MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 1383–1399 doi: 10.1111/mec.12182

Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations

STÉPHANE DE MITA,*† ANNE-CÉLINE THUILLET,* LAURÈNE GAY,‡ NOUROLLAH AHMADI,§ STÉPHANIE MANEL,¶ JOËLLE RONFORT§ and YVES VIGOUROUX*

**Institut de Recherche pour le Développement, UMR Diversité, Adaptation et Développement des Plantes (DIADE), Avenue Agropolis, BP 64501, 34394 Montpellier Cedex 5, France, †Institut National de la Recherche Agronomique, UMR Interactions-Arbres-Microorganismes, 54280 Champenoux, France, ‡Institut National de la Recherche Agronomique, UMR Amélioration Génétique et Adaptation des Plantes tropicales et méditerranéennes (AGAP), INRA 2 Place P. Viala, 34060 Montpellier, France, §Centre de Coopération Internationale en Recherche Agronomique pour le Développement, UMR Amélioration Génétique et Adaptation des Plantes tropicales et méditerranéennes (AGAP), Avenue Agropolis, 34398 Montpellier Cedex 5, France, ¶Université Aix-Marseille I, UMR Laboratoire Population Environnement Développement (LPED), 3 place Victor Hugo, 13331 Marseille cedex 3, France*

Abstract

Thanks to genome-scale diversity data, present-day studies can provide a detailed view of how natural and cultivated species adapt to their environment and particularly to environmental gradients. However, due to their sensitivity, up-to-date studies might be more sensitive to undocumented demographic effects such as the pattern of migration and the reproduction regime. In this study, we provide guidelines for the use of popular or recently developed statistical methods to detect footprints of selection. We simulated 100 populations along a selective gradient and explored different migration models, sampling schemes and rates of self-fertilization. We investigated the power and robustness of eight methods to detect loci potentially under selection: three designed to detect genotype–environment correlations and five designed to detect adaptive differentiation (based on F_{ST} or similar measures). We show that genotype–environment correlation methods have substantially more power to detect selection than differentiation-based methods but that they generally suffer from high rates of false positives. This effect is exacerbated whenever allele frequencies are correlated, either between populations or within populations. Our results suggest that, when the underlying genetic structure of the data is unknown, a number of robust methods are preferable. Moreover, in the simulated scenario we used, sampling many populations led to better results than sampling many individuals per population. Finally, care should be taken when using methods to identify genotype–environment correlations without correcting for allele frequency autocorrelation because of the risk of spurious signals due to allele frequency correlations between populations.

Keywords: adaptation, Bayesian methods, detection of selection, F_{ST}, genome scan

Received 16 September 2011; revision received 12 November 2012; accepted 21 November 2012

Introduction

Identifying genes underlying adaptation of populations is a key issue in evolutionary biology (Tenaillon & Tiffin 2008). Such studies are essential for understanding the mechanisms underlying the evolutionary response to environments that vary over time and space (Hansen *et al.* 2012). The identification of genes targeted by selection can help disentangle the molecular basis of traits that respond to ecological factors. Identifying such genes in domesticated species may provide opportunities for

Correspondence: Yves Vigouroux, Fax: +33 467416222; E-mail: yves.vigouroux@ird.fr

the genetic improvement of crops or livestock. The availability of large genetic data sets means that whole genomes can be scanned for footprints of natural or artificial selection without prior knowledge of the genetic determinants of the traits concerned, particularly in non-model organisms (Siol *et al.* 2010).

F_{ST}-based tests were among the earliest molecular genetics methods developed to detect selection (Lewontin & Krakauer 1973). These tests are based on the idea that, if local selection occurs at a given locus, differentiation (assessed by F_{ST}) will increase at this locus compared with what is theoretically expected at neutral loci. Differentiation-based tests require at least two subpopulations, but common applications can include more than two, and potentially many, subpopulations. The seminal work of Lewontin & Krakauer (1973) used an island model (IM) and many populations and proposed a simple formula for assessing selection using the variance of F_{ST}. This test was criticized because correlations across populations invalidated the estimation of the variance of F_{ST} (Nei & Maruyama 1975; Robertson 1975) and led to a high rate of false positives. The method was later improved by using stochastic simulations to obtain the null distribution of the F_{ST} test statistic (Bowcock *et al.* 1991; Beaumont & Nichols 1996). Although based on the same underlying model, the simulation scheme introduced by Beaumont and Nichols (BN) dramatically improved the performance of the test (especially its robustness). An extension of the Beaumont–Nichols test was designed to address the case of hierarchically structured populations, a problem encountered with a wide range of applications (Excoffier *et al.* 2009). Vitalis *et al.* (2001) followed the same line but focused on a single pair of populations and included a more complex demographic history. Such a complex model would not be possible when considering many populations because of the large number of parameters to be fixed. Beaumont and Nichols's approach was incorporated in a Bayesian framework in which F_{ST} is a model parameter (Beaumont & Balding 2004; Foll & Gaggiotti 2008). Recently, Bonhomme *et al.* (2010) improved the original Lewontin–Krakauer method to account for variations in population sizes and their inter-relationships under a pure divergence model described by a phylogenetic tree.

Other approaches to detect selection are based on the correlation of genetic variation with environmental data (Endler 1986). They can be used to detect selection along gradients or in heterogeneous environments. They use environmental data and are therefore expected to be powerful in detecting real correlations between genetic structure and environmental factors. For binary SNP markers, logistic regression (LR) appears to be a suitable approach (Joost *et al.* 2007). LR was extended to account for within-cluster autocorrelations (such as a sampling location) through generalized estimating equations (GEE; Poncet *et al.* 2010). In both of the last two approaches, the underlying generalized linear model can be extended to account for allele frequencies at each sampling site (or population). A more sophisticated regression model was introduced by Coop *et al.* (2010) in a Bayesian framework that also includes a correlation matrix between populations, therefore accounting for autocorrelation of allele frequencies.

A family of methods was also developed to specifically address the analysis of clines. In particular, methods have been devised to detect loci involved in reproductive isolation or adaptation along linear transects (Guillemaud *et al.* 1998; Mullen & Hoekstra 2008; Teeter *et al.* 2008). Most of these methods were designed to address one-dimensional sampling (but see Novembre *et al.* 2005). The aim of cline methods is generally to estimate selective coefficients using a few candidate genes. We do not discuss these methods here as we focus on two-dimensional sampling schemes and genome scans for signatures of selection.

The performance and the rate of false positives of tests for detecting selection have been addressed in only a few simulation studies (Pérez-Figueroa *et al.* 2010; Narum & Hess 2011; Vilas *et al.* 2012). To date, correlation-based methods have not been included in such comparisons, and different models of migration have not yet been evaluated. Indeed, the pattern of population structure is expected to affect the performance of most of the tests. Besides, self-fertilization is widespread in nature (Thornhill 1993; Keller & Waller 2002). Even though self-fertilization is expected to affect several population genetic parameters such as the level of diversity, linkage disequilibrium and population differentiation (Golding & Strobeck 1980; Cummings & Clegg 1998; Liu *et al.* 1999; Nordborg 2000), the effect of selfing on the performance of methods to detect selection has not previously been investigated. Another important question is how the sampling strategy influences the result (for instance, sampling many populations with only a few individuals per population vs. a few populations with many individuals per population).

In this study, we used a simulation approach to generate data sets under spatially, demographically and selectively explicit models (Epperson *et al.* 2010). We used these data sets to (i) compare eight methods to detect selection in terms of rates of false positives (type I error) and false negatives (type II error), including some recently developed methods that have yet not been tested such as correlation-based methods; (ii) analyse the efficiency of these methods under different migration models [an IM, a hierarchical IM and a two-dimensional stepping stone model (SS)]; (iii) determine the efficiency of the different methods in selfing vs. outcrossing populations; and (iv) test different sampling strategies maximizing either the number of populations sampled or the number of individuals sampled per population.

Material and methods

Simulations

We simulated data sets using QUANTINEMO (version 1.0.3; Neuenschwander *et al*. 2008), a flexible, time-forward, individual-centred simulation software. We modelled 100 populations arranged in a 10 × 10 regular grid (Fig. 1). Individuals were diploid and hermaphrodite and generations did not overlap. Populations had a constant size of 100 individuals each and never became extinct. Soft selection (population level regulation) was implemented. For one of the statistical tests, a reference (outgroup) population was required. We therefore added a larger (1000 individuals) isolated population that had never been connected to the others through migration.

Demographic models. We used a total of six combinations of demographic parameters. We used two rates of self-fertilization *S*: 0 (allogamy, where self-fertilization is prohibited) and 0.95 (predominantly but not complete selfing). We used three different migration models (Fig. 1G–I): an IM, a two-dimensional SS with reflective boundaries (SS) and a hierarchical model (HM). The population-based migration rate *m* was always 0.01 (1% of individuals migrate to other populations) but the pairwise migration rate m_{ij} from population *i* to population *j* depended on the migration model. In IM, m_{ij} was $m/(k-1)$ where *k* is the number of populations (here 100). In SS, populations were connected to their immediate neighbours and m_{ij} was $m/2$ if the source population was at one corner, $m/3$ if the source population was at a border but not in the corners and $m/4$ otherwise. In HM, m_{ij} was $m_{\text{within}} = m \times f/(d-1)$ if *i* and *j* belonged to the same quarter, and $m_{\text{between}} = m \times (1-f)/[(k-1) \times d]$ otherwise, where *f* is the proportion of local migration (0.9), *k* is the number of clusters (4) and *d* is the number of populations per cluster (25). In all three models, the migration rates to and from the outgroup population were 0.

Simulation of initial allele frequency at mutation, drift and migration equilibrium. We sampled initial allele frequencies from a simulated distribution at mutation–drift equilibrium. To achieve this, we first ran a preliminary analysis of 10^6 generations and 100 neutral loci with a mutation rate of 10^{-6} per generation and per locus. In time-forward simulations, it is not trivial to force the number of mutations occurring to be one per locus. Because time constraints did not allow an increase in the number of loci and a decrease in the mutation rate, we applied a finite-allele mutation model with a large number of alleles (256) to prevent homoplasy. In this way, sites at which multiple mutations occurred were more likely to be identified and loci with exactly two alleles were highly likely to have undergone only one mutation. We repeated the analysis using two initial conditions: either no polymorphism (all loci fixed) or maximum polymorphism. We ran these simulations using 100 populations and six different demographic models as described above. During the simulations, we monitored the number of fixed loci and the number of loci with exactly two alleles and observed that the mutation–drift equilibrium was reached after roughly 10^5 generations. As expected, the equilibrium values depended on the model but not on the starting conditions. We based our final simulations on the allele frequencies observed at equilibrium based on these preliminary results: we randomly selected global allele frequencies (merging all 100 populations) every 10^4 generation discarding the first 2×10^5 generations. Random noise was added to the allele frequencies sampled. These allele frequencies were used as the starting point of simulation of selected and neutral loci.

Simulation of neutral and selected loci. Our final simulations ran for 5000 generations. We monitored the convergence of *F*-statistics towards theoretical expectations and found that this number of generations (5000) was sufficient to reach the migration-drift equilibrium.

For each repetition, we simulated 751 unlinked loci with a mutation rate of 10^{-6} per generation and per locus. To emulate SNP markers, we only selected loci that exhibited exactly two alleles at the end of the simulation. We randomly selected 10 of these biallelic loci (as a result, all neutral loci of our data sets were necessarily polymorphic). The last locus was under selection. We defined a one-locus, fully additive, quantitative trait model without environmental variance. We set the number of alleles to two with an initial frequency of 0.5 in all populations.

We performed 100 repetitions of each simulation. So, for each parameter combination, we consequently generated 1000 (100 × 10) neutral and 100 (100 × 1) selected loci, all with exactly two alleles.

Selection. We applied a gradient of selection along the environment. The selective pressure was the same for

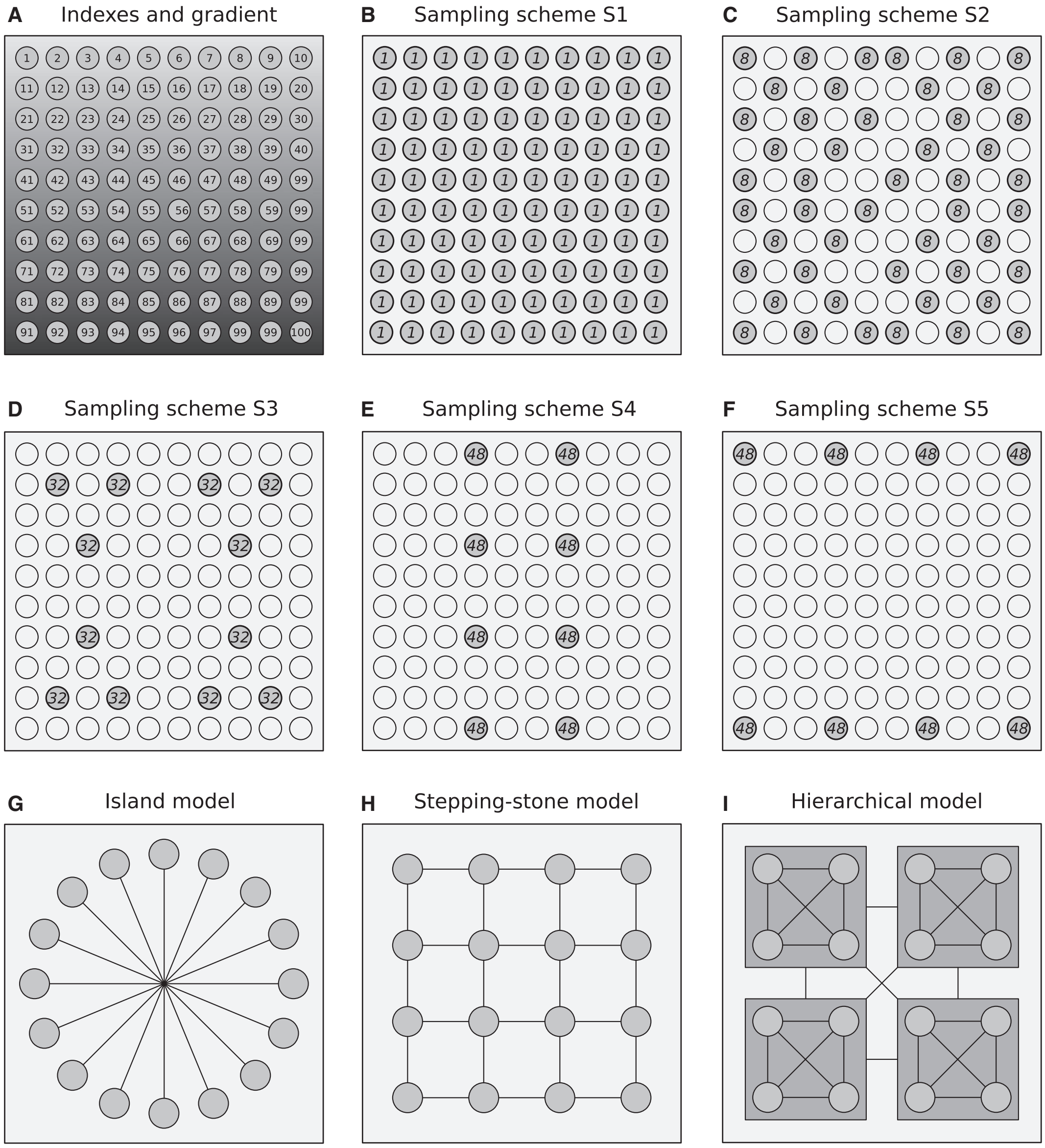

Fig. 1 Schematic diagram of the demographic model. (A) Index of all populations and their location in the virtual environment. The colour gradient symbolizes the selective gradient. (B–F) Sampling schemes. In these five panels, filled circles represent sampled populations and the number of sampled genes is in italics. For C, D and E, and F, the numbers correspond to 4, 16 and 24 sampled diploid individuals. (G–I) Illustration of the three models of migration. Lines represent opportunities for migration. In I, the darker frame represents regions with increased migration rates. For the sake of clarity, only 16 populations are presented in G–I. The outgroup (population 101) is not included and is located in the middle of the gradient, with the same number of samples as the other populations, and is never connected by migration to other populations.

populations labelled 1–10 and so on (Fig. 1A). With the two possible alleles denoted A and B, selection favoured B in populations 1–50 (most strongly in the first 10) and A in populations 51–100 (most strongly in the last 10). In any population from line i (i ranging from 1 to 10), the selective value of the genotypes AA, AB and BB were, respectively, 1, $1 - s_i$ and $1 - 2s_i$. In s_i, there was a fixed component (s_0) expressing the overall strength of selection and a component varying across populations to implement the gradient: $s_i = s_0 \times e_i$ where e_i is an *ad hoc* scaling parameter such as $e_i = (5.5 - i)/4.5$ for $i = 1, 2, \ldots, 10$ (e_i ranges from 1 to -1). The value e_0 for the outgroup (population 101) is 0 because there is no selection. We used four different selection intensities, defined by the maximum selection effect s_0. The values used were 0.05, 0.01, 0.005

and 0. We performed 100 replicated simulations (100 populations, 5000 generations) for each parameter combination, making a total of 2400 simulation runs.

Sampling. QUANTINEMO generates genotype data for all individuals. We defined five sampling strategies (Fig. 1B–F). In S1, one individual was sampled per population in all populations for a total of 100 alleles (one allele per individual). In all four other sampling modes, a total of 384 alleles were sampled, balancing the number of populations sampled by the number of samples per population. In S2, four random individuals (hence eight alleles) were sampled per population in 48 regularly sampled populations. In S3, 16 random individuals (hence 32 alleles) were sampled per population in 12 populations. In both S4 and S5, 24 random individuals (hence 48 alleles) were sampled per population in eight populations. In S4, the eight populations were organized as two transects parallel to the environmental gradient. In S5, the eight populations comprised four populations sampled at each extreme of the gradient. When data from the outgroup population were required, an identical number of samples was sampled in this population.

Implementation of methods for detecting selection

We used eight methods (Table 1). In the next two sections, we describe the implementation, settings and data samples used. The methods belong to two different categories: correlation-based methods that use genotype by environmental data and differentiation-based methods that use a measure of between-population differentiation (such as F_{ST}). For each combination of demographic parameters, selective pressure and sampling strategy, we generated one data set containing the 10 neutral loci from each of the 100 repetitions (1000 neutral loci) and the selected locus from each repetition (100 selected loci), which we analysed using the eight methods. Note that only sample schemes S2–5 allowed the methods to be compared, because S1 was only used for the LR method (Table 1), and the method designed by Excoffier *et al.* (2009) was applied only to the hierarchical IM.

For correlation-based methods. We used three variables (V1–3) to account for varying levels of noise around the true value. For each row i of the environment (i = 1, 2, …, 10), we defined the environmental value as ($100 + i + X$) where X was drawn randomly and independently for each population in a normal distribution of mean 0 and of standard deviation of either 0 (for V1), 2 (for V2) or 4 (for V3). Therefore, in V1, the environmental variable was supposed to be exactly known (and was the same for all populations in a given row), while in V2 and V3, two levels of noise were added.

Logistic regression method. Logistic regression (LR) is a generalized linear model for binary response data (McCullagh & Nelder 1989). Joost *et al.* (2007) applied LR to genome scans to look for signatures of local adaptation. We used R software (R Development Core Team 2011) to analyse data from all sampling modes and using all three environmental variables. For samples S2–5, the fact that several samples were collected in a single population was not taken into account. We fitted two models to data from each locus. In model M_1, genotypes were used as the response variable and the environmental variable was used as the explanatory variable. We used a logit link model with a binomial error variable for the regression. The null model (M_0) contained only the intercept (ignoring the environmental variable). To assess the goodness-of-fit of model M_1, we used two complementary tests (Joost *et al.* 2007): the likelihood ratio test (LRT) and the Wald test. A locus was significant if p_{LRT} was <0.05 and if p_W was <0.025 or >0.975 where p_{LRT} is the P-value of the LRT and p_W is the P-value of the Wald test.

Generalized estimating equations. Generalized estimating equations (GEE) are an extension of LR where a correlation matrix is added to the LR model. This matrix captures correlations between samples taken from a given population (Poncet *et al.* 2010). The GEE method was implemented in R, partially based on code from the APE package (Paradis *et al.* 2004). We used all five sampling modes and all three environmental variables. A null model M_0 was compared to a model M_1 accounting for the genotype–environment correlation. As the LRT is not applicable to GEE, we computed a quasi-likelihood criterion QIC as described in Pan (2001). The Wald test was computed in the same way as for LR, and a locus was declared significant if $QIC_1 < QIC_0$ and if p_W was <0.025 or >0.975.

Coop, Witonsky, Di Rienzo and Pritchard method. The metric used by Coop, Witonsky, Di Rienzo and Pritchard (CWDRP) is allele frequencies in populations (Coop *et al.* 2010). This method uses a Monte Carlo Markov chain to estimate the covariance matrix of allele frequencies. We used only neutral loci for this step. CWDRP uses the covariance matrix to control for between-population correlations and analyses the correlation between allele frequencies and the environmental variable in a Bayesian framework. A Bayes factor was computed to compare the model that allowed frequency–environment correlations with the null model. We used samples S1–5 and all three

Table 1 List of methods

Method and reference	Technique	Underlying model	Env. variable	Control loci	Sampling	S1	S2	S3	S4	S5
LR Joost *et al.* (2007)	GLM	Independence of observations	Yes	No	Individuals	+	+	+	+	+
GEE Poncet *et al.* (2010)	GEE	Independence of clusters	Yes	No	Several individuals per population	–	+	+	+	+
CWDRP Coop *et al.* (2010)	MCMC	Island model	Yes	Yes	Frequencies	–	+	+	+	+
FLK Bonhomme *et al.* (2010)	Forward simulations	Multiple divergence model	No	Yes	Frequencies	–	+	+	+	+
BN Beaumont & Nichols (1996)	Coalescent simulations	Island model	No	Yes	Frequencies	–	+	+	+	+
EHF Excoffier *et al.* (2009)	Coalescent simulations	Hierarchical island model	No	Yes	Frequencies	–	+	+	+	+
VDB Vitalis *et al.* (2001)	Coalescent simulations	Pairwise divergence model	No	Yes	Frequencies	–	–	A pair of populations of 24 individuals		
FG Foll & Gaggiotti (2008)	RJ-MCMC	Island model	No	No	Frequencies	–	+	+	+	+

BN, Beaumont and Nichols; CWDRP, Coop, Witonsky, Di Rienzo and Pritchard; EHF, Excoffier, Hofer and Foll; FG, Foll and Gaggiotti; GLM, generalized linear model; GEE, generalized estimating equations; LR, logistic regression; MCMC, Monte Carlo Markov chain; RJ-MCMC, reversible-jump Monte Carlo Markov chain; VDB, Vitalis, Dawson and Boursot. Columns S1–5 show which sample scheme was used in combination with the test.

environmental variables. We set the number of iterations to 10^3 for both stages of the procedure. A locus was declared significant when the Bayes factor reported by the program was >3, which is a common decision threshold (Jeffreys 1961).

A pair of Extended Lewontin and Krakauer (sFLK) method. The method of Lewontin & Krakauer (1973) was designed to provide critical values using F_{ST} as a test statistic under the IM. One assumption of this test is that populations are independent. The extension considered here relaxes this hypothesis and uses a phylogenetic tree connecting the populations (Bonhomme *et al.* 2010). This method is based on a statistic computed using the overall F_{ST} (among all populations). To apply the test, we used the R and Python code provided by the authors. We used samples S2–5, and we used the outgroup population to root the tree. We used control loci to reconstruct the tree, and we tested all loci using forward simulations (10^4 replicates). To avoid sampling bias, we treated the 100 repetitions independently, each time reconstructing the tree from 10 neutral loci. A locus was declared significant if the *P*-value was <0.025 or >0.975.

Beaumont and Nichols method:. The Beaumont and Nichols (BN) test is based on coalescent simulations to generate the distribution of F_{ST} conditioned on the expected heterozygosity under the IM (Beaumont & Nichols 1996). We implemented this test in a Python program using the population genetics package EGGLIB (De Mita & Siol 2012). Simulations featured a single mutation and an IM with the same number of populations as the tested data (100). The only nuisance parameter was the composite migration parameter $M = 4Nm$, which we estimated using the relation $M = (1 - F_{ST})/F_{ST}$ where N, m and F_{ST} are the theoretical values of respectively the size of a single population, the migration rate and the between-population differentiation. F_{ST} was estimated using the estimator of Weir & Cockerham (1984). We generated the joint distribution of the expected heterozygosity (*He*) and F_{ST} over 10^4 replicates, and we binned the distribution over 20 regular *He* categories. We used F_{ST} as a test statistic, and the test was performed conditionally to the *He* value of each locus. A locus was declared significant if the bilateral *P*-value was <0.05. This test was applied to data from sampling modes S2–5.

Excoffier, Hofer and Foll (EHF) method. The EHF test is an extension of the BN test in that it accounts for a HM of population structure (Excoffier *et al.* 2009). We thus applied this test only for data sets generated under the hierarchical migration model. We also used the Python package EGGLIB to perform coalescent simulations assuming the correct HM: four quarters of 25 populations each connected by two rates of migration. The within-quarter pairwise migration rate was set to $4 \times Ne \times m_{\text{within}}$ and the between-quarter pairwise migration rate was set to $4 \times Ne \times m_{\text{between}}$ where m_{within} and m_{between} have been defined earlier and Ne is the population size rescaled by the effect of selfing $N/(1 + F_{\text{IS}})$ where F_{IS} is itself $s/(2 - s)$, N is 100 and s is 0 or 0.95 depending on the model. The simulations were conditioned on observing only one mutation. We generated the joint distribution of the expected heterozygosity (He) and F_{ST} (computed between populations while ignoring the presence of four quarters) over 10^3 replicates, and we binned the distribution over 32 regular He categories. We used F_{ST} as a test statistic, and the test was performed with respect to each locus's He. A locus was declared significant if the bilateral P-value was <0.05. Samples S2–5 were used for this test.

Vitalis, Dawson and Boursot method. The Vitalis, Dawson and Boursot (VDB) method considers a single pair of populations and allows a demographic model that would not be computationally achievable with more than two populations (Vitalis *et al.* 2001). The model specifies divergence of two populations from an ancestral population immediately after a bottleneck. VDB uses two statistics, F_1 and F_2, which describe the amount of divergence of each population with respect to the ancestral population. To apply the VDB test, we used the DETSEL program (Vitalis 2012). To average the effect of selecting a particular pair of populations, we selected 10 random pairs of populations (sampled without replacement). In each population, a sample of 48 genes was taken as in sample scheme S3. We applied the test to each pair independently. The F_1 and F_2 statistics (computed from neutral loci) determine the size of the two populations in the model. The other demographic parameters are specified by the user. We disabled the ancestral population bottleneck by fixing its duration to 0 and we used a mutation rate of 10^{-6} per generation, a divergence time of 1000 generations and an ancestral population size of 5×10^4 individuals. We investigated other parameter values and checked that they resulted in similar or worse performances, as assessed by the fit of simulated F_1 and F_2 values to observed values. We performed 10^6 coalescent simulations and generated the distribution of test statistics F_1 and F_2. We then tested all loci and computed each P-value as the P-value of the observed F_1 and F_2 (F_{obs}) values. In the case of an SNP, F_1 and F_2 are discrete. This P-value was computed as $\Sigma[P(F_i) \leq P(F_{\text{obs}})]$ where $P(F_i)$ is the frequency of a given pair of F_1 and F_2, and $P(F_{\text{obs}})$ is the frequency of the observed set of values among simulations. For each locus, the test was repeated 10 times (for all population pairs). In cases where $P(F_{\text{obs}}) = 0$, the result of the test was ignored. When merging the results of the 10 tests for a given locus, we applied a Bonferroni correction: a locus was significant whenever at least one of the P-values was <0.005, corresponding to a threshold of 0.05 and 10 tests.

Foll and Gaggiotti (FG) method. The Foll and Gaggiotti (FG) method (Foll & Gaggiotti 2008) extends the method of Beaumont & Balding (2004). Both methods assume a model where allele frequencies are expressed as a function of F_{ST} which is itself modelled as the product of demographic and locus factors. The latter, α, encompasses the effect of selection. The null hypothesis is represented by $\alpha = 0$. FG uses a reversible-jump Monte Carlo Markov chain scheme to determine the probability of non-null α (expressed as a Bayes factor per locus). All neutral and selected loci were tested. We used the program BAYESCAN version 2.1 and the following settings: number of output points: 5×10^3, thinning interval: 10, burn-in period: 5×10^4, pilot runs: 10, pilot run length: 5×10^3, prior odds: 10. A locus was declared significant if the Bayes factor was >3. The test was applied to samples S2–5.

Results

Allele frequency clines caused by selection

We examined the effect of the selective gradients on allele frequency at selected loci under each combination of parameters. For this purpose, we generated plots of allele frequency clines under the different intensities of selection and under the different demographic models (Fig. 2).

The strongest selection setting ($s_0 = 0.05$) resulted in very steep clines at selected loci. With weaker selection ($s_0 = 0.01$ and 0.005), clines were less steep but, nevertheless, discernible. Long-range migration allowed by the IM prevented fixation of the favourable allele in extreme populations (Fig. 2A, B). In contrast, in the SS, fixation of the favourable allele was frequent even under weak selection, and the allele frequency cline was much steeper in all cases (Fig. 2C). The HM increased the efficiency of selection due to the matching of block limits with the switch of selection regime (Fig. 2D). As a result, the cline was very steep between the top and bottom blocks regardless of the strength of selection but almost vanished within blocks. However, long-range migration was permitted, albeit at a reduced frequency, which prevented the systematic fixation of the favourable allele.

Selfing tended to increase the effect of selection (Fig. 2B). Self-fertilization might hamper selection by random loss of the favourable allele. However, our

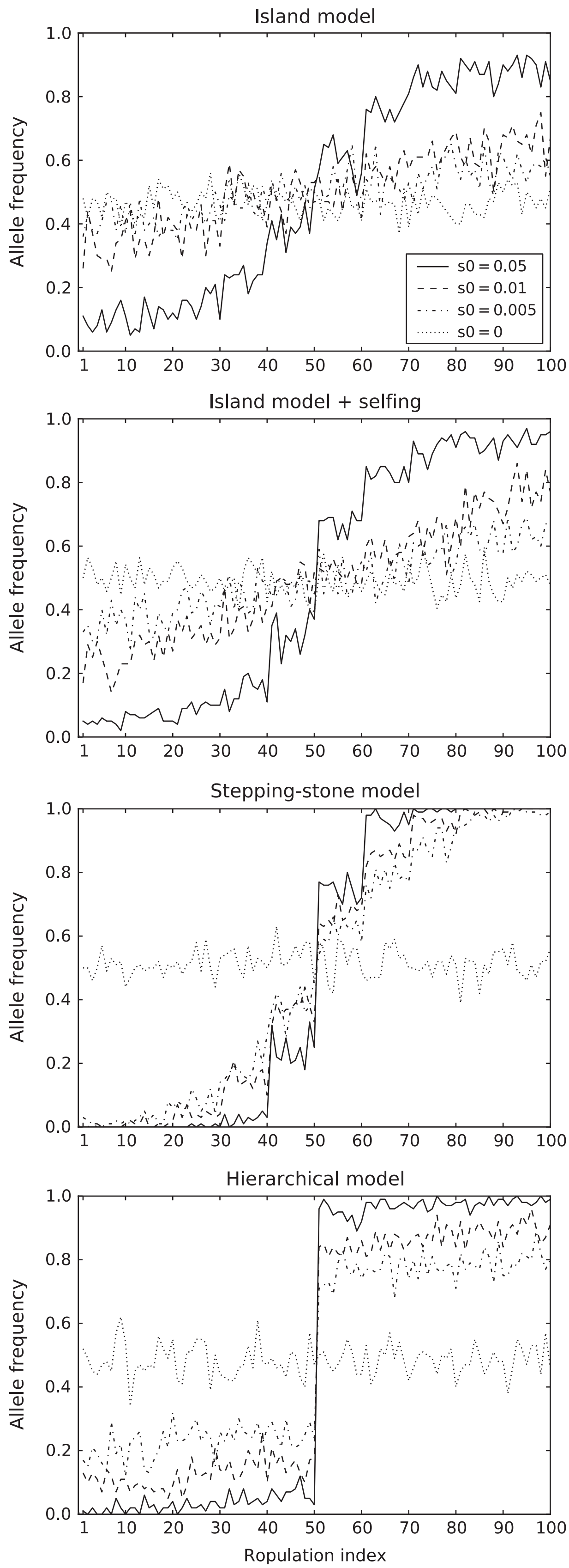

Fig. 2 Allele frequency cline caused by selection. Each panel represents the average frequency of allele A (using sample scheme S1) in the 100 populations, depending on the strength of selection (s_0 = 0 shown as a control). In A, C and D, the reproduction regime is allogamy.

setting with 100 populations interconnected by migration and balanced initial allelic frequencies prevented fixation of alleles at selected loci. On the other hand, self-fertilization decreases the proportion of heterozygotes. Alleles are therefore almost always exposed as homozygotes and can be more easily selected. As a result, non-adaptive alleles that are permanently added by migration can be more easily removed.

Detecting loci under selection

Effect of the migration model. With the highest level of selection (s_0 = 0.05) and allogamy, all methods produced fairly good results (Fig. 3). Under the IM, the worst method (FLK) still detected 92% of true positives and had a false positive rate of 10.9%. The stepping stone and HM both caused a major increase in false positive rates for LR and GEE (especially the SS) and to a lesser extent for CWDRP (with the HM in particular). The differentiation-based methods were not affected. Detection of true positives was not altered by the migration model, consistent with the fact that the two latter models of migration magnified the effect of selection.

Effect of selfing. The reproductive regime had consequences for the detection of selection. Combined with the IM of migration (Fig. 4), selfing reduced the power of the differentiation-based methods, except for BN, which showed a good rate of true positives (93%). Combined with other models of migration, selfing caused a dramatic loss of power (down to 0% in one case) for FLK and VDB. As the signal of selection is actually increased by the combination of selfing and non-IMs, this loss of power was probably caused by a technical artefact when applying the methods.

Logistic regression and GEE were also affected by selfing, exhibiting a consistent increase in false positives under all migration models in the case of selfing. The strongest relative increase was observed with the HM (increase from 9.6% to 24.1% for LR and from 20.1% to 38.5% for GEE). The CWDRP method behaved differently: selfing caused an increase in the rate of false positives under the IM, but a slight decrease under the two other models of migration.

Effect of the strength of selection. We monitored the loss of power caused by the decrease in selection strength

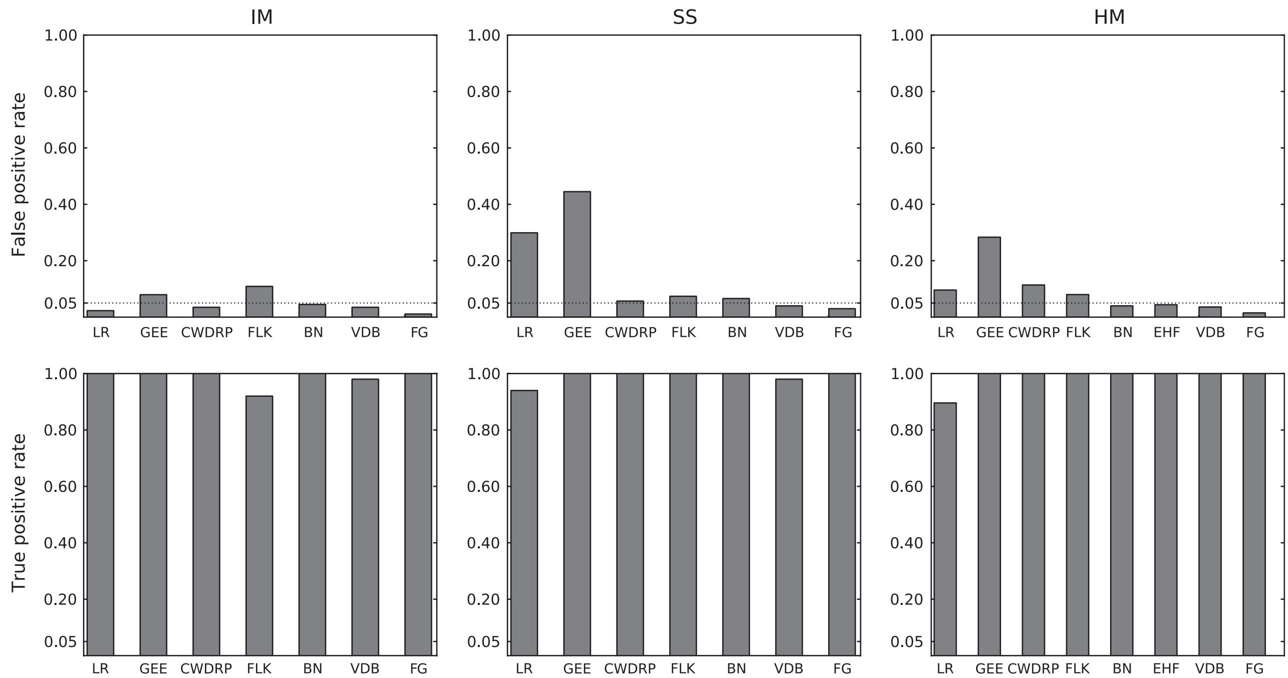

Fig. 3 False and true positive rates obtained with $s_0 = 0.05$ and allogamy. The top three panels give the false positive rates of the seven or eight tests, and the bottom three panels give the true positive rates. IM, island model; SS, stepping stone model; HM, hierarchical model. The dotted line at the top of three panels indicates an error rate of 0.05. Results were obtained using sample S2 except for LR for which S1 was used, FLK for which S3 was used and VDB which was based on pairwise comparisons. EHF appears only in HM.

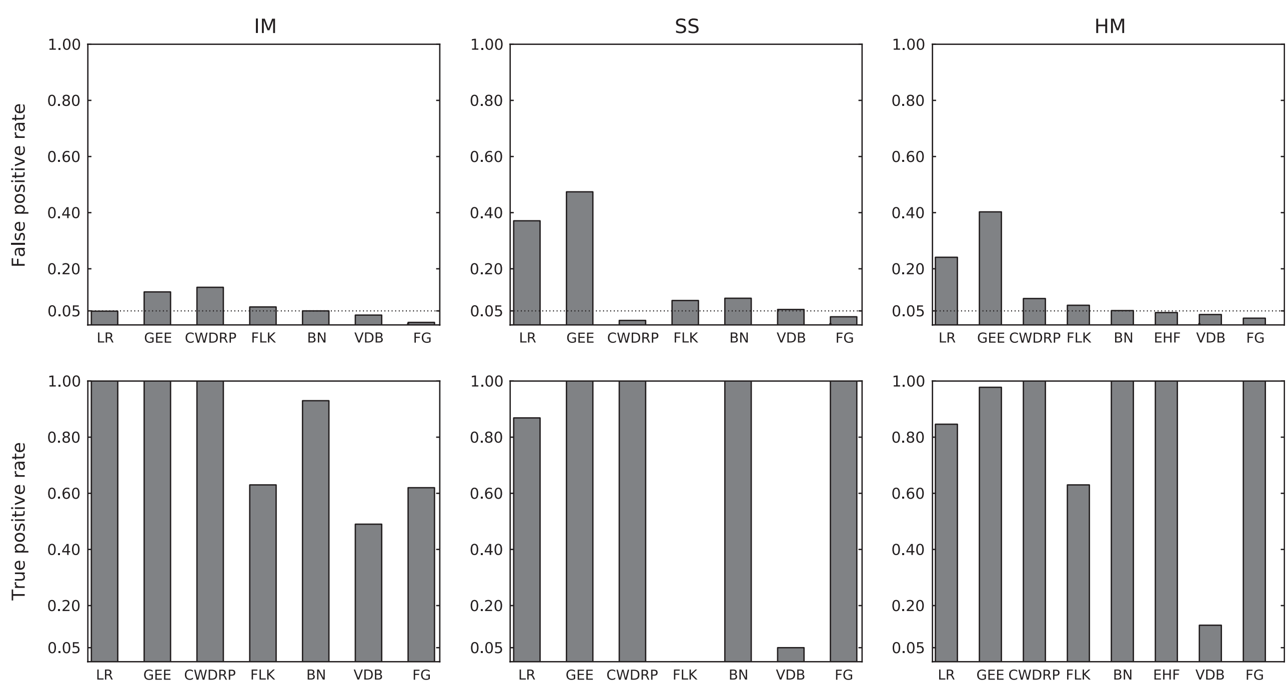

Fig. 4 False and true positive rates obtained with $s_0 = 0.05$ and selfing. The top three panels show the false positive rates of the seven or eight tests, and the bottom three panels give the true positive rates. IM, island model; SS, stepping stone model; HM, hierarchical model. The dotted line in the top three panels indicates an error rate of 0.05. Results were obtained using sample S2 except for LR for which S1 was used, FLK for which S3 was used and VDB which is based on pairwise comparisons. EHF appears only in HM.

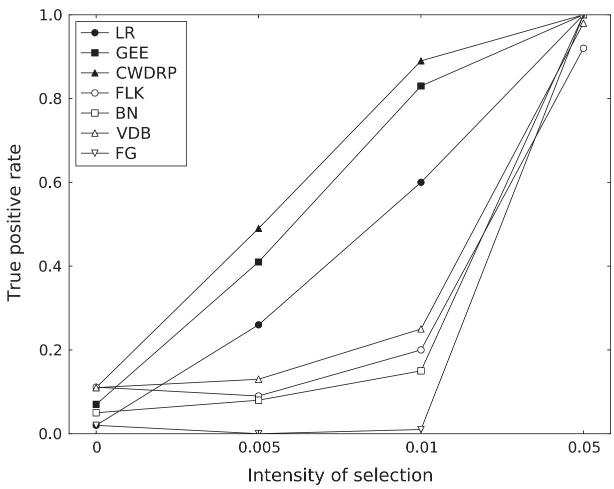

Fig. 5 Effect of the intensity of selection on true positive rates. Results are for the island model of migration, allogamy and using sample S2, except for LR for which S1 was used, FLK for which S3 was used and VDB, which is based on pairwise comparisons. Black symbols represent correlation-based methods. White symbols represent model-based methods.

under the IM of migration combined with allogamy (Fig. 5). The effect of selection intensity differed between the correlation-based methods (LR, GEE and CWDRP) and the differentiation-based methods. The former were less affected and still performed fairly well at the weakest level of selection ($s_0 = 0.005$). CWDRP appeared to have the best power when using discriminant selection intensities. The power of other methods decreased dramatically at $s_0 = 0.01$, especially FG which did not exhibit a true positive rate higher than that of the control ($s_0 = 0$). At $s_0 = 0.005$, none of the four differentiation-based methods showed more true positives than the control.

Effect of sampling. The effect of sampling (Fig. 6) was investigated using five sampling schemes (Fig. 1B–F). They were labelled in order of increasing number of samples per population (from 1 to 48) and of decreasing numbers of sampled populations (from 100 to 8). S4 and S5 differed only in the organization of the eight sampled populations. LR showed a dramatic increase in the rate of false positives for all sampling schemes incorporating more than one sample per population. With allogamy and the IM, the false positive rate increased from 2.3% with S1 to 14.2% with S2, 32.2% with S3 and similar values with S4 and S5. The effect was magnified in the case of selfing (Fig. 6, right panels). In both allogamy and selfing models, LR also underwent a marginal loss of power (98% or 99%) with sampling modes other than S1. The stepping stone and HM further increased the rates of false positives, but only marginally, while rates of false positives were strongly affected with sampling modes S3, S4 and particularly S5, which dropped to <20% or even to 0% (data not shown). In contrast to LR, GEE was designed to cope with multiple samples per population and prevented the excess of false positives, at least when passing from S1 to S2 (6.8–7.5% under allogamy, 11.7–11.1% under selfing, both with the IM). However, sampling modes with more samples per population caused an increase in the false positive rate (up to 17%) but only with the IM with allogamy. The same pattern appeared with CWDRP, suggesting that it is not caused by within-population correlations (since CWDRP only takes the population frequency into account). On the contrary, CWDRP generated fewer false positives when using many samples per population under selfing, probably due to the higher variance of allelic frequencies in that case. The effect of sampling on false positives was moderate with the other methods, which do not allow the use of S1. The most notable exception was FLK in which, in the case of allogamy and the IM (but also the stepping stone and HM, data not shown), S3 reduced the number of false positives (from 25.2% to 10.9%). Apart from this case, the effect on false positives was moderate. Further, in the case of the HM, S2 showed an excess of false positives with CWDRP, FLK or BN (data not shown). On the other hand, sampling modes with eight populations (especially S4) performed badly in terms of true positives. This was particularly true for FG for which all sampling modes with <48 populations yielded dramatically decreased rates of true positives. For a sample with only eight populations, sampling in the extreme of the gradient (S5) rather than along the gradient (S4) led to better true positive rate for FG and BN.

Discussion

Detecting signatures of selection has become a popular way of identifying loci involved in the response to environmental factors and other selective pressures, especially for non-model species where other approaches, such as candidate genes, are not easy to develop (Siol *et al.* 2010). New statistical tests have been developed to detect such loci under selection and there is now a pressing need for a critical assessment of their performances. We simulated data sets incorporating neutral loci and loci affected by a known gradient of selection, in an environment of 100 populations connected by migration. We examined several factors susceptible to affect the performance of tests in detecting selected loci. The factors investigated were the strength of selection, the pattern of migration, selfing and the sampling scheme. We compared the outcome of eight methods, three based on the correlation between genotypes and environmental variables (LR, GEE and CWDRP) and

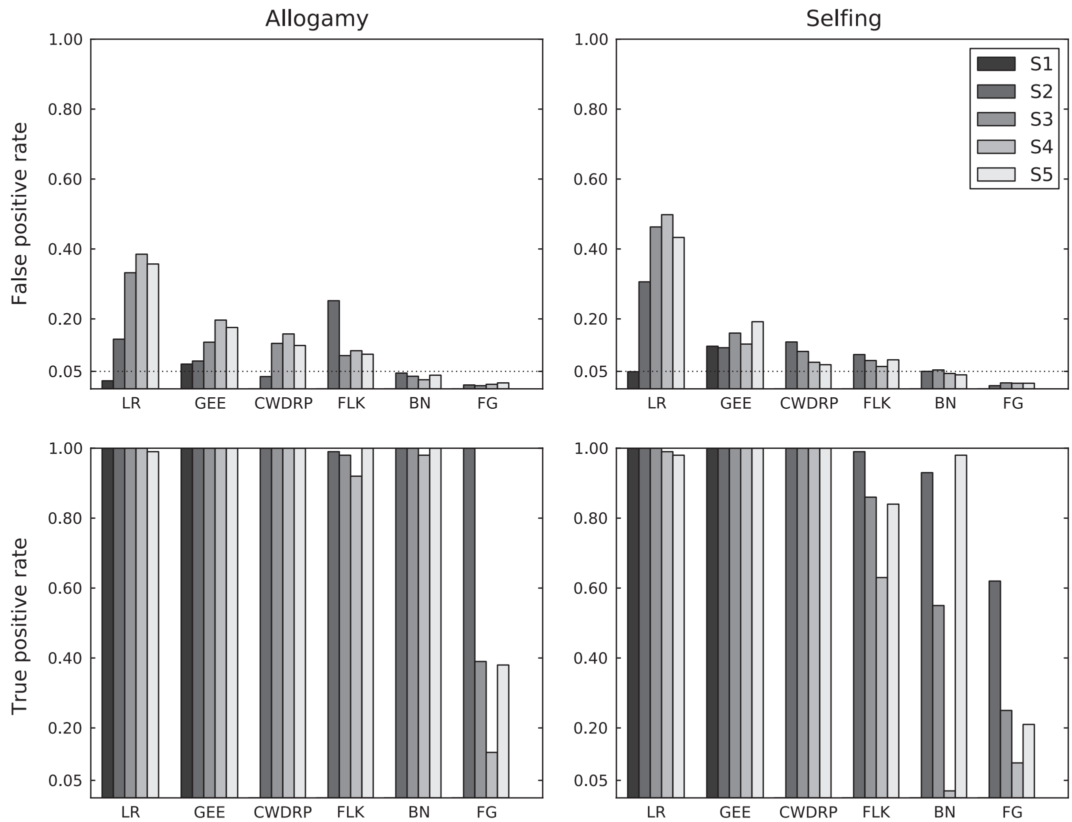

Fig. 6 Effect of sampling with $s_0 = 0.05$, island model of migration. The top two panels show the false positive rates, and the bottom two panels the true positive rates. The two left panels present results obtained under allogamy and the two right panels under selfing. The dotted line in the two top panels indicates an error rate of 0.05. Two tests are not shown: EHF, which was applied only to hierarchical models and on which the sampling modes had little effect (data not shown) and VDB, which was only applied to pairwise comparisons. S1 is applicable only with LR and GEE.

five that investigate genetic differentiation between populations (FLK, BN, EHF, VDB and FG).

Comparison with other simulation studies

Our analysis showed that FG has very low rates of false positives and a good power of detection of actual selection under several demographic models with many populations. This confirms the results obtained by Pérez-Figueroa *et al.* (2010) and Vilas *et al.* (2012). The first study used simulated AFLP data and a pair of populations exhibiting a differentiation of $F_{ST} = 0.025$, 0.1 and 0.3. In our simulations, the estimated value of F_{ST} ranged from 0.2 to 0.55 (depending on the demographic model). Pérez-Figueroa *et al.* (2010) used a selection of strength $s_0 = 0.25$, 0.025, 0.0025, whereas we used 0.05, 0.01 and 0.005 (although the values cannot be directly compared due to differences in the models). They observed that the power of VDB collapsed for a subset of parameter combinations, in agreement with our results. The power of all methods tended to decrease with increased differentiation and, as expected, with decreasing intensity of selection. This explains why, when we decreased the level of selection, we rarely detected any true positives in our system (which matches their highest level of differentiation). However, we showed that even in these unfavourable conditions, correlation-based methods were sufficiently powerful. The second study (Vilas *et al.* 2012) specifically investigated the performance of BN and FG methods in detecting neutral markers linked to selected loci. They also used a system with two populations and investigated a range of migration rates centred on 0.01, which also confirmed the good performance of FG. One notable difference between our study and these two studies is that they showed that the methods tested (BN and VDB, and BN and FG, respectively) resulted in a large excess of false positives that needed to be corrected using a false discovery rate (FDR) control procedure (Benjamini & Hochberg 1995). In contrast, we only observed a moderate excess of false positives, possibly as a result of (i) our system, which used a large number of populations which matches the underlying model of the BN method better, (ii) our implementation of the BN method for SNP data and (iii) our technique of integrating several pairwise tests for VDB with correction for multiple testing. One conclusion of the study of Pérez-Figueroa *et al.* (2010) was that combining different methods in a multitest approach did not improve results, chiefly because the three tests (BN, VDB and FG) yielded similar results. However, these methods are based on similar principles, and better results can be expected when complementary methods exploiting different sources of information are used together: typically, differentiation-based and correlation-based methods (Manel *et al.* 2009).

Another recent study compared the performance of four differentiation-based tests (Narum & Hess 2011): the BN method, EHF FG and a more descriptive approach based on global F_{ST} histograms (e.g. Luikart *et al.* 2003). Narum & Hess (2011) used a model with 10 populations arranged in a linear SS and two regimes of selection: one including a linear gradient matching the direction of the stepping stone migration (and thus matching our SS) and one where the gradient of selection was randomized with respect to the model of migration. As before, FG was the method with the lowest false positive rate, which confirms that this method presents very low false positive rates over a wide range of conditions. Similarly, the authors found that BN was as powerful as FG (provided that selection was strong) although it generated more false positives, as we also found with other demographic models.Narum & Hess (2011) found that EHF did not perform well, which is easily explained by the mismatch between the simulation model and the model assumed by the test. This suggests that EHF will be most efficient when data actually exhibit signatures of hierarchical structure.

Running time

The running time of a statistical method is a potentially crucial factor when dealing with very large data sets. We found that LR and GEE were considerably faster than the other methods tested (requiring a few minutes). All other tests required substantially more time (up to many hours). Compared to LR and GEE, the performance gain offered by CWDRP thus comes with a significant cost in computing time.

Factors affecting performances of tests

Several factors are likely to influence the choice of a method to detect signatures of selection: availability of environmental data suspected to induce variance in fitness, genetic structure between the populations studied and sampling strategy. In the following, we discuss how each of these factors affects the performance of the methods we tested.

Using environmental data. The correlation-based methods LR, GEE and CWDRP use environmental data to detect selection. It should be noted that, in the results presented above, the environmental variable was known without error, which might not be the case in real-life situations. The effect of blurring the signal by adding random noise around the true value had the expected effect of reducing the power of all three methods to detect selected loci (Fig. S1, Supporting Information) but it should be noted that the effect was marginal for the highest selection intensity, and that, for example, CWDRP exhibited a fairly decent rate of true positives at $s_0 = 0.01$ with the most blurred variable (61%). In our study, correlation-based methods consistently exhibited sufficient power to detect true selected loci, whatever the demographic model and even with the lower levels of selection. Nevertheless, LR and GEE methods appeared to generate slightly more false positives under standard conditions (allogamy and IM).

On the contrary, differentiation-based methods are based on the principle of rejecting a null model (detection of outliers). Consequently, they do not enable identification of the factor that causes the selective pressure. This may be an advantage when no environmental variable is available. In our study, they were less powerful than correlation-based methods. In particular, with the lowest level of selection ($s_0 = 0.005$), all the methods failed to detect selection (there was no difference between $s_0 = 0.005$ and the control $s_0 = 0$).

Deviations from the island model. We found that LR and GEE methods had higher false positive rates under deviations from the IM, particularly under patterns of isolation by distance. Such patterns probably resulted in spurious correlations of allele frequencies with the environmental variable. As a result, LR and GEE displayed a marked increase in the false positive rate under the stepping stone and HM. In practical applications, the use of a large number of neutral loci associated with a procedure to control false discovery is advisable. CWDRP was rather resistant, albeit not completely, to these effects and offered a satisfactory compromise between high rates of true positives and low rates of false positives. However, CWDRP works at the population level and requires estimation of allele frequencies, while LR and GEE can deal with genotype data and do not require extensive sampling at each location. Furthermore, the two latter methods can be extended to analyse relatedness among populations more accurately, for example using explanatory variables such as Moran's eigenvector maps, which use the spatial variation of data to control for spatial autocorrelation (Manel *et al.* 2010a,b; Schoville *et al.* 2012).

FG and (to a lesser extent) BN appeared to be robust to deviations from the IM (which is the underlying hypothesis of both tests). FG clearly requires a large number of populations, but we showed that 48 populations were sufficient to detect strong instances of selection. FG was the method with the lowest rate of false positives and was powerful under the best set of conditions (highest strength of selection and allogamy). However, FG lost a lot of power under weaker selection or when fewer populations were used (eight populations instead of 48).

With the HM, the BN test was essentially robust despite assuming a standard IM. The false positive rate was not higher with BN despite the deviation from the assumed IM (Figs 3 and 4, see also Fig. S2, Supporting Information). Instead, EHF appeared to be more powerful than BN in this case with lower selection intensity (Fig. S2, Supporting Information, third and fourth panels), although mostly for sampling modes S3–5). Excoffier *et al.* (2009) investigated cases in which between-population differentiation was stronger than between-cluster differentiation, and cases in which the reverse was the case. They found that analysing such data sets using the BN method (assuming a finite IM) led to inflated rates of false positives when between-cluster differentiation was stronger than within-cluster differentiation. The parameters of the HM we used led to stronger between-population differentiation than between-cluster differentiation, a situation to which BN might not be the most sensitive. Therefore, exacerbated between-cluster differentiation can invalidate the results of tests based on the IM. Such patterns can be predicted based on the biology of the species and otherwise can be detected from the data.

Deviations from the model assumed by FLK and VDB. FLK and VDB both underwent a striking loss of power under a few parameter combinations. VDB failed to detect any true positives under certain conditions: (i) selection was strong enough to frequently fix alternative alleles in the two populations being compared, and (ii) the demographic model caused sufficient population differentiation to render neutral allele fixation predictable. Under these conditions, simulations frequently resulted in the fixation of neutral alleles and yielded a non-significant test for selected alleles. This problem was partly caused by the limited amount of information conveyed by a single SNP marker assessed in only two populations, as VDB was originally designed for a higher number of alleles. One possible solution would be to merge information from adjacent markers to build more complex and refined genotypes, either by reconstructing haplotypes or by performing genome scans making use of linkage between markers. A similar problem occurred with FLK when populations on either side of the gradient inflection point were fixed or nearly fixed with respect to the allele concerned. In this case, the FLK statistic was within or close to the limit of the simulated values, yielding poor power estimates despite overwhelming evidence. It should be noted that VDB and FLK both allow for complex models of population history, far beyond those we simulated (population divergence with a bottleneck in the ancestral population for VDB and populations structured according to a phylogenetic tree). For example, FLK was designed for populations that diverged along a phylogenetic tree. In that case, this test was shown to perform better when information concerning the tree is taken into account (Bonhomme *et al.* 2010).

Population structure. Logistic regression does not account for the correlation between samples collected in the same population. Although GEE was designed to cope with the effect of within-population autocorrelation, the performance gain was partial with our simulation setting. GEE allows for a variety of models for within-population structures but further investigation is needed to understand their effects. However, we showed that strong within-population autocorrelations led to more contrasted performances by LR and GEE. We performed coalescent simulations with a lower migration rate scaled parameter $M = 4Nm = 0.01$, 0.5 and 4 (corresponding to $m = 0.000025$, 0.00125 and 0.01 in our forward simulation scheme; $m = 0.01$ is the value used in our main simulations). The results of these complementary simulations showed that GEE had a lower rate of false positives than LR only if migration was low ($M = 0.01$ and 0.5) under the IM (Fig. S3, Supporting Information). The SS caused a major increase in false positives with both the LR and GEE methods. This is because this model creates a spurious correlation between allele frequency and the environmental variable, as discussed earlier. GEE was severely affected, especially when the migration rate was low, but tended to outperform LR when the number of samples per population was high. LR performed best when only a few samples were collected per population. LR was almost always better than GEE for the highest rate of migration. Below a certain migration rate, populations might have been so highly differentiated that correlation between adjacent populations vanished. In both migration models, selfing had little effect, except with the highest migration rate and in LR, where it increased the rate of false positives, presumably by increasing within-population autocorrelation (Fig. S3, Supporting Information).

Regime of reproduction. Selfing appeared to have several effects on the performance of most tests, and some effects were quite drastic. The interplay between several interacting factors makes it difficult to decipher the mechanisms involved, but we can suggest several effects that are likely to occur in real-life applications. As we sampled both alleles per individual, and selfing has the property of increasing homozygosity, we expect that the effect of sampling modes will shift in data sets generated under selfing, with sampling modes with more samples being more efficient. However, this is not a major feature of our results, except for CWDRP, for which the effect of sampling modes was clearly opposite with allogamy with respect to selfing models (Fig. 6). With the exception of CWDRP with non-island models, selfing tended to increase the rates of false positives, but not necessarily much and chiefly with correlation-based methods. This could be a consequence of the increase in genetic variance between populations caused by a decrease in effective population size. Finally, rather striking losses of power were observed. In the case of FLK and VDB particularly,

we mentioned possible explanations earlier. The case of all differentiation-based methods in the IM (Fig. 4, bottom-left panel) is remarkable, despite good resistance by BN. The explanation might also be the decrease in effective population sizes and, hence, in the increase in relative between-population differentiation. The detection of outliers is based on the comparison of candidate loci with a distribution based on neutral loci. Whenever the basal level of differentiation is high, it will be obviously more difficult to detect outliers. The interaction between high levels of differentiation and selfing or other situations with low effective population size must therefore be dealt with.

Sampling schemes. Results obtained under the sampling scheme with many populations (S2) were generally better than those obtained with the sampling schemes with many samples per population (S3–5) in terms of both false and true positives. For LR, the sample including 100 populations with only one sample per population (S1) logically provided the best results. However, this sampling scheme is only applicable with the LR method. In some cases, however, samples with more populations enabled a decrease in false positive rates, probably due to more accurate estimation of allele frequency. There were a few cases in which S3 produced far fewer false positives than S2 (with FLK). With correlation-based methods, more populations are expected to result in better performances. Indeed, although more within-population samples allow better estimation of allele frequencies, investigating fewer populations makes the analysis more sensitive to high between-population variance. This is likely to be crucial with high rates of differentiation between populations (for example in the case of selfing), where individuals within populations are likely to be redundant. In that case, sampling a few individuals per population and many populations would allow a good estimate of overall allele frequency. To control for false positives with correlation-based methods, it would be wise to select a sampling scheme that avoids confounding autocorrelation effects, for example, by sampling distant populations or by investigating several unrelated environmental gradients (Poncet *et al.* 2010).

Conclusion

Our analysis revealed the effects of different biological and methodological factors on the performance of several methods used to detect natural selection. It would be impossible to address all possible combinations of factors, and our conclusions are necessarily limited to the demographic models, the level of differentiation and the form of selection we addressed. It should be noted that a wide range of phenomena exceeding the demographic factors we investigated in this study are likely to cause a correlation between genetic variation and environmental variables without direct selection. Examples include isolation by distance (Vasemägi 2006), interference between hitchhiking and structuration (Bierne 2011), coupling of incompatibility factors with dispersion barriers (Bierne *et al.* 2011) and genomic variation in the recombination rate (Roesti *et al.* 2012), among many others. For a further discussion of possible confounding factors and possible solutions, see Barrett & Hoekstra (2011). Nevertheless, general conclusions can be drawn from our analysis and previous similar studies. First, it is essential to clearly identify the hypotheses underlying statistical tests before applying them to data sets and to ensure that the data match these hypotheses. It should be noted that correlation-based methods can be very sensitive to genetic correlations between or within populations when these are not accounted for. LR and GEE run very fast, which is a great advantage in the current context of data generation techniques. However, these advantages come with the drawback of a high risk of spurious associations. Controlling for population structure is a prerequisite and should be combined with an efficient way to control for false positives (typically using control loci). Differentiation-based methods have the advantage of accommodating complex population history and structure when these are known, and they do not require environmental data. The low rate of false positives of some of the methods (even when the model does not exactly match reality) is also an advantage. Our results highlight the benefits of applying different methods jointly or serially. For example, Manel *et al.* (2009) used a correlation-based method to validate loci detected using differentiation-based methods. As we intended to examine the intrinsic properties of tests, we have not employed an FDR approach (Benjamini & Hochberg 1995), but this would be advisable for most applications, especially data sets with many markers.

In conclusion, our results allow us to formulate general guidelines for selecting a method to detect selection:

1. Differentiation-based methods have lower false positive rates but (with some exceptions) have the disadvantage of being rather time-consuming. They are less powerful than correlation-based methods but still perform well in detecting strong instances of selection.
2. As a corollary, the results of correlation-based methods should be interpreted with caution, especially in situations that lead to between-population correlations. In such case, recently developed statistical tools accounting for underlying correlation structure are recommended.
3. It is better to sample a few individuals in many populations than many individuals in a few populations. We found that eight samples per population are sufficient, whereas in real-life applications, 10 or more are preferable. However, the BN method is robust to the decrease in the number of populations. With the LR method, it is clearly better to sample only one individual per location.
4. Under allogamy and the IM, all methods are efficient and comparable.
5. Under allogamy and the stepping stone or the HM, differentiation-based methods are better thanks to their lower rate of false positives (FG is best).
6. Under selfing and the IM, the best option is to choose LR with sample S1 (one individual per population).
7. Under selfing and the stepping stone or the HM, the best methods are LR with S1, BN with S3 (few populations) and FG with S2 (many populations).

Acknowledgements

The authors thank Jean-Louis Pham, Mathieu Siol and Maxime Bonhomme for helpful discussion and three anonymous reviewers for their detailed and constructive suggestions. SDM was supported by the Agropolis Fondation. SM was supported by the Institut Universitaire de France. This study is part of the flagship project Agropolis Resource Center for Crop Conservation, Adaptation and Diversity (ARCAD) funded by the Agropolis Fondation.

References

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **263**, 1619–1626.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate–a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society series B-Methodological*, **57**, 289–300.

Bierne N (2011) The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. *Evolution*, **64**, 3254–3272.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

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

Bowcock AM, Kidd JR, Mountain JL *et al.* (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 839–843.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Cummings M, Clegg MT (1998) Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barley (*Hordeum vulgare* ssp. *spontaneum*): an evaluation of the background selection hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5637–5642.

De Mita S, Siol M (2012) EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genetics*, **13**, 27.

Endler JA (1986) *Natural Selection in the Wild*. Princeton University Press, Princeton.

Epperson BK, McRae BH, Scribner K *et al.* (2010) Utility of computer simulations in landscape genetics. *Molecular Ecology*, **19**, 3549–3564.

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

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Golding GB, Strobeck C (1980) Linkage disequilibrium in a finite population that is partially selfing. *Genetics*, **94**, 777–789.

Guillemaud T, Lenormand T, Bourguet D, Chevillon C, Pasteur N, Raymond M (1998) Evolution of resistance in

Culex pipiens: allele replacement and changing environment. *Evolution*, **52**, 443–453.

Hansen MM, Olivieri I, Waller DM, Nielsen EE, GeM Working Group (2012) Monitoring adaptive genetic responses to environmental change. *Molecular Ecology*, **21**, 1311–1329.

Jeffreys H (1961) *Theory of Probability*, 3rd edn. Oxford University Press, Oxford.

Joost S, Bonin A, Bruford MW *et al*. (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955–3969.

Keller LF, Waller DM (2002) Inbreeding effects in wild populations. *Trends in Ecology and Evolution*, **17**, 230–241.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Liu F, Charlesworth D, Kreitman M (1999) The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus *Leavenworthia*. *Genetics*, **151**, 343–357.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Manel S, Conord C, Després L (2009) Genome scan to assess the respective role of host-plant and environmental constraints on the adaptation of a widespread insect. *BMC Evolutionary Biology*, **9**, 288.

Manel S, Joost S, Epperson BK *et al*. (2010a) Perspective on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**, 3760–3772.

Manel S, Poncet BN, Legendre P, Gugerli F, Holderegger R (2010b) Common factors drive adaptive genetic variation at different scales in *Arabis alpina*. *Molecular Ecology*, **19**, 2896–2907.

McCullagh P, Nelder JA (1989) *Generalized Linear Models*. Chapman and Hall, London.

Mullen LM, Hoekstra HE (2008) Natural selection along an environmental gradient: a classic cline in mouse pigmentation. *Evolution*, **62**, 1555–1569.

Narum SR, Hess JE (2011) Comparison of F_{ST} outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**, 184–194.

Nei M, Maruyama T (1975) Lewontin-Krakauer test for neutral genes—comment. *Genetics*, **80**, 395.

Neuenschwander S, Hospital F, Guillaume F, Goudet J (2008) quantiNEMO: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. *Bioinformatics*, **24**, 1552–1553.

Nordborg M (2000) Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics*, **154**, 923–929.

Novembre J, Galvani AP, Slatkin M (2005) The geographic spread of the CCR5 Δ32 HIV-resistance allele. *PLoS Biology*, **3**, 1954–1962.

Pan W (2001) Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120–125.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Pérez-Figueroa A, García-Pereira MJ, Saura M, Rolán-Alvarez E, Caballero A (2010) Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology*, **23**, 2267–2276.

Poncet BN, Herrmann D, Gugerli F *et al.* (2010) Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Molecular Ecology*, **19**, 2896–2907.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Robertson A (1975) Lewontin-Krakauer test for neutral genes–comment. *Genetics*, **80**, 396.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.

Schoville SD, Bonin A, François O, Lobreaux S, Melodelima C, Manel S (2012) Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 23–43.

Siol M, Wright SI, Barrett SCH (2010) The population genomics of plant adaptation. *New Pythologist*, **188**, 313–332.

Teeter KC, Payseur BA, Harris LW *et al*. (2008) Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Research*, **18**, 67–76.

Tenaillon MI, Tiffin PL (2008) The quest for adaptive evolution: a theoretical challenge in a maze of data. *Current Opinion in Plant Biology*, **11**, 110–115.

Thornhill NW (1993) *The Natural History of Inbreeding and Outbreeding: Theoretical and Empirical Perspectives*. University of Chicago Press, Chicago, Illinois.

Vasemägi A (2006) The adaptive hypothesis of clinal variation revisited: single-locus clines as a result of spatially restricted gene flow. *Genetics*, **173**, 2411–2414.

Vilas A, Pérez-Figueroa A, Caballero A (2012) A simulation study on the performance of differentiation-based methods to detect selected loci using linked neutral markers. *Journal of Evolutionary Biology*, **25**, 1364–1376.

Vitalis R (2012) DETSEL: a R-package to detect marker loci responding to selection. In: *Data Production and Analysis in Population Genomics* (eds Bonin A, Pompanon F), pp. 277–293. Humana Press, New York, New York.

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

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

S.D.M., A.-C.T., L.G., N.A., J.R., and Y.V. planned the study. S.D.M. performed the analyses. S.D.M. and S.M. applied the LR and GEE methods and intepreted their results. S.D.M., J.R., and Y.V. analyzed all results and wrote the manuscript with contributions of all other authors.

Data accessibility

The data used in this study are the result of simulations. The parameters of these simulations as well as pieces of software and version numbers are fully described in the text. Tests used in the manuscript are implemented in software referred to in the text, except tests LR and GEE, which are based on a R program that is available as a supplementary data file.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Effect of environmental variables on the results of correlation-based tests.

Fig. S2 Comparison of BN and EHF tests with several sampling modes.

Fig. S3 False positive rate of logistic regression and generalized estimating equations methods estimated with coalescent simulations with low migration rates.

Data S1 Code of a R script performing LR and GEE tests for all loci of a given data set.