size parameters; (2) a neutral model with two groups of loci with different migration rates and/or effective sizes; and (3) a model with selection against gene flow affecting a subset of the loci.

The first set of simulations allows us to quantify the false positive rate of detecting multiple groups of loci when in fact all loci are affected by only a single demographic process. With the second set of simulations we can assess the power of the method to classify loci into two groups, as well as the accuracy of the migration rates and effective size estimates of each group. Finally, the third set of simulations allows us to examine to what extent selection against gene flow is detectable and quantifiable with our inference framework.

Data sets were simulated under a two-population isolation-with-migration model, assuming the infinite-sites mutation model, with 15 gene copies sampled from each population, with 10 loci and $\theta_1 = \theta_2 = \theta_A = 10$, $t = 5$. We used the SIMDIV program (Wang and Hey 2010) to generate all the data sets for the first two cases and a Python program implementing a forward simulation for the third case (see details below). For each of the three simulation cases we applied our three-step inference procedure in which we (1) estimated the assignment, (2) estimated the demographic parameters, and (3) conducted the likelihood-ratio tests. Each analysis requires a particular full model to be specified, and we considered two different such models: a "migration only" (MO) model in which a_θ was set to include only a single group of loci, while a_m was free to vary with a maximum of two groups of loci, and a "migration and population size" (MAPS) model in which the assignment

Table 1 Classification of loci under the neutral model with a single group of loci

		Partition distance		Proportion rejection, null model		Mean *P*-value	
	$2NM$	Mean assignment[a]	Bayes factor[b]	Drift[c]	Mig[d]	Drift[c]	Mig[d]
MO model	0.25	0.10	0.00	—	0.00	—	0.7454
	2.50	0.05	0.00	—	0.10	—	0.2518
MAPS model	0.25	0.95	0.15	0.05	0.05	0.5476	0.5910
	2.50	0.85	0.30	0.10	0.05	0.4128	0.3377

Mean partition distance between the true assignment and the estimated assignment based on the mean assignment and based on Bayes factors, proportion of runs in which the null model was rejected, and mean *P*-values for rejecting the null model. Each cell corresponds to the mean among 20 runs.

[a] Classification based on the mean assignment.

[b] Classification based on Bayes factors >1.0 (log10 scale) of belonging to the group with lower migration rates.

[c] Nested model: no differences between the effective sizes among groups (chi-square d.f. = 3).

[d] Nested model: no differences between the migration rates among groups (chi-square d.f. = 2). Mig, migration.

vector for both types of parameters ($a = a_\theta = a_\mathrm{m}$) was free to vary with a maximum of two groups of loci. Data simulated under the conventional neutral model (*i.e.*, case 1) were analyzed under both of these models; while data simulated under case 2 were analyzed under the model that is true for those data, *i.e.*, either the MO or the MAPS model; and data simulated under case 3 were analyzed under the MO model, *i.e.*, the model that would approximate a scenario of selection against gene flow.

Neutral model with one group of loci: We investigated the performance of the method as a function of the migration rate (scaled by the mutation rate) for the neutral loci, ranging from $m_{1\to2} = m_{2\to1} = 0.05$ to $m_{1\to2} = m_{2\to1} = 0.50$, which corresponds to $2NM$ values ranging from 0.25 to 5.00. A total of 20 simulations were performed under each combination of parameters considered (see Table 1), which were analyzed under the MO and MAPS models.

Neutral model with two groups of loci: We used these simulations to examine scenarios where a subset of loci has altered demographics, as if caused by linkage to other selected loci: (1) two groups of loci for migration and a single group of loci for effective population size parameters (*i.e.*, all loci share the same effective population sizes, but they may have different migration rates) and (2) two groups of loci for both migration and effective population sizes (*i.e.*, each locus may fall into one of two groups, with the assignment being shared for migration rates and effective population sizes, $a = a_\theta = a_\mathrm{m}$).

For the first case, loci in group g_1 were simulated with migration rates varying from $m = 0.05$ to $m = 0.5$ ($2NM$ values of 0.25 and 2.5, respectively), and loci in group g_2 were simulated with a migration rate of zero. This approximates a scenario where a group of loci is affected by selection against gene flow due to linkage with genomic incompatibilities in the hybrids or to genes undergoing divergent selection. A migration rate of zero for group g_2 would mimic strong selection against gene flow such that the probability of sampling a migrant allele approaches zero. Data sets were analyzed under the MO model, *i.e.*, the correct model for this case.

For the second case, loci in group g_2 were simulated with a migration rate of zero and also with half the effective size of the other loci ($\theta_1 = \theta_2 = \theta_A = 5$). This approximates a scenario where a group of loci are affected through linkage by divergent selection in different populations (reducing gene flow), accompanied by selective sweeps within each population (reducing the effective sizes) (Charlesworth *et al.* 1997; Galtier *et al.* 2000). Data sets were analyzed under the MAPS model, allowing for variation in migration rates and effective sizes among groups of loci. We investigated the effect of increasing the number of loci in group g_2 as well as the effect of the migration rate of loci in group g_1 (see Table 2 for combinations of parameters tested).

Model of selection against gene flow: In this model an ancestral haploid population of size $2N$ evolves for T_a generations, after which it splits into two populations of size $2N$ that continue to exchange migrants for T generations at a symmetric and constant rate M. Selection against gene flow is modeled by partitioning each of the nonancestral populations into two distinct gene pools: residents with fitness $w_\mathrm{r} = 1$ and immigrants with fitness $w_\mathrm{i} = 1 - s$, where s is the selective coefficient. Each generation includes mutation, migration, and reproduction, which are treated as independent stochastic processes that occur sequentially. Mutation and migration follow Poisson processes with rates $2N\mu$ and $2NM$, respectively, where μ is the mutation rate (per loci per generation) and M is the migration rate per generation. Only one mutation is allowed at each site (infinite-sites model). Migrants are then randomly sampled without replacement from the source population and added to the sink population. Note that migrants from the immigrant pool of the source population are added to the resident pool of the sink population, and vice versa. Reproduction occurs via binomial sampling, according to the fitness of the resident and immigrant gene pools. Data sets were generated with an individual-based forward simulator that precisely implements this model. We tested the program by comparing simulated data with theoretical expectations and with data simulated under ms (Hudson 2002) (see Supporting Information, Figure S1). All simulations were performed assuming a population size of $2N = 1000$ and a mutation

Table 2 Classification of loci under the true models with two groups of loci

		Partition distance		Proportion rejection, null model		Mean *P*-value	
	$2NM$	Mean assignment[a]	Bayes factor[b]	Drift[c]	Mig[d]	Drift[c]	Mig[d]
MO model							
1 locus, g_2	0.25	1.50	1.35	—	0.00	—	0.5020
	1.00	2.00	1.10	—	0.35	—	0.1815
	2.50	0.60	0.75	—	0.80	—	0.0275
	5.00	0.70	0.45	—	0.90	—	0.0405
3 loci, g_2	2.50	0.15	0.05	—	1.00	—	0.0002
5 loci, g_2	0.25	4.80	4.35	—	0.10	—	0.4691
	1.00	1.35	1.20	—	0.85	—	0.0346
	2.50	0.20	0.05	—	1.00	—	0.0000
MAPS model							
1 locus, g_2	1.00	1.85	1.15	0.00	0.15	0.3900	0.1924
	2.50	1.00	0.80	0.40	0.40	0.0880	0.1343
5 loci, g_2	1.00	0.75	0.45	0.15	0.90	0.3697	0.0284
	2.50	0.20	0.05	0.45	1.00	0.0966	0.0006

Mean partition distance between the true assignment and the estimated assignment based on the mean assignment and based on Bayes factors, proportion of runs in which the nested model was rejected, and mean *P*-values for rejecting the nested model. Each cell corresponds to the mean among 20 runs.

[a] Classification based on the mean assignment.

[b] Classification based on Bayes factors >1.0 (log10 scale) of belonging to the group with lower migration rates.

[c] Nested model: no differences between the effective sizes among groups (chi-square d.f. = 3).

[d] Nested model: no differences between the migration rates among groups (chi-square d.f. = 2).

rate per locus (50-kb sequence) of $\mu = 0.005$ ($\theta = 4N\mu = 10$). The ancestral population evolved for $T_a = 5000$ generations, followed by $T = 1000$ generations after the split ($t = T\mu = 5.0$). We examined the effects of varying the neutral migration rates, selection coefficients, and number of loci under selection (see Table 3 for combinations of parameters tested). Three migration rates were considered: $M = 0.00025$, $M = 0.001$, and $M = 0.0025$ that correspond to $m = M/\mu = 0.05$ ($2NM = 0.25$), $m = 0.2$ ($2NM = 1.0$), and $m = 0.50$ ($2NM = 2.5$), respectively. We considered three selection coefficients for loci under selection: $s = 0.01$, $s = 0.05$, and $s = 0.10$ that correspond to $2Ns = 10$, $2Ns = 50$, and $2Ns = 100$, respectively. Neutral loci were simulated by setting $s = 0.0$.

Likelihood-ratio tests: In addition to these analyses we also assessed the performance of likelihood-ratio tests by comparing the distribution of the test statistic, for cases when the null model is true, to a χ^2-distribution. Groups of 50 data sets were generated according to the null model and then for each the maximum likelihood was found under both a model with two groups of loci (the alternative model) and a model with one group of loci (the null, or nested, model) and the difference used to calculate $-2\hat{\Lambda}$. The LRTs were performed either on the space of the migration rate parameters or on the space of the population size parameters, as the efficiency of the optimization algorithm to find the maximum likelihood was reduced when considering jointly all parameters. The search of the drift parameters that maximize the likelihood was done by setting a lower bound of 1.0, instead of zero, as the estimated posterior of the ancestral population tended to be quite flat but to have a high variance at values close to zero (resulting in local maximum). The distribution of $-2\hat{\Lambda}$ was assessed for the case of low ($m = 0.05$) and high ($m = 0.50$) migration rates for data sets of 10 loci under both the MO and the MAPS models. We performed these analyses based on the likelihood ratios irrespective of the assignment ($f_N(X|\Phi_N)/f_F(X|\Phi_F)$) and based on the likelihood ratios conditional on a given fixed assignment ($f_N(X|\Phi_N, a)/f_F(X|\Phi_F, a)$). For the likelihood-ratio test conditional on the assignment, we fixed a to be a random vector with 5 loci belonging to each group. We assessed whether the empirical $-2\hat{\Lambda}$ distribution converged to a χ^2 with the number of degrees of freedom equal to the difference between the number of parameters in the full and nested models, as expected for bounded parameters under regularity conditions (Chernoff 1954). Given the uncertainty of whether our models follow the required conditions, we also compared $-2\hat{\Lambda}$ to a χ^2 with an extra degree of freedom.

Data analysis: All parameter estimations and likelihood-ratio tests were carried out using a modified version of the IMa2 program, which implements a Markov chain simulation under an IM model (Hey 2010). We used uniform prior distributions for demographic parameters, $\Theta \sim U[0, 30]$, $t \sim U[0, 15]$, and $m \sim U[0, 0.75]$, where U denotes uniform distribution, Θ the effective size parameters, and m the migration rate parameters. In the analyses of the convergence of the empirical distribution of $-2\hat{\Lambda}$ to the χ^2 with $m = 0.05$ we used a narrower prior for the migration rates, $m \sim U[0, 0.15]$. For the data sets simulated with one selected locus and neutral migration rate $m = 1.0$, we used a wider prior limit of $m \sim U[0, 1.50]$. The MCMC runs began with a burn-in period ranging from 10^5 to 10^6 steps, followed by a sampling period of $5 \times 10^5 - 2 \times 10^6$ steps. A total of

Table 3 Classification of loci under the IM model with selection against gene flow

			Partition distance			
	2NM	*2Ns*	Mean assignment[a]	Bayes factor[b]	Proportion rejection, null model[c]	Mean *P*-value[c]
1 selected locus	0.25	10	1.20	1.20	0.00	0.6065
		100	1.40	1.40	0.10	0.5713
	2.50	10	1.00	1.00	0.00	0.2648
		50	0.80	1.00	0.40	0.1023
		100	0.60	0.70	0.60	0.0753
3 selected loci	0.25	50	2.30	1.40	0.70	0.0528
		100	2.20	1.50	0.60	0.0829
	2.50	10	3.00	3.00	0.40	0.1851
		50	1.10	1.00	0.80	0.0166
		100	0.30	0.30	1.00	0.0002
5 selected loci	0.25	10	4.70	4.20	0.10	0.4911
		100	4.50	3.50	0.20	0.4715
	1.00	50	2.30	3.20	0.50	0.1069
		100	1.90	2.20	0.80	0.0576
	2.5	10	5.00	5.00	0.00	0.3260
		50	1.00	1.40	0.90	0.0126
		100	0.10	0.50	1.00	0.0005

Mean partition distance between the true assignment and the estimated assignment based on the mean assignment and based on Bayes factors, proportion of runs in which the nested model was rejected, and mean *P*-values for rejecting the nested model. Each cell corresponds to the mean among 10 runs.

[a] Classification based on the mean assignment.

[b] Classification based on Bayes factors >1.0 (log10 scale) of belonging to the group with lower migration rates.

[c] Nested model: no differences between the migration rates among groups (chi-square d.f. = 2).

5000−10,000 assignment vectors were sampled for each run, together with t and G, every 30–100 steps after the burn-in.

Application to European rabbit data

We applied the new method to the study of two subspecies of European rabbit, *O. cuniculus cuniculus* and *O. cuniculus algirus*, that occur in parapatry in the Iberian Peninsula and that are thought to have diverged ~2 million years ago (Branco *et al.* 2000; Carneiro *et al.* 2009). These two subspecies exhibit contrasting patterns of differentiation at multiple loci, resulting in a bimodal distribution of differentiation (Geraldes *et al.* 2008). While the majority of loci show low differentiation consistent with high levels of gene flow, some loci exhibit high levels of differentiation, suggestive of little or no gene flow. The latter group of loci includes the mitochondrial (mt)DNA (Branco *et al.* 2000); the Y chromosome (Geraldes *et al.* 2008); and loci near the centromeres of both the X chromosome (Geraldes *et al.* 2006) and autosomes 8, 13, and 14 (Carneiro *et al.* 2009). We analyzed a data set of 44 loci ranging in length from 421 to 815 bp, sampled from primarily intronic regions of the genome, from multiple locations for each subspecies (for details see Carneiro *et al.* 2010). For each subspecies samples from multiple locations were pooled, yielding total sample sizes of 10 individuals for *O. c. algirus*, and 12 individuals for *O. c. cuniculus*. Inheritance scalars were set to 1.0 for autosomal loci and to 3/4 for X-linked loci, with the assumption that the sexes occur in equal numbers with equal variances in reproductive success (see Table S1 for details).

Genetic differentiation between the two subspecies was estimated using the F_{ST} estimator of Hudson *et al.* (1992) as implemented in the program SITES (Hey and Wakeley 1997). The mutation rate per locus per generation was estimated for each locus based on the net nucleotide differences D_A (Nei 1987) of *Oryctolagus* and *Lepus*, assuming a divergence time of $T = 11.8$ MYA (Matthee *et al.* 2004), as in Carneiro *et al.* (2010).

We began with an analysis using the conventional neutral model in which all loci shared the same effective sizes and migration rates. The prior distributions on parameters were as follows: $\Theta \sim U[0, 12]$, $m \sim U[0, 3]$, and $t \sim U[0, 6]$. Three independent runs were performed, with each providing a sample of 10,000 genealogies. Convergence was achieved with runs with 100–160 Metropolis-coupled chains (Geyer 1991) after a few million steps ($1 - 10 \times 10^6$ steps).

We then applied the inference framework described above under the MO model with two groups of migration rate parameters and one group of effective population sizes. We used a pooled sample of 100,000 genealogies, population split times and assignment vectors from 10 independent MCMC runs, each with an initial random assignment, a burn-in of 10^6 steps, and a sample size of 10,000 collected over the course of 10^6 steps of the Markov chain simulation. In the first step we estimated the mean assignment $\bar{a}$, the coassignment probabilities for pairs of loci, and marginal posteriors of assignment for each locus. In the second step we estimated the marginal posteriors for the demographic parameters of each group. And in the third step, we used likelihood-ratio tests to ask whether the MO model with different groups of loci for migration fitted the data better than models where loci share some or all of the same migration parameters. Three nested models were considered, including a model in which there is only a single migration rate in each direction shared by all loci

(*i.e.*, there is only one group of loci) and two models in which there are two groups of loci, but for one of the migration directions all of the loci share the same migration rate, while in the other direction the migration rates of the two groups of loci can vary. These latter two models differ only in the direction in which all loci share the same migration rate.

When the likelihood-ratio tests indicated the presence of two groups of loci, we identified the group affected by selection as the one with reduced migration rates. To account for the assignment uncertainty in the identification of the selected group, we evaluated the Bayes factors for the coassignment of pairs of loci ($\hat{a}_{\mathrm{BF}_{\mathrm{coassign}}}$) and Bayes factors for the marginal assignment of each locus $\hat{a}_{\mathrm{BF}}$ in the selected group. We took a "conservative assignment", by classifying loci into the selected group only if the $\log_{10}(\mathrm{BF}) > 1.0$.

As a further test to verify whether the loci showing the higher Bayes factors were involved in differences between the migration rates among groups of loci, we performed an extra MCMC run to estimate the demographic parameters of each group of loci, conditioning on the estimated assignment $\hat{a}_{\mathrm{BF}}$. In that case five independent MCMC runs were performed with a burn-in of 2×10^5 iterations, saving a total of 100,000 genealogies. The priors were the same as used to estimate the assignment. Finally, we also performed the same likelihood-ratio tests described above.

Results

To test the performance of the new methods we analyzed simulated data sets that were generated under classical neutral models, models with differences in the migration rates and drift parameters between groups of loci, and models with loci under selection against gene flow. For each data set we recorded (1) the estimated mean assignment $\bar{a}$ and the estimated assignment based on Bayes factors of marginal assignment $\hat{a}_{\mathrm{BF}}$, (2) the marginal posterior distribution for parameters of the model (effective sizes, migration rates, and times of split), and (3) the significance (*P*-values) of the likelihood-ratio tests.

Assignment of loci

For each data set we used the partition distance to quantify the difference between the estimated assignment ($\bar{a}$ and $\hat{a}_{\mathrm{BF}}$) and the true assignment. The partition distance takes values ranging from zero, if all loci are correctly classified, to $L/2$ if half of the loci are incorrectly classified. Tables 1–3 show the mean partition distance (among 10–20 runs) and the proportion of runs where the null model was rejected in the likelihood-ratio tests for the different scenarios considered.

The data sets simulated under the null model with a single group of loci had mean partitions close to zero, suggesting that the method is correctly classifying all loci into a single group, even though data were analyzed under a model with multiple groups (Table 1). Also as expected if the method is working correctly, the proportion of runs in which the null model was rejected in the likelihood-ratio tests is close to 0.05.

For data sets simulated under the true alternative model, including loci simulated with zero gene flow (group g_2), we found that the mean partition distances decrease toward zero when increasing the migration rate of loci in group g_1 and increasing the number of selected loci (Table 2). Overall, the classification based on the Bayes factors tended to return lower partition distances, suggesting that taking into account the uncertainty of assignment improves the classification. For most data sets generated with a migration rate of $m = 0.05$, corresponding to $2NM = 0.25$, all loci were classified into a single group and the null model was not rejected. This suggests that it is difficult to detect groups of loci when migration rates in both groups are low. In contrast, for higher differences in the migration rates between the two groups, *i.e.*, $m = 0.5$ ($2NM = 2.5$) and $m = 1.0$ ($2NM = 5$) for group g_1 and $m = 0.0$ for group g_2, the null model was rejected for at least 80% of the data sets, and the mean partition distances were close to zero, indicating that loci were correctly classified. With $m = 0.2$ ($2NM = 1$), the two groups were distinguishable in most runs with five loci in group g_2 but not with a single locus (Table 2). Overall, these results fit the expected pattern that it is easier to correctly classify loci when the difference between the migration rates of the two groups is higher. Also, it suggests that increasing the number of loci in each group increases the accuracy of the classification. Similar results were obtained under the MO and MAPS models (Tables 1 and 2).

For the data sets simulated with loci under selection against gene flow, the partition distance decreases toward zero for higher selective coefficients and higher neutral migration rates (Table 3). This suggests that the higher the neutral migration rates and the selection coefficients are, the easier it is to detect loci under selection. Note that the results with $2Ns = 100$ approximate those obtained for neutral scenarios simulated with zero migration rate (Table 2), which mimics very strong selection ($2Ns \approx \infty$). For cases where the neutral migration rate was low ($2NM = 0.25$), even with a high selective coefficient of $2Ns = 100$, all loci tended to be assigned into a single group without rejecting the null model, showing that it is difficult to detect loci under selection in those situations (Table 3).

Estimation of demographic parameters

Figure 1 shows the distribution of the modes of the marginal posteriors obtained for each demographic parameter under the different combination of parameters considered for simulations with two groups of loci under the MO model. As can be seen, there is no apparent bias in the estimation of the effective sizes (Figure 1, A–C). Most runs exhibited modes close to the true value (dashed horizontal line), with the exception of the ancestral effective sizes, for which there was a larger variance, in agreement with the fact that it is more difficult to estimate this parameter (Figure 1C). The times of split were reasonably well estimated, with the exception of the scenarios with a single group of loci and with one locus in group g_2 and higher migration rates (*i.e.*,

scenarios S0m0.5, S1m0.5, and S1m1.0, Figure 1D). This is probably because of the combination of high gene flow and relatively old separation time between the two populations, which decreases the information in the data about the time of split. The migration rates were also reasonably well estimated (Figure 1, E–H). For each migration direction ($m_{1\rightarrow 2}$ and $m_{2\rightarrow 1}$), we show the estimates for the group g_1 (m_{g_1}) in Figure 1, E and G and for the group g_2 (m_{g_2}) in Figure 1, F and H. We defined g_1 and g_2 as corresponding to the groups with higher migration rate estimates and lower migration rate estimates, respectively. Note that the true migration rates for group g_1 varied between 0.05 and 1.0, depending on the scenario examined, whereas the true value for g_2 was zero for all runs simulated with one, three, and five loci in group g_2 (S1, S3, and S5). Despite the high variance, the group g_1 modes exhibited medians (Figure 1, solid line within each box) close to the true values, ranging from 0.014 to 0.048, from 0.10 to 0.25, and from 0.45 to 0.75, for scenarios with $m = 0.05$, $m = 0.2$, and $m = 0.5$, respectively. The modes of group g_2 showed a reduced variance, with most runs with modes at zero, especially for runs with three and five loci in that group (scenarios S3 and S5) and the ones with higher migration rate and a single locus (scenarios S1m0.5 and S1m1.0), with medians ranging from 0.000 to 0.015.

For the simulations analyzed under the MAPS model, there were not only two sets of migration rate parameters (one for each group), but also two sets of effective population sizes. The modes of the posteriors of the effective sizes were close to the true values, *i.e.*, 10.0 for loci in group g_1 and 5.0 for loci in group g_2, especially for θ_1 and θ_2 (see Figure S2, A–D). As expected, for the data sets simulated with a single group of loci, with no differences between groups (scenario S0m0.5), both sets of modes were close to 10.0. Similar results were found with a single locus in group g_2 and $m = 0.2$ (Table 2). Note that the modes of the ancestral population size exhibited a higher variance and appear biased toward values close to zero for the scenarios with five loci with zero migration and half effective sizes (S5m0.2 and S5m0.5). For the times of split and migration rates the estimates were similar to the ones obtained under the MO model. Overall, these results suggest that most parameters are reasonably well estimated and that the differences between the two groups are clearer with five selected loci.

The marginal posteriors showed a high density close to the true parameters, which is reflected in narrow credible intervals, as seen in Figure 2. Figure 2 shows the sum of the posterior distributions of runs where the mean assignment was correctly inferred, under three scenarios with five selected loci with increasing selective coefficients ($2Ns = 50$ and $2Ns = 100$) and with five loci simulated with $m = 0.0$ (*i.e.*, $2Ns \approx \infty$). The posteriors for the migration rate of group g_1 were flatter than for group g_2, suggesting that it is difficult to precisely estimate the gene flow rates in g_1, which were relatively high. However, it is noteworthy that the variance of the distributions of g_1 in models with two groups of loci is similar to the one obtained with a single group of loci, suggesting that the shape of the posteriors and the apparent high uncertainty are not due to the existence of two groups. The posteriors for the migration of group g_2 loci showed distributions concentrated around low migration values and increasingly close to zero for higher selective coefficients. Actually, we found that the mean of the peaks of the posterior distributions for group g_2 decreased with increasing the selective coefficient (Figure 3A). Similarly, the difference between the peaks of the migration rate estimates for groups g_1 and g_2 increased almost linearly as a function of the selective coefficients (Figure 3B), in agreement with the expectation that the stronger the selection pressure is, the higher the reduction in the effective migration rate (Petry 1983; Barton and Bengtsson 1986; Fusco and Uyenoyama 2011).

Figure 4 shows one example of the posterior densities obtained for runs in which the mean assignment was incorrect. In these cases, the estimates for all the parameters still showed high densities close to the true parameter values, but the posteriors for the migration rates of the two groups overlapped considerably (Figure 4, C and D). This is in agreement with the results of the likelihood-ratio tests (Tables 2 and 3), indicating that it is harder to distinguish loci in the two groups with limited neutral gene flow ($m = 0.05$).

Likelihood-ratio tests

We examined the empirical distribution of $-2\hat{\Lambda}$ by analyzing data generated according to the null model, *i.e.*, with a single group of loci. We assessed the correspondence between the likelihood-ratio test statistic $-2\hat{\Lambda}$ and the χ^2-distributions for both the ratio of the likelihoods integrating over assignment, $f_N(X|\Phi_N)/f_F(X|\Phi_F)$, and the ratio of likelihoods conditional on a random fixed assignment, $f_N(X|\Phi_N, a)/f_F(X|\Phi_F, a)$. For the MO model we tested a full model comprising two groups of loci with specific migration rates (four migration rate parameters) against a nested model where migration rates of both groups of loci were identical, which corresponds to a model with a single group (Figure 5, A and B). The empirical distribution of the $-2\hat{\Lambda}$ statistic converges reasonably well to a χ^2-distribution with 2 d.f. when conditioning on a given fixed assignment for data sets simulated with low ($m = 0.05$) and high ($m = 0.5$) migration rates. However, when performing the likelihood ratio integrating over the assignment, the distribution was shifted to the left of the $\chi^2_{\text{d.f.}=2}$ curve for the lower migration rate (Figure 5A), indicating that using a $\chi^2_{\text{d.f.}=2}$ to obtain the *P*-values would result in a conservative test. In contrast, for the higher migration rate (Figure 5B), the distribution was slightly shifted to the right and seemed to fit a χ^2-distribution with 3 rather than 2 d.f. In any case, based on the $\chi^2_{\text{d.f.}=2}$ and at a statistical level of 0.05, the observed proportion of rejections of the null model ranged from 0.00 to 0.04, close to the expected value of 0.05.

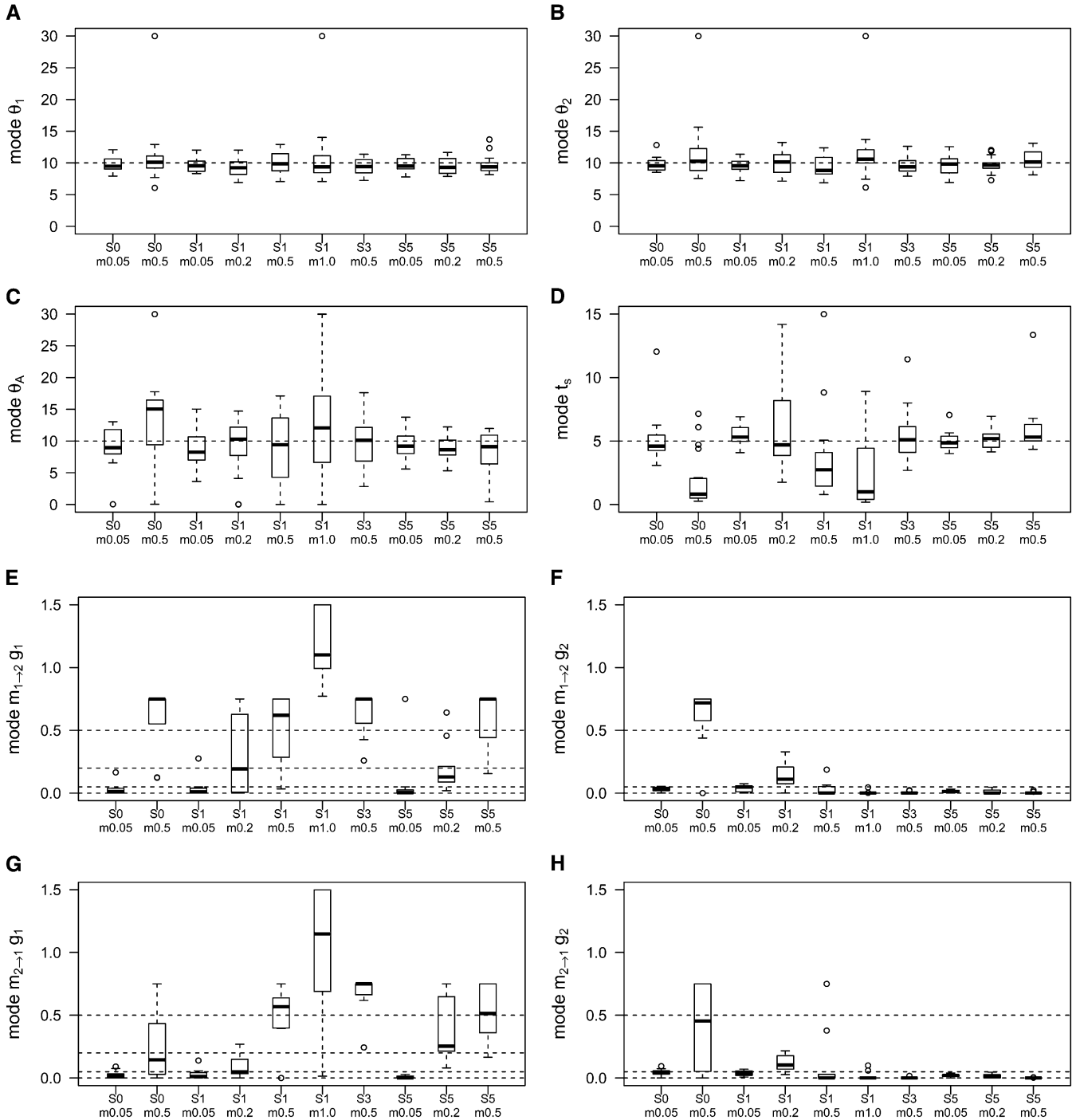

Figure 1 (A–H) Distribution of the posterior modes for the demographic parameters of each scenario considered, under the "migration only" (MO) model. The simulated scenarios are coded as follows: (1) S0, S1, S3, and S5 correspond to cases with zero, one, three, and five loci in group g_2 ($m = 0.0$), respectively; and (2) m values (*i.e.*, m0.05, m0.2, m0.5, and m1.0) are the migration rates for loci in group g_1. Horizontal dashed lines correspond to the true parameter values used to simulate the data. For scenarios S1, S3, and S5 the loci in group g_2 were simulated with a migration rate of zero to mimic the effects of strong selection against gene flow.

For the MAPS model we tested a full model with two groups of loci for the effective sizes (six population size parameters) against a nested model where effective sizes of both groups were identical (Figure 5, C and D). Again, the empirical distribution of the $-2\hat{\Lambda}$ statistic fitted reasonably well the χ^2-distribution with 3 d.f., with a proportion of rejection of the nested model ranging from 0.02 to 0.04. The exception was the distribution obtained for data sets simulated with higher migration rates ($m = 0.5$) integrating over the assignment. In this case, the curve was slightly shifted to the right. Using the $\chi^2_{\text{d.f.}=3}$, the nested model was rejected in 0.12 of times at a significant level of $\alpha = 0.05$,

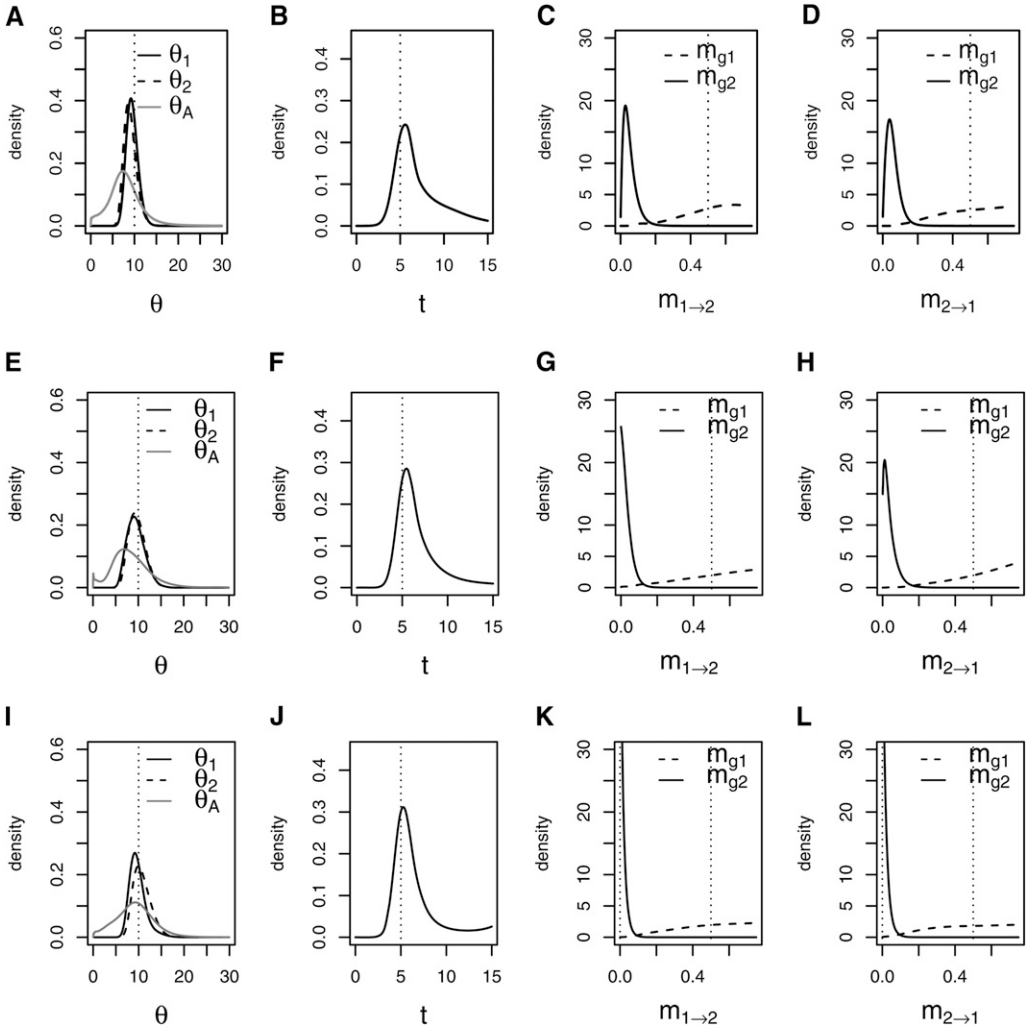

Figure 2 Marginal posterior distribution of demographic parameters under the MO model. Densities were obtained by summing the posteriors of the runs where the mean assignment corresponded to the correct classification of loci, under three scenarios with neutral migration rate of $m = 0.5$ ($2NM = 2.5$). (A–D) Five selected loci with $2Ns = 50$; (E–H) five selected loci with $2Ns = 100$; (I–L) five loci in group g_2 simulated with $m = 0.0$ ($2Ns \approx \infty$).

resulting in a higher rate of false positives. Note that this curve seems to fit a χ^2-distribution with 4 d.f., suggesting that a more conservative test can be performed by comparing the obtained statistic with the χ^2-distribution with an extra degree of freedom.

Overall, these results suggest that the χ^2-distributions can be used to obtain approximate *P*-values for the likelihood-ratio tests, as previously shown for IMa2 with simpler models without groups of loci (Hey and Nielsen 2007).

Application to European rabbits

Assignment of loci into groups: Figure 6 shows the values of F_{ST} for each locus, together with the assignment of loci estimated with our method under the MO model. Figure 6B shows the mean assignment (represented by solid and open bars) and the uncertainty of assignment, quantified as the BFs for being classified into group 2 (g_2), which corresponds to the group exhibiting lower migration rates. Note that using the mean assignment, all loci with posterior assignment probabilities >0.50 are classified into group g_2. In contrast, the BFs give more weight to loci with posteriors close to 1.0, accounting for the prior of assignment. Given that the prior probability is close to 0.75, loci with posterior probabilities ranging from 0.50 to 0.75 are assigned to group g_2 according to the mean assignment, but have negative BFs. Information about the loci can be found in Table S1.

Although there is a high variance among F_{ST} values, there is a clear correspondence between F_{ST} values and the corresponding BFs for assignment into g_2 (Figure 6). The BFs of the majority of loci were >1.00 or <-1.00, suggesting high support for assignment into group g_2 or group g_1, respectively. However, some loci, such as 28, 33, 8, and 24, exhibited BFs close to zero, reflecting a high uncertainty in their classification. We took a conservative approach by classifying loci into group g_2 only if their BFs were >1. All other loci were classified into group g_1, corresponding to the group exhibiting higher migration rates (neutral loci). Of the 44 loci, 12 were considered to be potentially linked to sites under selection based on the strong support of belonging to group g_2 (log10(BF) > 1.0): loci 25, 12, 6, 19, 4, 23, 13, 15, 18, 17, 36, and 43. Ten of these loci were found in the X chromosome and 2 in the autosomes (chromosomes 13 and 14).

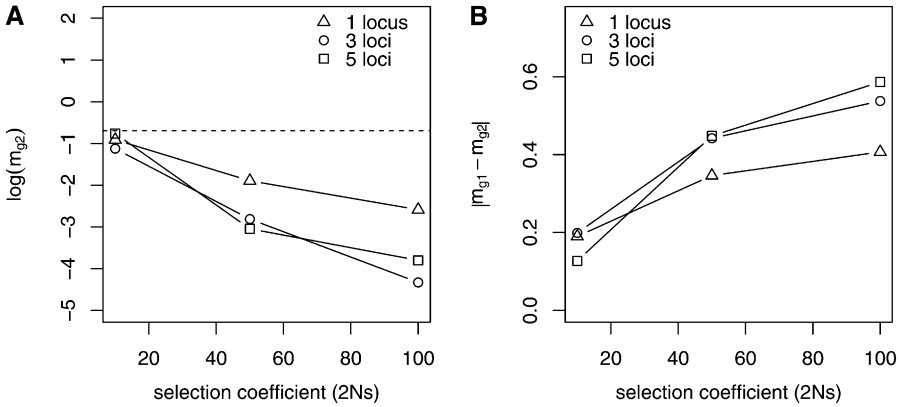

Figure 3 Posterior distribution of migration rates as a function of the selective coefficients. (A) Mean of the modes of the posterior for migration rates of loci in group g_2 (in log scale), defined as the group with lower rates, as a function of the selected coefficients for scenarios with one, three, and five selected loci. (B) Mean difference between the posterior modes of migration rates of groups g_1 and g_2, as a function of the selective coefficients for scenarios with one, three, and five selected loci.

As a further test, we examined whether the same loci were grouped together based on the pairwise coassignment probabilities (see Figure S3). These results also support the presence of two groups, with the same 12 loci grouped together. Thus, using two different approaches to summarize the posterior of assignment we found the same set of loci assigned to group g_2.

Estimates of migration rates: The marginal posterior density estimates obtained under a model with only one group of loci suggest symmetric gene flow, with similar rates in both directions ($m_{1\to2} \approx m_{2\to1} \approx 0.37$, Figure 7). However, when analyzed assuming two groups of loci, there is a clear difference between the posterior curves for group g_1 and those for group g_2 in the migration rate from *O. c. cuniculus* to *O. c. algirus*, with little overlap of the two distributions (Figure 8A). Loci in group g_1 are estimated to have high gene flow, with a 95% high posterior density (HPD) between 0.49 and 2.50 and a peak at 1.85 (95% HPD: 0.49–2.50); whereas loci in group g_2 have a peak at zero (95% HPD: 0.00–0.06). Note that we consider the peak of the posterior as a point estimate and the 95% HPD as an estimate of the credible intervals. In the other direction, from *O. c. algirus* to *O. c. cuniculus*, the estimates for groups g_1 and g_2 overlap. The group g_1 posterior has a nonzero peak at 0.18 (95% HPD: 0.00–1.08), whereas the group g_2 has a peak at zero (95% HPD: 0.00–0.07), both consistent with limited gene flow (Figure 8B). Note that these results were obtained conditional on a fixed assignment of loci, $\pi(\Phi|X, \hat{a}_{\mathrm{BF}})$, assuming that the 12 loci with BFs > 1.0 (Figure 6B) belong to group g_2 and the remaining loci to group g_1. Similar qualitative results were found when looking at the marginal posteriors, integrating over the assignment $\pi(\Phi|X)$ (see Figure S4). In that case, for the migration rate from *O. c. cuniculus* to *O. c. algirus* we obtained estimates of 2.30 (95% HPD: 0.52–2.99) and 0.07 (95% HPD: 0.02–0.16), and for the migration rate from *O. c. algirus* to *O. c. cuniculus* we obtained estimates of 0.64 (95% HPD: 0.13–2.72) and 0.06 (95% HPD: 0.01–0.16), for groups g_1 and g_2, respectively.

Estimates of effective sizes and times of split: The posterior distributions for the effective sizes suggest that *O. c. algirus* have a larger effective size than *O. c. cuniculus*. The posterior estimates obtained under the assumption that all loci share the same demography were ~4.2 (95% HPD: 3.4–5.3) and 2.8 (95% HPD: 2.2–3.5), respectively (Figure 7). These were similar to the marginal posteriors obtained when assuming two groups of loci, integrating over assignment (Figure 9A), with estimates of 4.1 (95% HPD: 3.3–5.1) and 2.6 (95% HPD: 2.0–3.3), respectively. The posteriors of the ancestral population size indicated a population size similar to that of *O. c. cuniculus* (Figure 9A), with a peak of 2.4 (95% HPD: 1.3–3.7) that is similar to the value of 2.6 (95% HPD: 1.8–3.5) found under a model with a single group of loci. The posteriors conditional on the assignment $\hat{a}_{\mathrm{BF}}$, with 12 loci classified into group g_2, were also similar (see Figure S5), with modes of 3.6 (95% HPD: 2.9–4.7), 2.8 (95% HPD: 2.2–3.4), and 2.3 (95% HPD: 0.1–8.1) for the effective sizes of populations 1, 2, and ancestral, respectively. Assuming a geometric mean mutation rate of 1.18×10^{-6} (estimated based on nucleotide differences between *Oryctolagus* and *Lepus* for the loci in this study), these posterior modes point to effective sizes on the order of hundreds of thousands (800,000, 600,000 and 500,000, respectively).

The posteriors for the time of split were affected by the model assumptions. When assuming equal migration rates for all loci (one group of loci), the posterior had a peak of 1.24 (95% HPD: 0.91–1.66), suggesting a split ~1.05 MYA (Figure 7). In contrast, the marginal posterior for the time of split obtained with the MO model with two groups of loci showed a high density between 1.14 and 2.72 (95% HPD, Figure 9B), with a peak of 1.60, which suggests a split ~1.35 MYA. This is close to the estimate of 1.57 (95% HPD: 1.26–5.30) found for the posterior conditional on the assignment $\hat{a}_{\mathrm{BF}}$, suggesting a split ~1.33 MYA. Taken together, these results are in agreement with previous estimates of divergence of the two subspecies ~1.0–2.0 MYA (Branco *et al.* 2000; Carneiro *et al.* 2009).

Likelihood-ratio tests: We used likelihood-ratio tests to examine whether the rabbit data could be explained by simpler models. When comparing the likelihoods calculated by integrating over the assignment, the nested model with identical migration rates among groups was strongly rejected

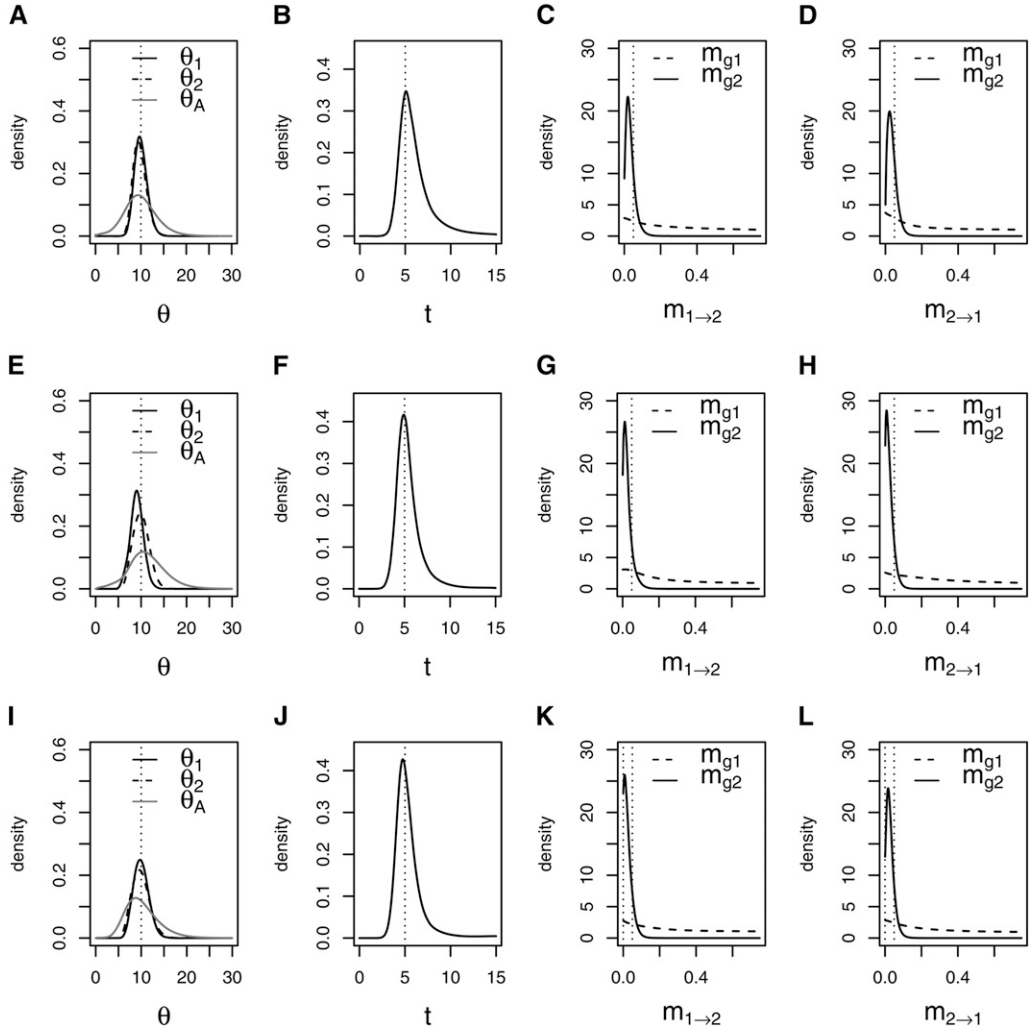

Figure 4 Marginal posterior distribution of demographic parameters obtained with the MO model for runs where the estimated mean assignment was incorrect, *i.e.*, loci were not classified into the correct groups. Densities were obtained by summing the posteriors under the case with neutral migration rate of $m = 0.05$ ($2NM = 0.25$). (A–D) Five selected loci with $2Ns = 10$; (E–H) five selected loci with $2Ns = 100$; (I–L) five loci in group g_2 simulated with $m = 0.0$ ($2Ns \approx \infty$).

(P-value $= 9.18 \times 10^{-26}$). Table 4 shows the results of tests for likelihoods calculated by conditioning on the estimated assignment $\hat{a}_{\mathrm{BF}}$. All nested models were rejected, except the model where the migration rate $m_{2\to1}$ from *O. c. algirus* to *O. c. cuniculus* was identical in both groups of loci ($m_{2\to1_{g_1}} = m_{2\to1_{g_2}}$). Thus, the major barrier to gene flow in loci of group g_2 appears to occur from *O. c. cuniculus* to *O. c. algirus*, but not in the opposite direction. This is in agreement with the posteriors for the migration rates (Figure 8), suggesting that the effects of genetic incompatibilities on backcrossed hybrid fitness depend on the genetic background of the parental population.

Discussion

When new species arise while gene exchange is occurring, there must be some kind of selective barrier to the movement of genes, associated with the adaptations of the respective incipient species (Bush 1975; Endler 1977; Felsenstein 1981b; Templeton 1981; Rice 1984; Rieseberg 2001; Butlin 2005; Pinho and Hey 2010). Whether selection is associated with reduced hybrid fitness or with other environmental factors that differentiate the habitats of the incipient species, reduced gene flow is expected over those portions of the genome linked to loci that are the selective targets (Petry 1983; Barton and Bengtsson 1986; Charlesworth *et al.* 1997).

To study these selective processes within a larger demographic framework we extend the now classic approach of assuming altered neutral model processes for loci linked to those under selection (*e.g.*, Hudson and Kaplan 1988; Hey 1991; Charlesworth *et al.* 1993, 1995, 1997; Slatkin 1995; Gillespie 2001). Our method clusters loci into distinct groups characterized by different sets of parameters and is applicable to a wide range of biological questions. By grouping loci the method allows for the analysis of data sets with many loci, some of which may be linked to sites under selection, without introducing a very large number of parameters (*e.g.*, as would be the case if each locus had a set of demographic parameters). This approach helps to serve two important goals. First, it makes possible estimates of the neutral demographic parameters even when the data set

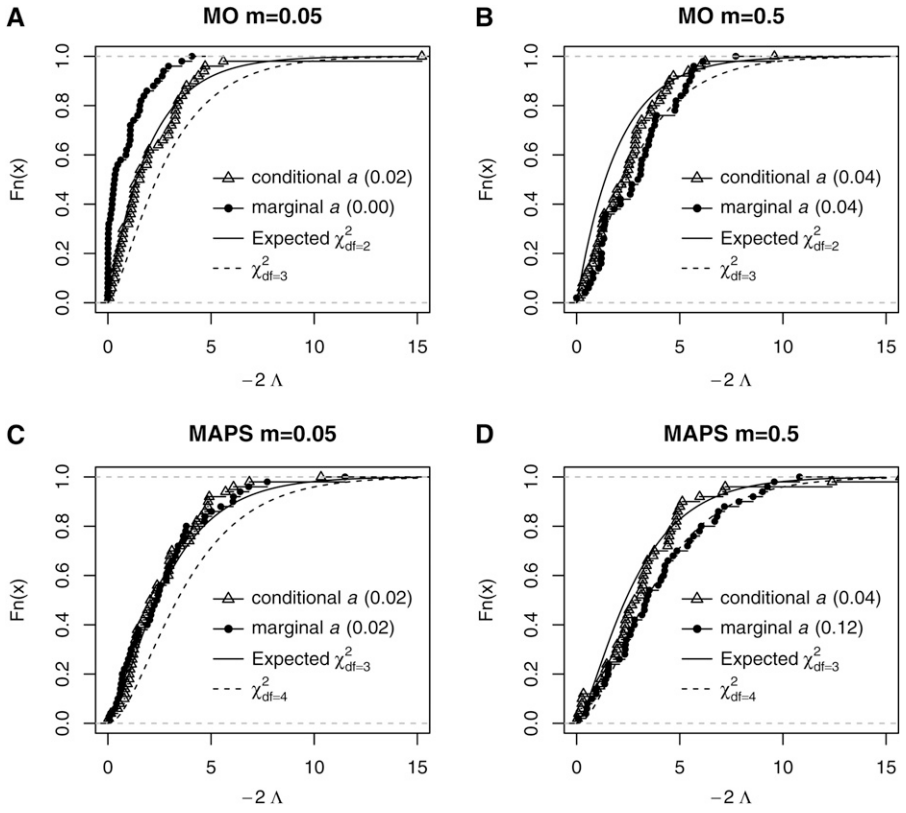

Figure 5 Fit of the estimated likelihood-ratio test statistic ($-2\hat{\Lambda}$) to the expected χ^2-distributions (solid line). Shown are the ratio of likelihoods integrating over assignment, $f_F(X|\Theta)/f_N(X|\Theta)$ ("marginal *a*"), and the ratio of likelihoods conditional on a fixed random assignment, $f_F(X|\Theta, a)/f_N(X|\Theta, a)$ ("conditional *a*"). (A) MO model with low migration rate $m = 0.05$. Shown is comparison of the full model with four migration rate parameters, with the nested model with two parameters where migration rates are equal in both groups; *i.e.*, $m_{0\to1} = m_{0\to1_{g_1}} = m_{0\to1_{g_2}}$ and $m_{1\to0} = m_{0\to1_{g_1}} = m_{0\to1_{g_2}}$. (B) Same as A but with higher migration rate of $m = 0.50$. (C) MAPS model with low migration rate $m = 0.05$. Shown is comparison of the full model with six effective size parameters, with the nested model with four parameters where effective sizes are equal in both groups; *i.e.*, $\theta_1 = \theta_{1_{g_1}} = \theta_{1_{g_2}}$, $\theta_2 = \theta_{2_{g_1}} = \theta_{2_{g_2}}$, and $\theta_A = \theta_{A_{g_1}} = \theta_{A_{g_2}}$. (D) same as C but with higher migration rate $m = 0.50$. The values within parentheses are the proportion of times the nested model was rejected (expected value of 0.05 for a significance level of $\alpha = 0.05$). Dashed line corresponds to a χ^2-distribution with one extra degree of freedom. Each empirical distribution was obtained by analyzing 50 data sets, simulated under the null model (all loci share the same parameters) with an effective size of $\theta = 10$ for all populations, time of split $t = 5$, and migration rate of $m = 0.05$ or $m = 0.50$.

contains loci affected by selection through linkage. Second, it allows the identification of those genomic regions linked to genes under selection. Moreover, the MCMC sampler of assignment vectors described here allows classification and likelihood-ratio tests without the need to specify *a priori* a list of candidate loci. Furthermore, it is possible to obtain estimates of demographic parameters, integrating over the uncertainty of the assignment of loci, as well as testing for differences between groups given a fixed assignment.

Throughout this article we have assumed that differences between the drift and/or migration rates among groups of loci are caused by the action of natural selection. However, it should be clear that we do not take selection into account explicitly, but rather consider only its indirect effects on the migration rates and effective sizes. This approach is explicit in part of our simulation study, in which data for loci in the selected group were generated using a neutral model simulation with zero gene flow and half the effective sizes of other loci. It is important to recognize that there may be other processes leading to similar genetic signatures. For instance, our basic isolation-with-migration model assumes that migration rates and effective sizes were constant since the time of the split. It is possible that in some species, gene flow and/or the effective sizes may have changed through time in ways not accommodated by a basic isolation-with-migration model and they therefore introduce additional variation in genealogies among loci. For instance, the method of Li and Durbin (2011) uses the variation of the time to the most recent common ancestor (TMRCA) among linked loci to infer changes in population sizes. Although the IM MCMC sampler that was used here has been shown to be fairly robust to deviations from the model assumptions (Strasburg and Rieseberg 2010), these aspects would need to be further investigated to quantify its effects on the false positive rate of loci under selection.

We also emphasize that because selection is not directly a part of the analysis, the actual labeling of which group of loci is the "selected" group and the interpretation of the mode of selection are up to the investigator with the aid of the estimates obtained from the analysis. In this light it is important to recognize that reduced gene flow is not necessarily the only consequence of selection during divergence. It is possible in principle for some loci to have higher gene flow because of selection, as in cases of spread of beneficial alleles from one population to another. Also, shared balanced polymorphisms at some loci would lead to estimates of high gene flow levels in some loci. Likewise, selective sweeps, local adaptation, and background selection can all lead to estimates of reduced effective sizes, but regions under balancing selection could exhibit higher effective sizes.

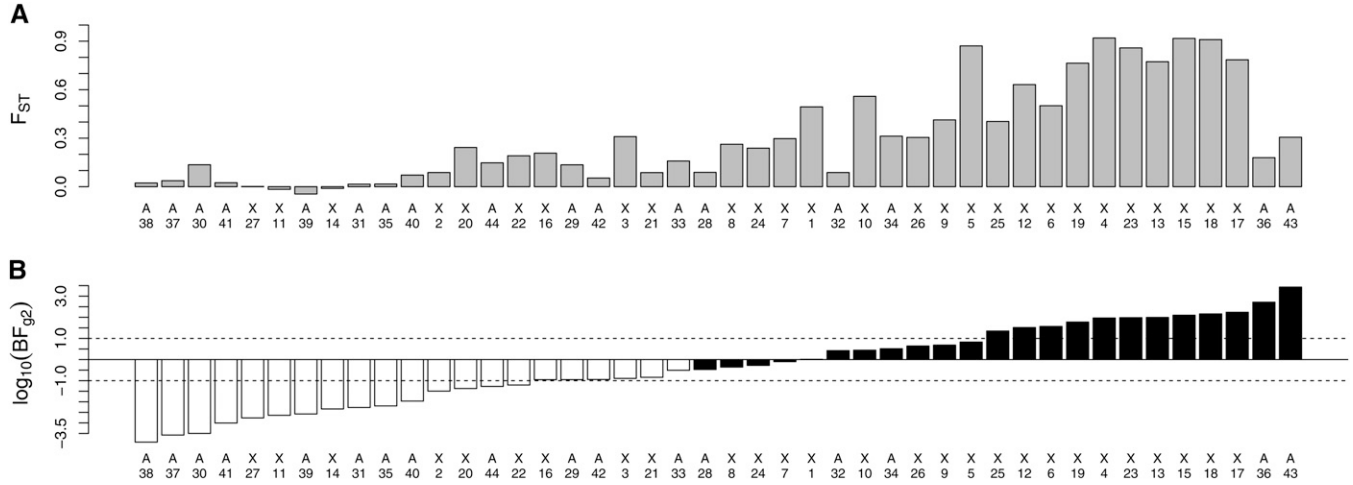

Figure 6 Distribution of F_{ST} and assignment of loci into groups for European rabbit data. (A) F_{ST} estimates obtained for each locus. (B) Bayes factors (BF) for assignment of loci into group g_2 in logarithmic scale. Loci belonging to group g_2 exhibited posteriors indicating low migration rates, identifying this as the selection group. In B, solid and open bars correspond to the groups inferred according to the mean assignment. Each bar corresponds to one locus, and its height represents the BF for the hypothesis that it belongs to group g_2. Horizontal dashed lines represent log10(BF) = 1.0 and log10(BF) = −1.0, which according to Jeffrey's scale (Kass and Raftery 1995) indicate strong support for being classified into group g_2 and group g_1, respectively. Loci with log10(BF) < 1.0 were classified into group g_1 (neutral loci). Note that loci classified into group g_2 according to the mean assignment (solid bars) with probabilities between 0.50 and 0.75 will have negative BF because the prior probability of assignment is close to 0.75. The loci were ordered according to increasing BF values. See Table S1 for information on these loci. Results were obtained by pooling together 10 independent runs (100,000 assignment vectors). Runs were performed with the following priors: $\theta \sim U[0, 12]$, $m \sim U[0, 3]$, and $t \sim U[0, 6]$.

The class of models introduced here is general as it encompasses the two extreme cases of all loci sharing the same demographic parameters ($K = 1$ group) and of each locus having its own set of migration rates and effective sizes ($K = L$ groups). Here, we focused on models with two groups of loci, with the key assumption that loci classified into the same group are similar enough to be approximated with a single set of demographic parameters. This implies, in the case of loci linked to selection, that these have a similar combination of selection coefficients and degrees of linkage to actual selection targets. However, it is likely that selective coefficients and (certainly) recombination distances vary along the genome, and in some cases it may be useful to consider a larger number of groups.

Overall, the results of the simulation study suggest that the MCMC assignment sampler is able to correctly classify loci into groups, mainly when there is a considerable difference in the migration rates among groups (*e.g.*, due to higher selective pressures) and there is more than a single locus in one of the groups. Furthermore, the method had power to distinguish groups of loci with likelihood-ratio tests, as the null model was rejected in most runs with highly significant *P*-values, especially with higher migration rate differences (Tables 2 and 3). This is in agreement with general results obtained in statistical studies of mixture models, showing that power increases with the spacing between components and decreases for asymmetric proportions (*e.g.*, Mendell *et al.* 1991; Lo *et al.* 2001; Hall and Stewart 2005). Thus, in principle, the power of our inference framework to distinguish groups of loci should increase with the number of loci. A desirable property of the assignment sampler is to not detect more than one group when data have in fact been shaped only by demography. We examined this by analyzing data simulated with a single group (*i.e.*, with no selected loci), allowing us to quantify the proportion of runs with loci incorrectly classified into two groups (false positives). Our results revealed that in most runs all loci were correctly classified into one group based on the $\hat{a}_{BF}$ (40 and 33 out of 40 for MO and MAPS, respectively). Interestingly, in the few cases where loci were classified into two groups (false positives), the null models were rejected in ∼5% of times when performing the likelihood-ratio tests, as expected if demography was the only factor responsible for the genetic patterns.

Population divergence in European rabbits

Previous studies of the European rabbit subspecies have found differences in differentiation patterns among loci, suggesting that some of them could be, or could be linked to, sites of genomic incompatibilities causing partial reproductive isolation (Carneiro *et al.* 2009, 2010).

Our results confirm that the data favor two groups of loci over one, as loci were assigned into two groups with high Bayes factor support for the majority of loci and likelihood-ratio tests were highly significant. Moreover we infer, based on posterior distributions for migration rates and by likelihood-ratio tests, that gene flow between populations and among groups is asymmetric. In the direction of *O. c. cuniculus* to *O. c. algirus* there is a clear difference between the migration rates of the two groups of loci, with loci in group g_1 indicating high gene flow and loci in group g_2 indicating reduced gene flow. This suggests that some loci in our data

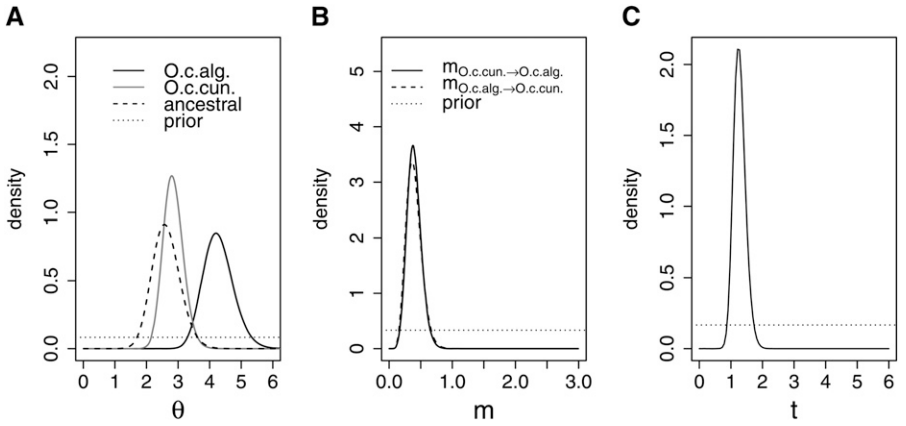

Figure 7 Posterior distributions for demographic parameters obtained by assuming a single group of loci for the European rabbit data. (A) Effective size parameters (Θ); (B) migration rates (m); (C) time of split (t). Results were obtained by pooling together three independent runs, in a total of 30,000 genealogies.

set are likely to be associated with hybrid incompatibilities only in the *O. c. algirus* background and hence prevent introgression from *O. c. cuniculus* into *O. c. algirus*. The analysis of gene flow in the other direction, from *O. c. algirus* to *O. c. cuniculus*, tells a different story. While group g_2 loci were estimated to have zero gene flow, the estimate of gene flow for group g_1 loci was much lower than in the other direction, and we could not reject a model in which there is only one single rate of (low) gene flow rate from *O. c. algirus* to *O. c. cuniculus*. One explanation is that barriers to gene flow are stronger in this direction, so that more loci are effectively linked to sites of incompatibility.

Overall, these results indicate gene flow variation among groups of loci and an asymmetry depending on the direction of migration. These are in agreement with previous studies (Carneiro *et al.* 2010) and with the expected asymmetrical nature of Dobzhansky–Muller incompatibilities under reciprocal introgression (Wu and Palopoli 1994; Gavrilets 1997; Turelli and Orr 2000). The X chromosome location of most g_2 loci is also in agreement with an enrichment of loci contributing to reproductive isolation on this chromosome (Coyne and Orr 2004), and with the expectation of several speciation models in which incompatibility loci accumulate in regions of low recombination (*e.g.*, Rieseberg 2001; Navarro and Barton 2003a). Previous estimates suggest that in European rabbits the X chromosome has on average low recombination rates (Carneiro *et al.* 2010). However, it is also the case that many of the X chromosome loci were assigned to g_1 with high estimated gene flow levels and that two autosomal loci (36 and 43) also fell within g_2. An important implication of these results is that they provide testable predictions for future studies of the rabbit hybrid zone, including controlled experimental crosses. A reasonable alternative explanation for the high proportion of loci with reduced gene flow on the X chromosome is that the X-linked loci are affected differently from autosomal loci by some demographic processes, such as sex-biased dispersal. Our model does allow for a reduced effective size for X-linked genes, and each locus has its own mutation rate scalar; however, our model does not otherwise parameterize sex-limited processes. The fact that we found loci assigned with high credibility into two groups in both the autosomes and the X chromosome argues against biased sex dispersal as the sole cause of this observation, as in that case we would expect to see all loci in the X chromosome in a separate group. Nevertheless, we performed extra analyses, looking at the autosomes and X chromosome loci independently. We found that, with the exception of one locus on the X chromosome, the same set of loci was classified with high support into group g_2 (see Figure S6). This suggests that our conclusions are not affected by jointly considering all loci under our model and that indeed we gain statistical power by jointly analyzing all loci.

Potential applications and limitations

The methods described here provide a mean for identifying loci affected by selection in the context of population divergence. In contrast with other approaches that rely upon summary statistics, the method fully accounts for differences in mutation rates and inheritance modes (autosomal, X chromosome, Y chromosome, and mtDNA) among loci, and it

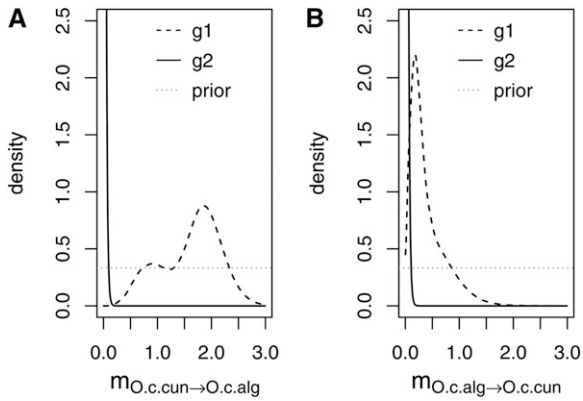

Figure 8 Posterior distributions for the migration rates in both directions for the European rabbit data, conditional on the assignment of loci based on Bayes factors ($\hat{a}_{BF}$). (A) Posterior distributions for the migration rate from *O. c. cuniculus* to *O. c. algirus*; (B) posterior distributions for migration rate from *O. c. algirus* to *O. c. cuniculus*. Results were obtained under the MO model by pooling together 10 independent runs, in a total of 100,000 genealogies.

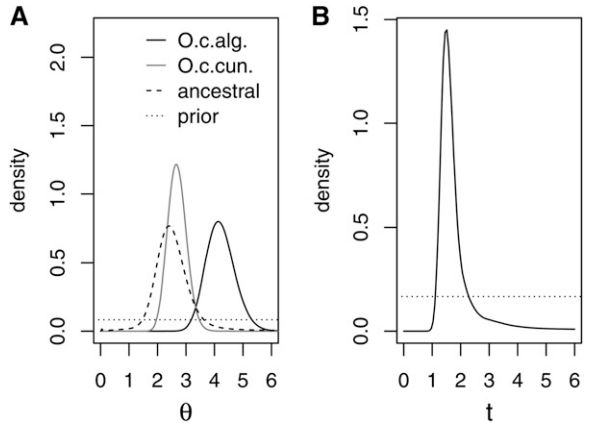

Figure 9 (A and B) Marginal posterior distributions for (A) effective size parameters (Θ) and (B) times of split (t). Results were obtained with the European rabbit data, integrating over the assignment. Results were obtained under the MO model by pooling together 10 independent runs, in a total of 100,000 genealogies.

allows for the identification of locus groups while also estimating times of population split, effective sizes, and migration rates. The general approach described here could also be extended to other families of demographic models, such as island models (*e.g.*, Beerli and Felsenstein 2001) and admixture models (*e.g.*, Chikhi *et al.* 2001).

From a practical standpoint, one of the major limitations of the current method is that it relies on a computationally intensive MCMC algorithm. For instance, for the European rabbit data set comprising 44 loci, the average computation time was 182 hr for runs with 2×10^6 iterations in a CPU with 2.80 GHz, and this is expected to grow with the number of loci. A related limitation is that the model assumes that all loci are independent, with free recombination among loci and no recombination within loci. This assumption imposes constraints on the length of the sequences that can be analyzed. In practice, large data sets (*e.g.*, genomic data) may be broken down into smaller data sets, focusing on a small number of sufficiently distant genomic regions that can be considered independent. As shown by Strasburg and Rieseberg (2010) in a simulation study, the IMa method is robust to modest levels of recombination when data are reduced to haplotype blocks that do not show evidence of recombination. In this way it may still be possible to analyze genomic data sets by focusing on a few hundred independent and nonrecombining loci distributed across the genome.

The results of the simulation study suggest that the power to detect groups of loci depends on the migration rates, the selective coefficients, and the number of loci in each group. In cases where the effective migration rates are similar for the two groups the method may fail to identify them. Note that the posterior of the assignment of loci depends not only on the information in the data (expected to increase with the number of loci), but also on the prior distribution. Thus, the apparent lack of power for cases with a single selected locus

Table 4 Likelihood-ratio tests for different nested models based on European rabbit data

Model	$-2\hat{\Lambda}$	*P*-value	d.f.	ESS
$m_{1\to 2_{g_1}} = m_{1\to 2_{g_2}}$, $m_{2\to 1_{g_1}} = m_{2\to 1_{g_2}}$	108.1	3.36×10^{-24}	2	1.00
$m_{1\to 2_{g_1}}$, $m_{1\to 2_{g_2}}$, $m_{2\to 1_{g_1}} = m_{2\to 1_{g_2}}$	2.015	0.156	1	8.07
$m_{1\to 2_{g_1}} = m_{1\to 2_{g_2}}$, $m_{2\to 1_{g_1}}$, $m_{2\to 1_{g_2}}$	28.51	9.32×10^{-8}	1	1.98

Results obtained by pooling together 10 independent runs (100,000 genealogies). Runs were performed with the following priors: $\theta \sim U[0, 12]$, $m \sim U[0, 3]$, and $t \sim U[0, 6]$. d.f., degrees of freedom; ESS, effective sample sizes. Results were obtained for likelihoods by conditioning on the assignment based on the Bayes factors ($\hat{a}_{BF}$), with 12 loci classified into group 2 ($f(X|\Phi, \hat{a}_{BF})$).

and low migration rate can be related with the reduced number of loci analyzed. In principle, increasing the number of loci increases the ability to distinguish between groups. The choice of the prior may also affect to some extent the power to detect groups. Here, we considered that all the partitions with a given number of loci in each group are equally likely (*i.e.*, a uniform distribution on the number of loci classified into each group). This prior gives the same probability to a partition where all loci are grouped together as it does, for example, to the set of all the partitions with one locus in group g_1 and the remaining loci in group g_2. For a case with a maximum of two groups of loci, the shape of this distribution is similar to the Dirichlet process prior, used by Huelsenbeck and Andolfatto (2007) to assign individuals into populations. It is straightforward to implement other priors in this inference framework. For example, more informative priors could be considered to account for the case when most loci are expected to be affected by demography and only a small subset of loci are expected to be under selection. Finally, it is also straightforward for the investigator to fix the assignment of some loci beforehand and to infer group assignment for only a subset of the loci.

In addition to estimated group assignments, which may suggest one or more than one group of loci, the method provides for likelihood-ratio tests to detect differences between the drift or migration rates among groups. However, because the data are considered to have been sampled from a mixture of $K = 2$ models, with the assignment parameter a identifying which portions correspond to which model, finding the expected distribution for the likelihood-ratio statistic is not straightforward (*e.g.*, Garel 2007; Azaïs *et al.* 2009). We attempted to circumvent this problem by not including the assignment space in the maximization of the likelihood. Instead, we either conditioned the likelihood on a fixed assignment, $f(X|\Phi, a)$, or looked at the likelihood integrating over the assignment, $f(X|\Phi)$. We thus focused on finding the maximum likelihoods over the drift or migration rate parameter space, for which the likelihood-ratio test statistic was shown to follow a χ^2-distribution (Hey and Nielsen 2007). Our simulation results show that the likelihood-ratio test statistic converges reasonably well to a χ^2-distribution with the number of degrees of freedom given by the difference in the number of parameters of the full and nested models. However, in one of the scenarios examined, the empirical distribution appeared to converge to a distribution

with more degrees of freedom (Figure 5D). Thus, to be conservative, we recommend obtaining the *P*-values based on distributions with one extra degree of freedom, especially when performing tests involving drift parameters.

The methods presented here also hold potential for informing on the selection coefficients for the loci in the group under selection. Relevant theory has shown that the expected reduction in the effective migration rate is proportional to the strength of selection and the recombination rate between the neutral locus marker and the site under selection (*e.g.*, Petry 1983; Barton and Bengtsson 1986; Fusco and Uyenoyama 2011). Our results confirm this expectation, as we found a correlation between the migration rate estimates and the selection coefficients. This suggests the possibility of translating differences in migration rates among groups of loci into selection coefficients. However, we did not include recombination in our model, and hence it remains challenging to establish a direct connection between these two quantities. Further theoretical developments that explicitly model recombination are required before selection coefficients can be inferred from differences in demographic parameters.

Acknowledgments

We thank two anonymous reviewers for their valuable comments that improved the quality of the manuscript. We thank Sang Chul Choi, Aude Grelaud, Sruti Patoori, and Vijay Ravikumar for helpful discussions and Janeen Pisciotta for programming assistance. This research was supported by grants from the National Institutes of Health (GM078204) and the National Science Foundation (DEB-0949561) and by the Portuguese Science Foundation (Fundação para a Ciência e a Tecnologia) postdoctoral grant to M.C. (SFRH/BPD/72343/2010) and project grant PTDC/BIA-EVF/111368/2009. The authors declare no conflict of interest.

Literature Cited

Arnold, M., 1997 *Natural Hybridization and Evolution*. Oxford University Press, London/New York/Oxford.

Azaïs, J., É. Gassiat, and C. Mercadier, 2009 The likelihood ratio test for general mixture models with or without structural parameter. ESAIM Probab. Stat. 13: 301–327.

Barton, N., 2001 The role of hybridization in evolution. Mol. Ecol. 10: 551–568.

Barton, N., and B. Bengtsson, 1986 The barrier to genetic exchange between hybridising populations. Heredity 56: 357–376.

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

Beaumont, M., 1999 Detecting population expansion and decline using microsatellites. Genetics 153: 2013–2029.

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

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

Beerli, P., and J. Felsenstein, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA 98: 4563–4568.

Branco, M., N. Ferrand, and M. Monnerot, 2000 Phylogeography of the European rabbit (*Oryctolagus cuniculus*) in the Iberian Peninsula inferred from RFLP analysis of the cytochrome b gene. Heredity 85: 307–317.

Bull, V., M. Beltrán, C. Jiggins, W. McMillan, E. Bermingham *et al.*, 2006 Polyphyly and gene flow between non-sibling *Heliconius* species. BMC Biol. 4: 11.

Bush, G., 1975 Modes of animal speciation. Annu. Rev. Ecol. Syst. 6: 339–364.

Butlin, R., 2005 Recombination and speciation. Mol. Ecol. 14: 2621–2635.

Carneiro, M., N. Ferrand, and M. Nachman, 2009 Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (*Oryctolagus cuniculus*). Genetics 181: 593–606.

Carneiro, M., J. Blanco-Aguiar, R. Villafuerte, N. Ferrand, and M. Nachman, 2010 Speciation in the European rabbit (*Oryctolagus cuniculus*): islands of differentiation on the X chromosome and autosomes. Evolution 64: 3443–3460.

Charlesworth, B., 2009 Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10: 195–205.

Charlesworth, B., M. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303.

Charlesworth, B., M. Nordborg, and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. 70: 155–174.

Charlesworth, D., B. Charlesworth, and M. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. Genetics 141: 1619–1632.

Chernoff, H., 1954 On the distribution of the likelihood ratio. Ann. Math. Stat. 25: 573–578.

Chikhi, L., M. Bruford, and M. Beaumont, 2001 Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. Genetics 158: 1347–1362.

Choi, S., and J. Hey, 2011 Joint inference of population assignment and demographic history. Genetics 189: 561–577.

Coyne, J. A., and H. A. Orr, 2004 *Speciation*. Sinauer Associates, Sunderland, MA.

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

Dobzhansky, T., 1951 *Genetics and the Evolution of Species*. Columbia University Press, New York.

Endler, J., 1977 *Geographic Variation, Speciation, and Clines*, Vol. 10. Princeton University Press, Princeton, NJ.

Felsenstein, J., 1981a Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17: 368–376.

Felsenstein, J., 1981b Skepticism towards Santa Rosalia, or why are there so few kinds of animals? Evolution 35: 124–138.

Felsenstein, J., 1988 Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22: 521–565.

Fusco, D., and M. Uyenoyama, 2011 Sex-specific incompatibility generates locus-specific rates of introgression between species. Genetics 189: 267–288.

Galtier, N., F. Depaulis, and N. Barton, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 155: 981–987.

Garel, B., 2007 Recent asymptotic results in testing for mixtures. Comput. Stat. Data Anal. 51: 5295–5304.

Gavrilets, S., 1997 Hybrid zones with Dobzhansky-type epistatic selection. Evolution 51: 1027–1035.

Geraldes, A., N. Ferrand, and M. Nachman, 2006 Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (*Oryctolagus cuniculus*). Genetics 173: 919–933.

Geraldes, A., M. Carneiro, M. Delibes-Mateos, R. Villafuerte, M. Nachman *et al.*, 2008 Reduced introgression of the Y chromosome between subspecies of the European rabbit (*Oryctolagus cuniculus*) in the Iberian Peninsula. Mol. Ecol. 17: 4489–4499.

Geyer, C. J., 1991 Markov chain Monte Carlo maximum likelihood, pp. 156–163 in *Computing Science and Statistics*, *Proceedings of the 23rd Symposium on the Interface*, edited by E. M. Keramidas. Interface Foundation of North America, Seattle.

Gillespie, J., 2001 Is the population size of a species relevant to its evolution? Evolution 55: 2161–2169.

Gossmann, T., M. Woolfit, and A. Eyre-Walker, 2011 Quantifying the variation in the effective population size within a genome. Genetics 189: 1389–1402.

Grelaud, A., J. M. Marin, C. Robert, F. Rodolphe, and F. Tally, 2009 Likelihood-free methods for model choice in Gibbs random fields. Bayesian Anal. 2: 427–442.

Gusfield, D., 2002 Partition-distance: a problem and class of perfect graphs arising in clustering. Inf. Process. Lett. 82: 159–164.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5: e1000695.

Hall, P., and M. Stewart, 2005 Theoretical analysis of power in a two-component normal mixture model. J. Stat. Plann. Inference 134: 158–179.

Hey, J., 1991 A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. Theor. Popul. Biol. 39: 30–48.

Hey, J., 2005 On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol. 3: 0965–0975.

Hey, J., 2006 Recent advances in assessing gene flow between diverging populations and species. Curr. Opin. Genet. Dev. 16: 592–596.

Hey, J., 2010 Isolation with migration models for more than two populations. Mol. Biol. Evol. 27: 905–920.

Hey, J., and C. Machado, 2003 The study of structured populations: new hope for a difficult and divided science. Nat. Rev. Genet. 4: 535–543.

Hey, J., and R. Nielsen, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167: 747–760.

Hey, J., and R. Nielsen, 2007 Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc. Natl. Acad. Sci. USA 104: 2785–2790.

Hey, J., and J. Wakeley, 1997 A coalescent estimator of the population recombination rate. Genetics 145: 833–846.

Hudson, R., and N. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics 120: 831–840.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Hudson, R. R., M. Slatkin, and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.

Huelsenbeck, J., and P. Andolfatto, 2007 Inference of population structure under a Dirichlet process model. Genetics 175: 1787–1802.

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

Kronforst, M., L. Young, L. Blume, and L. Gilbert, 2006 Multilocus analyses of admixture and introgression among hybridizing *Heliconius* butterflies. Evolution 60: 1254–1268.

Kuhner, M., 2006 LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22: 768–770.

Kuhner, M., J. Yamato, and J. Felsenstein, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149: 429–434.

Lee, K., K. Mengersen, J. M. Marin, and C. Robert, 2009 Bayesian inference on mixtures of distributions, pp. 165–202 in *Statistical Science and Interdisciplinary Research: Perspectives in Mathematical Sciences I*, Vol. 7, edited by N. S. Sastry, T. S. S. R. K. Rao, M. Delampady, and B. Rajeev. World Scientific, Hackensack, NJ.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

Lo, Y., N. Mendell, and D. Rubin, 2001 Testing the number of components in a normal mixture. Biometrika 88: 767–778.

Matthee, C., B. Van Vuuren, D. Bell, and T. Robinson, 2004 A molecular supermatrix of the rabbits and hares (Leporidae) allows for the identification of five intercontinental exchanges during the Miocene. Syst. Biol. 53: 433–447.

Maynard Smith, J., 1966 Sympatric speciation. Am. Nat. 100: 637–650.

Mendell, N., H. Thode, Jr., and S. Finch, 1991 The likelihood ratio test for the two-component normal mixture problem: power and sample size analysis. Biometrics 47: 1143–1148.

Nachman, M., and B. Payseur, 2012 Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos. Trans. R. Soc. B 367: 409–421.

Nadachowska, K., and W. Babik, 2009 Divergence in the face of gene flow: the case of two newts (Amphibia: Salamandridae). Mol. Biol. Evol. 26: 829–841.

Navarro, A., and N. Barton, 2003a Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. Evolution 57: 447–459.

Navarro, A., and N. Barton, 2003b Chromosomal speciation and molecular divergence—accelerated evolution in rearranged chromosomes. Science 300: 321–324.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Neuhauser, C., and S. Krone, 1997 The genealogy of samples in models with selection. Genetics 145: 519–534.

Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158: 885–896.

Onogi, A., M. Nurimoto, and M. Morita, 2011 Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. BMC Bioinformatics 12: 263.

Orr, H., 1996 Dobzhansky, Bateson, and the genetics of speciation. Genetics 144: 1331–1335.

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

Pinho, C., and J. Hey, 2010 Divergence with gene flow: models and data. Annu. Rev. Ecol. Evol. Syst. 41: 215–230.

Rice, W., 1984 Disruptive selection on habitat preference and the evolution of reproductive isolation: a simulation study. Evolution 38: 1251–1260.

Rieseberg, L., 2001 Chromosomal rearrangements and speciation. Trends Ecol. Evol. 16: 351–358.

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

Sabeti, P., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.

Slatkin, M., 1995 Hitchhiking and associative overdominance at a microsatellite locus. Mol. Biol. Evol. 12: 473–480.

Smadja, C., and R. Butlin, 2011 A framework for comparing processes of speciation in the presence of gene flow. Mol. Ecol. 20: 5123–5140.
Sousa, V., A. Grelaud, and J. Hey, 2011 On the nonidentifiability of migration time estimates in isolation with migration models. Mol. Ecol. 20: 3956–3962.
Strasburg, J., and L. Rieseberg, 2010 How robust are "Isolation with Migration" analyses to violations of the IM model? A simulation study. Mol. Biol. Evol. 27: 297–310.
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.
Teeter, K., B. Payseur, L. Harris, M. Bakewell, L. Thibodeau *et al.*, 2008 Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res. 18: 67–76.
Templeton, A., 1981 Mechanisms of speciation—a population genetic approach. Annu. Rev. Ecol. Syst. 12: 23–48.
Turelli, M., and H. Orr, 2000 Dominance, epistasis and the genetics of postzygotic isolation. Genetics 154: 1663–1679.
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., 2008 *Coalescent Theory: An Introduction*. Robert & Company, Greenwood Village, CO.
Wang, Y., and J. Hey, 2010 Estimating divergence parameters with small samples from a large number of loci. Genetics 184: 363–379.
Won, Y., A. Sivasundar, Y. Wang, and J. Hey, 2005 On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence. Proc. Natl. Acad. Sci. USA 102: 6581–6586.
Wu, C., 2001 The genic view of the process of speciation. J. Evol. Biol. 14: 851–865.
Wu, C., and M. Palopoli, 1994 Genetics of postmating reproductive isolation in animals. Annu. Rev. Genet. 28: 283–308.

Communicating editor: J. Wall

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.149211/-/DC1

Identifying Loci Under Selection Against Gene Flow in Isolation-with-Migration Models

Vitor C. Sousa, Miguel Carneiro, Nuno Ferrand, and Jody Hey

DOI: 10.1534/genetics.113.149211

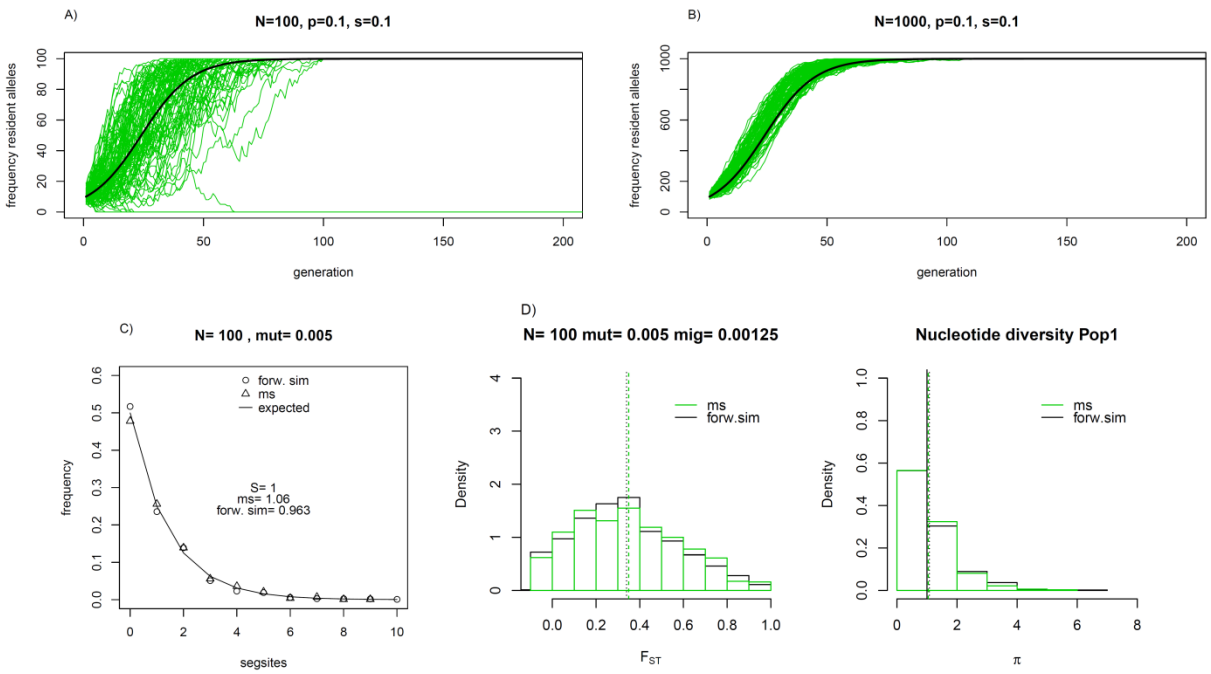

Figure S1 Comparison of simulations with theoretical expectations.
A) Trend of 1000 independent simulations of the frequency of resident alleles with fitness *w*=1+*s*, with *N*=100, *s*=0.1 and an initial frequency of *p*=0.1. Solid black line represents the expectation under a process with no drift.
B) Trend of 1000 independent simulations of the frequency of resident alleles with fitness *w*=1+*s*, with a larger effective size of *N*=1000 (*s*=0.1$ and *p*=0.1). Solid black line represents the expectation under a process with no drift.
C) Distribution of the number of pairwise differences in a single stationary population obtained with our forward simulator, *ms* and theoretical expectation (obtained as $f(k|\theta) = \int_0^\infty f(k|t_c,\theta) f(t_c)\, dt_c$, where k is the number of pairwise differences, t_c is the coalescent time, $f(k|t_c,\theta)$ is the distribution of k given t_c, i.e. Poisson with rate $\theta\ t_c$, and $f(t_c)$ is the distribution of t_c, i.e. exponential with rate 1.0). Simulations performed with *2N*=100 and $\mu = 0.005$, which corresponds to a $\theta = 1.0$. The expected number of segregating sites is *S*=1. The means obtained with ms ("ms") and our forward simulator ("forw.sim") are shown within the plot.
D) Comparison of the distribution of F_{ST} and nucleotide diversities obtained over 1000 simulations with our simulator ("forw.sim"), and with ms ("ms"), under an IM model. Vertical dotted and dashed lines correspond to the mean of the distribution.

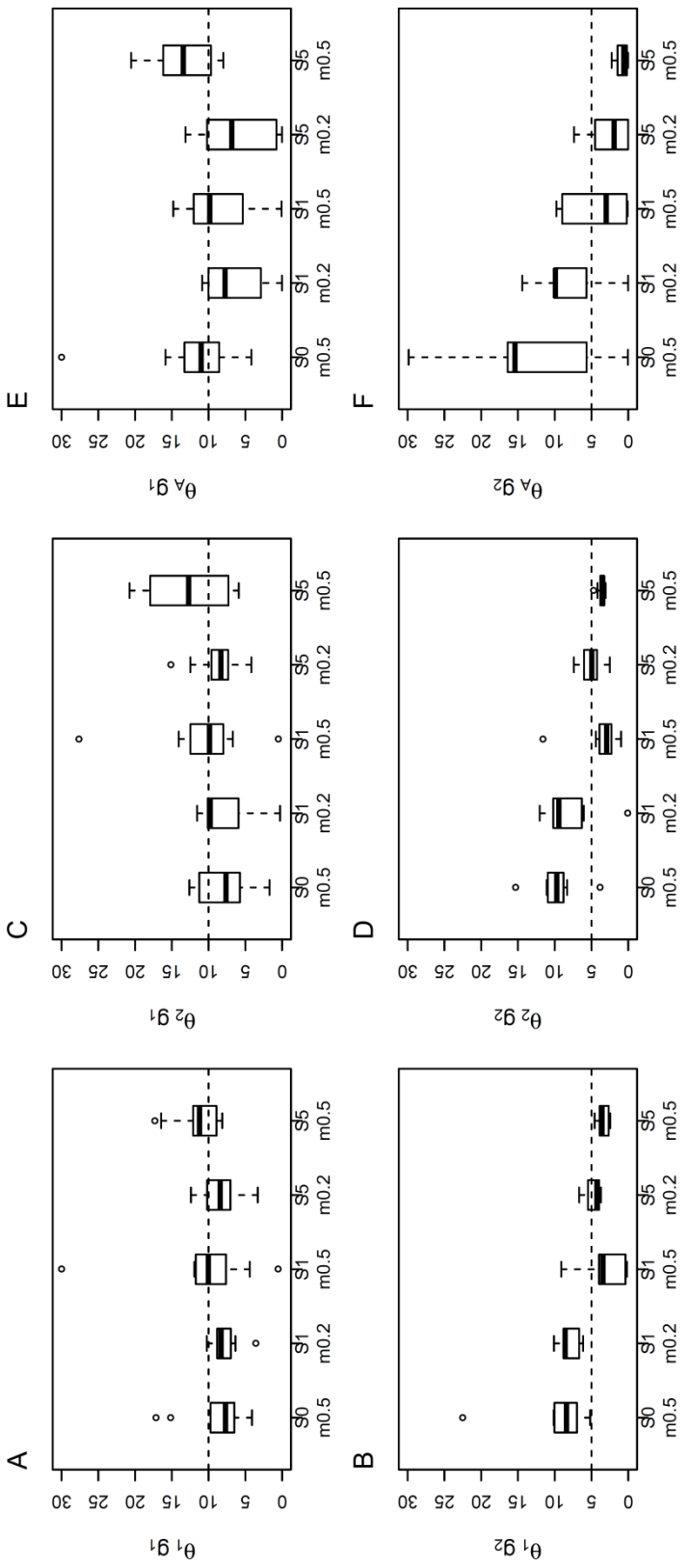

Figure S2 Distribution of the posterior modes for the demographic parameters of each scenario considered, under the "Migration and population size" (MAPS) model. The scenarios are coded as in Figure 1. Horizontal dashed lines correspond to the true parameter values used to simulate the data. Loci in group g_2 were simulated with a migration rate of zero, and with half the effective sizes of loci in group one, to mimic the effects of selection against gene flow.

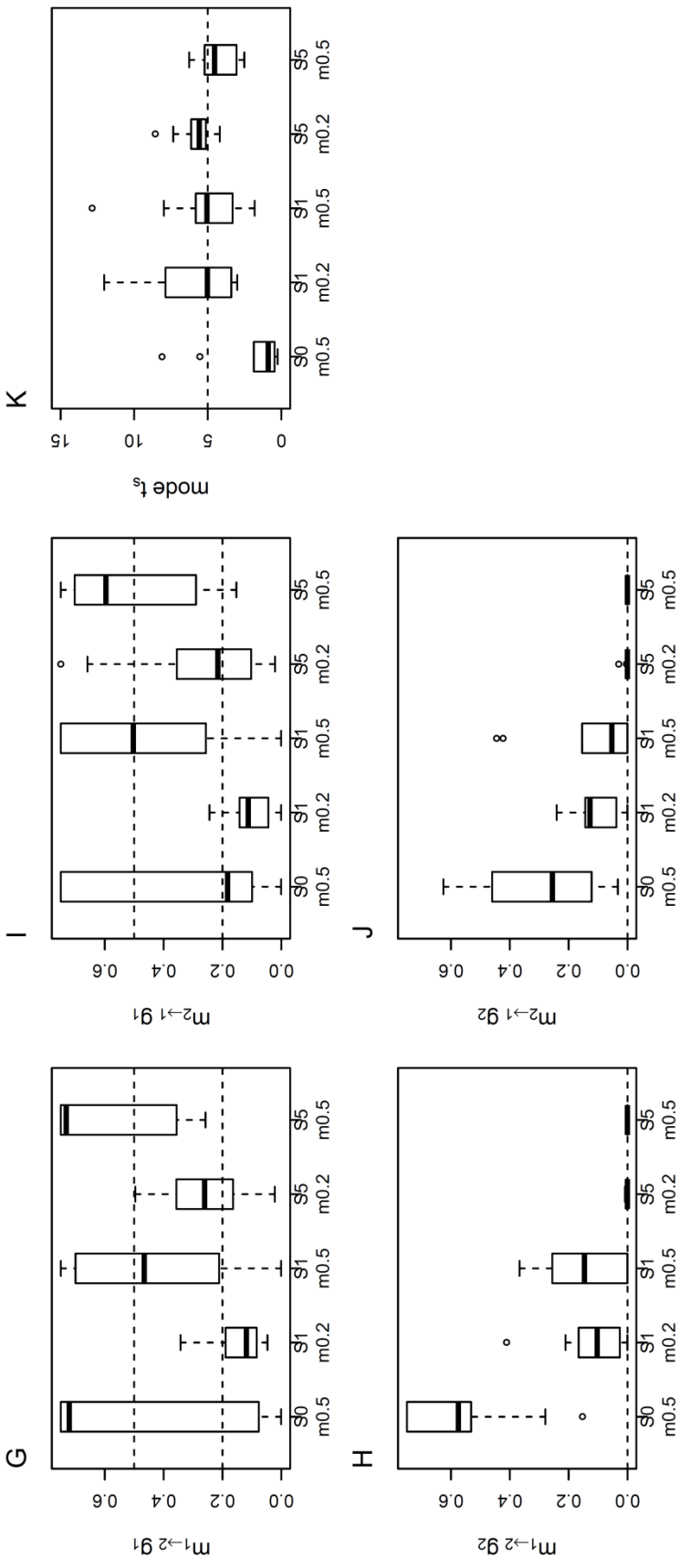

Figure S2 Continued.

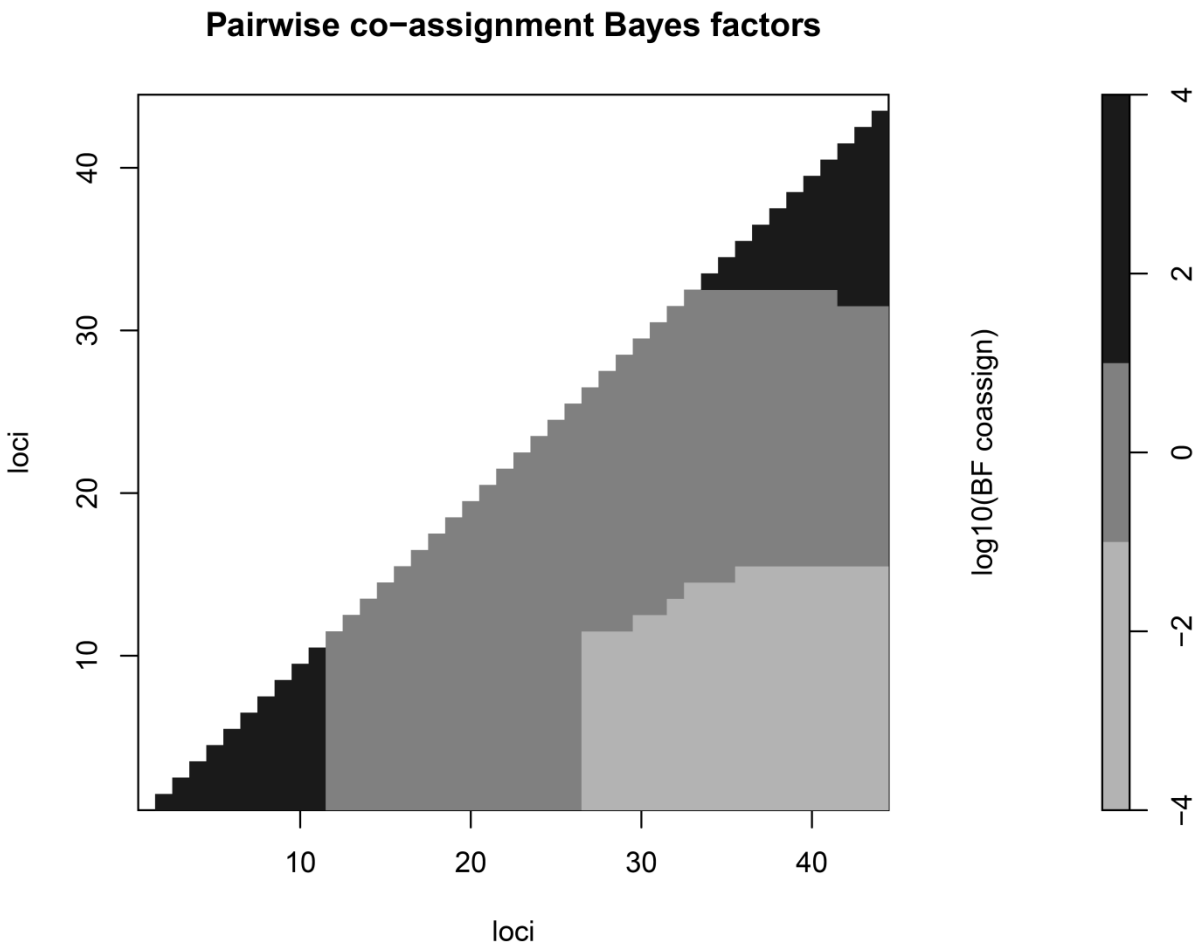

Figure S3 Pairwise Bayes factors (BF) for co-assignment of loci into the same group. The colors indicate the values of the BF in logarithmic scale (log10). A BF larger than 1.0 is considered strong evidence for the hypothesis that both loci are classified into the same group, whereas a BF lower than -1.0 favors the hypothesis that each locus is classified into a different group. The order of loci is the same as in Fig. 6.

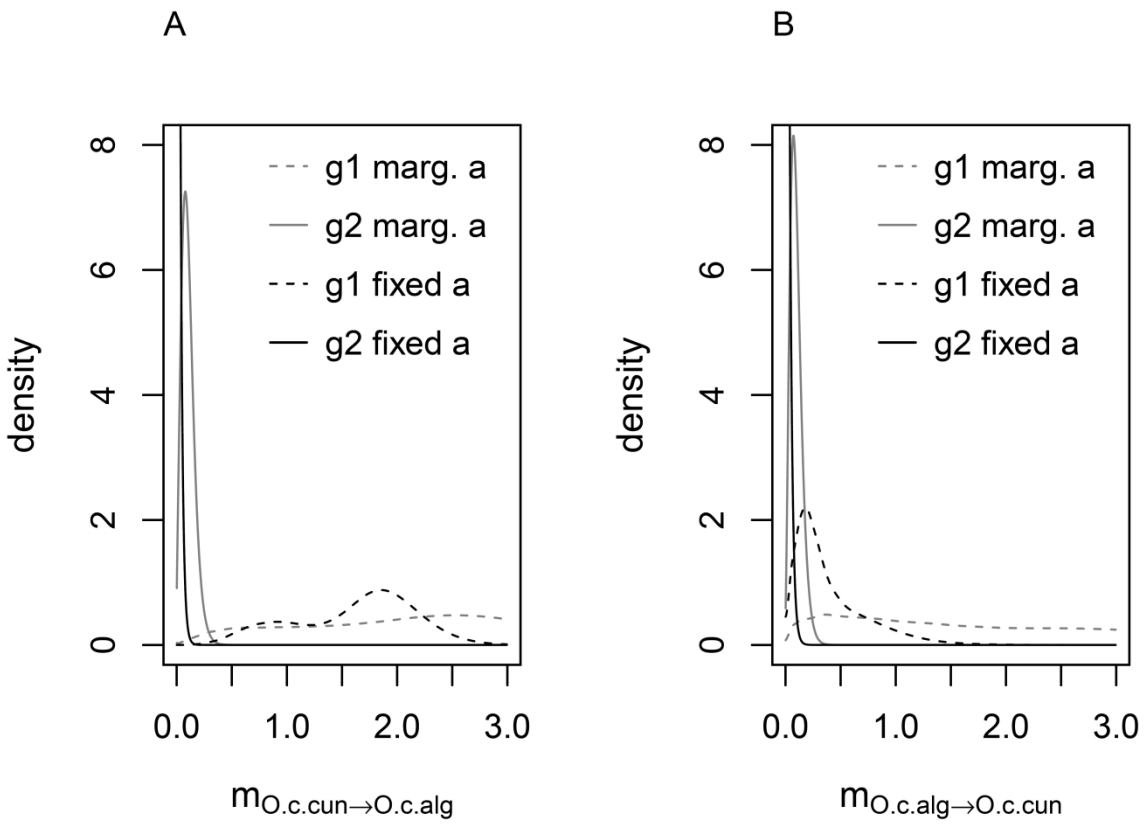

Figure S4 Comparison of posterior distributions for the migration rates in both directions for the European rabbit data, conditional on the assignment of loci based on Bayes factors ("fixed a") and integrating over the assignment ("marg. a").
A) Posterior distributions for migration rate from *O. c. cuniculus* to *O. c. algirus*;
B) Posterior distributions for migration rate from *O. c. algirus* to *O. c. cuniculus*.

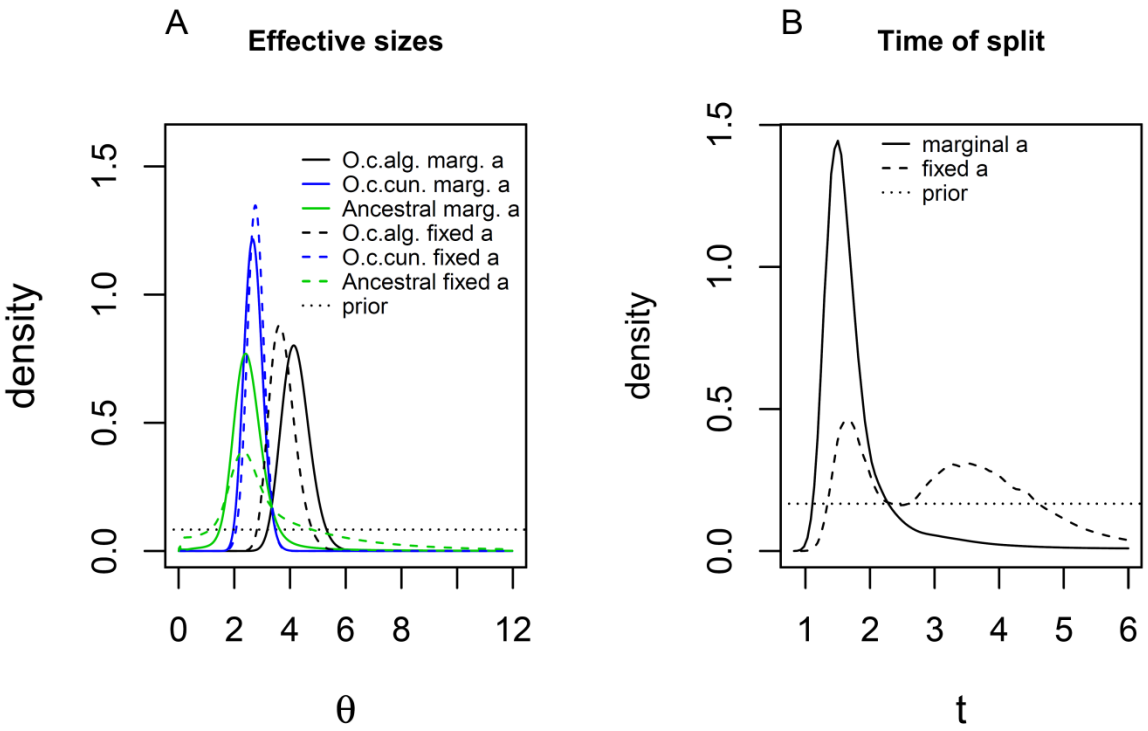

Figure S5 Comparison of posterior distributions obtained conditional on the assignment of loci based on Bayes factors ("fixed a") and integrating over the assignment ("marg. a"). A) effective size parameters (Θ), and B) times of split (t).

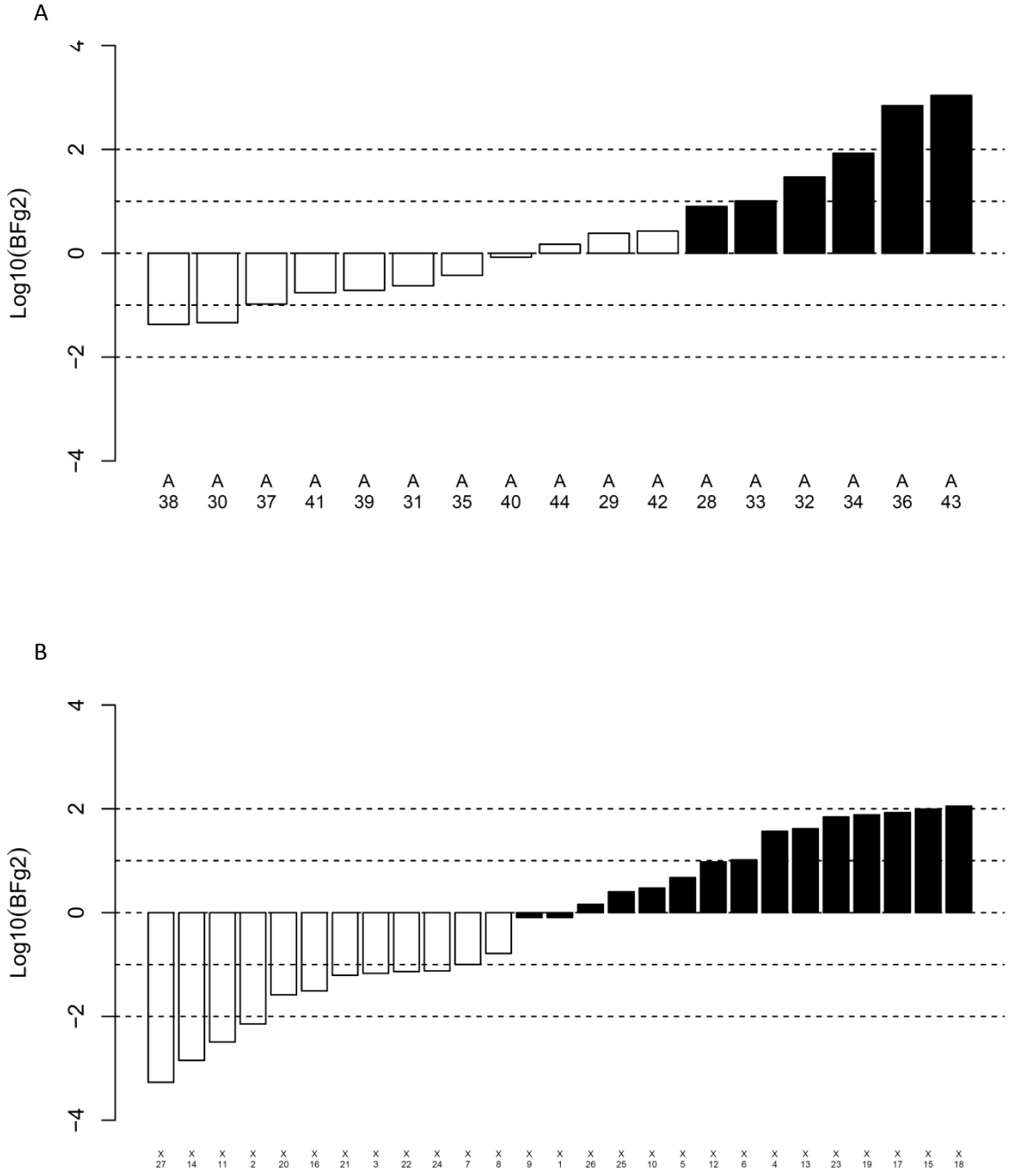

Figure S6 Assignment of loci into groups for European rabbit data.
A) Bayes factors (BF) for assignment of loci into group g_2 in logarithmic scale obtained for the autosomal loci.
B) Bayes factors (BF) for assignment of loci into group g_2 in logarithmic scale obtained for the X-linked loci.
Loci belonging to group g_2 exhibited posteriors indicating low migration rates, identifying this as the selection group. The black and white bars correspond to the groups inferred according to the mean assignment. Each bar corresponds to one locus, and its height represents the BF for the hypothesis that it belongs to group g_2. Loci with $\log_{10}$(BF)<2.0 for autosomes and $\log_{10}$(BF)<1.0 for X-chr were classified into group g_1 (neutral loci). The loci were ordered according to increasing BF values. See Supplementary Table 1 for information on these loci. Results obtained by pooling together three independent runs (30,000 assignment vectors). Runs performed with the following priors: $\theta \sim U[0,12]$, $m \sim U[0,3]$, $t \sim U[0,6]$.

Table S1 Information about the loci of European rabbit used in this study (for details see Carneiro *et al.* 2010).

Locus ID	Locus name[a]	Chromosome	n_1[b]	n_2[c]	L[d]	h[e]
1	PGK1	X	6	10	514	0.75
2	TIMP1	X	8	7	717	0.75
3	MAOA	X	9	9	754	0.75
4	OGT	X	9	11	807	0.75
5	ARHGEF9	X	9	11	759	0.75
6	DGKK	X	9	10	813	0.75
7	OPHN1	X	8	9	574	0.75
8	KLHL4	X	10	8	802	0.75
9	TMEM47	X	9	12	767	0.75
10	POLA1	X	7	11	670	0.75
11	TNMD	X	8	12	544	0.75
12	NRK	X	8	9	717	0.75
13	KLHL13	X	8	11	688	0.75
14	IL1RAPL1	X	9	11	723	0.75
15	AMOT	X	10	10	445	0.75
16	DIAPH2	X	10	10	764	0.75
17	F9	X	10	11	753	0.75
18	FMR1	X	10	11	739	0.75
19	G6PD	X	8	11	573	0.75
20	GRIA3	X	9	11	701	0.75
21	OCRL	X	8	9	456	0.75
22	PABPC5	X	9	12	748	0.75
23	SHOX	X	8	10	536	1*
24	GPC4	X	9	10	512	0.75
25	CYLC1	X	10	12	789	0.75
26	GLRA2	X	8	11	778	0.75
27	PDHA1	X	9	8	804	0.75
28	SLC4A7	14	17	18	695	1
29	MYNN	14	20	16	616	1
30	GK5	14	16	15	719	1
31	GBE1	14	18	24	631	1
32	KPNA4	14	18	22	753	1
33	NAALADL2	14	17	23	505	1
34	ATP12A	8	15	19	930	1
35	CYCT	4	12	19	612	1
36	MGST3	13	16	20	483	1
37	EXT1	3	13	23	1077	1
38	LUM	4	15	20	585	1
39	TIAM1	14	12	17	703	1
40	UD14	7	16	22	551	1
41	PRL	12	18	15	1110	1
42	SIAH2	14	13	17	491	1
43	STAG1	14	15	24	1049	1
44	T	12	8	19	427	1

[a] Locus name as in Table 2 of Carneiro et al. (2010)

[b] Sample size (number of gene copies) for *Oryctolagus cuniculus algirus*

[c] Sample size (number of gene copies) for *Oryctolagus cuniculus cuniculus*

[d] Sequence length in bp (after discarding regions with evidence of recombination)

[e] Inheritance scalar

* SHOX is located in the pseudo-autosomal region of X-chromosome