MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 2178–2192 doi: 10.1111/mec.12725

Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests

KATIE E. LOTTERHOS* and MICHAEL C. WHITLOCK

Department of Zoology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4

Abstract

F_{ST} outlier tests are a potentially powerful way to detect genetic loci under spatially divergent selection. Unfortunately, the extent to which these tests are robust to non-equilibrium demographic histories has been understudied. We developed a landscape genetics simulator to test the effects of isolation by distance (IBD) and range expansion on F_{ST} outlier methods. We evaluated the two most commonly used methods for the identification of F_{ST} outliers (FDIST2 and BayeScan, which assume samples are evolutionarily independent) and two recent methods (FLK and Bayenv2, which estimate and account for evolutionary nonindependence). Parameterization with a set of neutral loci ('neutral parameterization') always improved the performance of FLK and Bayenv2, while neutral parameterization caused FDIST2 to actually perform worse in the cases of IBD or range expansion. BayeScan was improved when the prior odds on neutrality was increased, regardless of the true odds in the data. On their best performance, however, the widely used methods had high false-positive rates for IBD and range expansion and were outperformed by methods that accounted for evolutionary nonindependence. In addition, default settings in FDIST2 and BayeScan resulted in many false positives suggesting balancing selection. However, all methods did very well if a large set of neutral loci is available to create empirical *P*-values. We conclude that in species that exhibit IBD or have undergone range expansion, many of the published F_{ST} outliers based on FDIST2 and BayeScan are probably false positives, but FLK and Bayenv2 show great promise for accurately identifying loci under spatially divergent selection.

Keywords: Bayenv2, BayeScan, FDIST2, FLK, F_{ST} outlier, genome scan, local adaptation, non-equilibrium, range expansion, refugia

Received 19 October 2013; revision received 13 March 2014; accepted 14 March 2014

Introduction

A major goal of evolutionary biology is to understand the molecular basis for adaptive differences between populations. F_{ST} outlier tests—tests that identify larger values of F_{ST} than expected by drift alone—have become a popular way of using genomic data to identify genes that have evolved under spatially-divergent selection.

F_{ST} is a standardized measure of the variance of allele frequencies among populations (Wright 1949). The foundation of the F_{ST} outlier test is to identify loci with F_{ST}s that are unusually high (divergent selection) or unusually low (balancing selection). However, nonselective evolutionary forces can affect the distribution of F_{ST} values from loci across a genome. Dispersal tends to reduce the differences among populations, while genetic drift increases differences on average. Most genes in the same genome experience these neutral processes relatively equally, and therefore, most neutral genes have approximately the same expected F_{ST}. However, in subdivided populations, by chance the measured F_{ST} can differ substantially from this expectation, causing even neutral genes to vary, sometimes substantially, in their F_{ST}s. F_{ST} outlier tests attempt to account for this neutral variation in F_{ST} and determine which

Correspondence: Katie E. Lotterhos, Fax: (336) 758-6008;
E-mail: lotterke@wfu.edu
*Present address: Department of Biology, Wake Forest University, Winston Salem, NC 27109, USA

loci have F_{ST} large enough or small enough to show significant evidence of selection. The challenge with outlier tests is to identify how much variation in F_{ST} among loci would be expected (i.e. the null distribution of F_{ST}) in the absence of selection.

In the original presentation of F_{ST} outlier tests by Lewontin & Krakauer (1973), the distribution of $(n_{demes}-1)F_{ST}/\bar{F}_{ST}$ was approximated by a χ^2 distribution. Difficulties with the method were immediately recognized because the variation in F_{ST} depends on sample sizes and the degree of independence of the evolutionary histories of sampled populations (Lewontin & Krakauer 1975; Nei & Maruyama 1975; Robertson 1975).

A number of F_{ST} outlier tests have been since been developed and used extensively for candidate gene discovery (Beaumont & Nichols 1996; Vitalis *et al.* 2001; Beaumont & Balding 2004; Foll & Gaggiotti 2008; Excoffier *et al.* 2009b; Bonhomme *et al.* 2010; Günther & Coop 2013). These tests fall into two general categories: methods that make strict assumptions about the demographic history of the samples and methods that estimate and account for evolutionary nonindependence among samples. In the first category are methods that (i) simulate a specific demographic history as a null distribution to test for significance (the island model of Beaumont & Nichols 1996 or the hierarchical model of Excoffier *et al.* 2009b) or (ii) assume that samples have diverged independently from a common ancestor (Bayesian methods assuming a multinomial Dirichlet distribution: Beaumont & Balding 2004; Foll & Gaggiotti 2008). In the second category are methods that estimate coancestry (Bonhomme *et al.* 2010) or covariance (Günther & Coop 2013) among populations and account for population structure in the test statistic.

The potential problem with methods that simulate a specific population history is that the results may be very sensitive to the specific history, and the true population history is rarely known with confidence. Assuming an island model (as in the FDIST2 methods of Beaumont & Nichols 1996), however, will often lead to a high number of false positives in real populations, because it results in a much narrower range of F_{ST}s than more complicated models of demographic histories [e.g. hierarchical model in humans (Excoffier *et al.* 2009b; Hofer *et al.* 2012), two-refugia model in pines (Eckert *et al.* 2010) and fractal networks in rivers (Fourcade *et al.* 2013)].

As an alternative to simulating a specific population history, a Bayesian method was developed (BAYESFST of Beaumont & Balding 2004; BayeScan of Foll & Gaggiotti 2008). The Bayesian method assumes that the gene frequencies under any neutrally structured population model can be approximated by a multinomial Dirichlet distribution (Beaumont 2005). The Dirichlet distribution describes the gene frequencies under a wide range of demographic models (Beaumont 2005; Charlesworth & Charlesworth 2010; pp. 341–350). Even though the Dirichlet distribution holds when sampled populations receive unequal number of migrants from the ancestral pool, these models assume that demes have evolved independently from an ancestral gene pool (Beaumont 2005; Excoffier *et al.* 2009b). Therefore, the Dirichlet distribution would not be appropriate if different samples are drawn from the same population, if some sampled populations share more recent ancestry than others, if there is unequal or recent migration among sampled populations or if there is a hierarchical population structure (Excoffier *et al.* 2009b).

Recently, two methods have been developed that relax the assumption that samples follow a particular demographic history. Bonhomme *et al.* (2010) developed an extension of the Lewontin–Krakauer test for structured populations, implemented in the program FLK. Their method uses a phylogenetic tree to estimate coancestry among samples and accounts for this coancestry in the calculation of the test statistic ($T_{F\text{-}LK}$), which follows a χ^2 distribution. Similarly, Günther and Coop calculate an F_{ST} analog called X^TX, which has been standardized by the covariance among populations. These methods explicitly adjust for evolutionary nonindependence among samples.

There have been a few independent simulation studies that have tested and compared current methods (Pérez-Figueroa *et al.* 2010; Narum & Hess 2011; Vilas *et al.* 2012; De Mita *et al.* 2013). Despite the fact that each study tested different scenarios (Table 1), simulations were typically conducted on small landscapes and few demographic histories were simulated. In addition, the choice of different criteria to assess significance among methods makes it difficult to compare error rates among methods on common ground (for example, the *P*-value cut-off used in FDIST2 was not statistically equivalent to the Bayes-factor cut-off used in BayeScan). So, although some previous simulation studies have found one method may outperform the other, this may be due to the different criteria that were used to assess significance. In the present study, we compare methods on common grounds by transforming probabilities (*P*-values or Bayes factors) to *q*-values (Storey & Tibshirani 2003) and use the same *q*-value cut-off to control for the false discovery rate (FDR) at the same level.

In this study, we compare the false-positive rates and power of four leading methods to detect loci that have differentiated by spatially heterogeneous selection: FDIST2, BayeScan, FLK and BAYENV2. We focus on methods that do not require environmental data or do not assume that the researcher knows a priori which environmental axes are important. We chose these methods because they are either in wide use (FDIST2

Table 1 A summary of simulation studies that have examined false-positive rates and power of various F_{ST} outlier tests

Reference	No. neutral loci	No. selected loci (directional, divergent)	Marker type	Landscape	No. demes	N_e/deme	No. replicate data sets	No. generations	Control for mean F_{ST}?
Pérez-Figueroa *et al.* (2010)	1000, 9990, 9970, 9950, 9900	0, 10, 30, 50, 100	AFLP/dominant	Island Model	2	500	10	Theoretical distribution	Yes
Narum & Hess (2011)	95	5	QTL	10 × 1 grid	10	500	1	2000	No
Vilas *et al.* (2012)	Depended on mutation rate (1400–33 000)	1–10	QTL	Island Model	2	500	100	10 000	Not stated
De Mita *et al.* (2013)	750	1	SNP	10 × 10 grid	100	100	100	5000	No
This study	10000, 9900, 9000	0, 100, 1000	SNP	360 × 360 grid	129 600	Varied	3	Varied	Yes

and BayeScan) or are recent and show real promise (FLK and BAYENV2).

We extend previous analyses comparing these methods in several ways. First, we explore both equilibrium and nonequilibrium scenarios with isolation by distance (IBD), as well as the island model for a baseline. In particular, we draw attention to the biologically common but statistically problematic case of recent range expansion. We model the expansion of a species from one or two refugia, mimicking the post-glacial expansion of temperate species or the expansion of a species into a new geographical area after a vicariance event. Such cases, while common in nature, are poorly understood in terms of the genetic patterns they can create (Hewitt 2000), and the performance of F_{ST} outlier tests in such cases has never been explored. In such cases, sampled populations have not yet reached equilibrium, and many loci may have experienced allele surfing during demographic expansion (Klopfstein *et al.* 2006; Excoffier *et al.* 2009a). Moreover, these scenarios exhibit IBD and are not easily grouped into discrete populations. In this case, the island model is obviously not exact, and it may be difficult to determine the correct hierarchical structure (e.g. Narum & Hess 2011).

We consider these methods under three different scenarios, depending on the type of data that is available. First, we consider what we refer to as the 'default settings' case, where there is no a priori information about which loci in the data set are likely to be under selection or not. This is perhaps the main case considered by existing methods, and to analyse such data, we use the default settings of each program.

In the second scenario, researchers may have a set of loci that were chosen a priori to be less likely to be directly affected by selection, by creating a set of genetic markers that are not in coding sequences or in any likely regulatory region. These putatively neutral loci can be used to parameterize the neutral null model in the following programs: FDIST2 requires the mean F_{ST} of neutral loci; FLK needs the *F*-matrix of coancestry; and BAYENV2 estimates the among-population covariance matrix Ω. Each of the methods is in principle improved if these parameters are estimated from truly neutral data without the influence of the selected loci. We call this scenario 'neutral parameterization'. In addition, BayeScan requires the user to input the prior odds that a locus is neutral, and we investigate the effects of varying this parameter from the default.

Finally, in the third scenario, if the set of neutral loci is large enough, the statistical significance of a putatively selected outlier locus could be assessed by its quantile in the empirical distribution of differentiation measures obtained from the neutral set. We therefore also explore this 'empirical *P*-value' approach.

We find that IBD and range expansion can cause some methods to have large false-positive rates and FDRs, but the two newer methods are much improved in this regard. In most cases, the methods can be improved with neutral parameterization, and FDRs for all methods are very low with the empirical *P*-value approach. Detecting true F_{ST} outliers is possible with modern methods, but for species with IBD, many of the loci previously detected in the literature are likely false positives.

Methods

Landscape simulations

We developed a haploid landscape genetics simulator in the R (R Core Team 2013) and C programming languages. The simulator used recurrence equations to model the evolution of a single biallelic locus on a quasi-continuous square landscape composed of 129 600 (360 × 360) demes. Our landscape simulator is designed to efficiently simulate a species with large effective population sizes and widespread geographical ranges, differing from previously developed simulators in this regard [e.g. Nemo (Guillaume & Rougemont 2006) or Simupop (Peng & Kimmel 2005)]. Because we modelled large populations, we assumed that linkage decayed rapidly in the genome and that several thousand independent SNPs could be ascertained. Each haploid locus was simulated independently, with a starting allele frequency randomly chosen from a uniform distribution between 0 and 1. Source code for the landscape simulator is located in the Dryad repository (doi: 10.5061/dryad.v8d05), and datasets are available from the authors upon request. Details of the simulator are provided in the Supporting information, Appendix S1.

We modelled four demographic histories, as described below. For each of the four scenarios, we chose parameters and the time of sampling that would result in approximately the same overall F_{ST} equal to ~0.05. This was an important step, because it causes the F_{ST} distribution of all loci to reflect differences in demographic history rather than differences in overall genetic differentiation (see Table 1 for a summary of studies that controlled for mean F_{ST}). Thus, we controlled for an easily measured parameter (mean F_{ST}) and varied aspects of demographic history that are more difficult to estimate.

Demographic histories and dispersal models

We simulated four population histories. Note that we chose different parameters for each demographic history such that they would all give the same mean F_{ST} (Table 2). This is because mean F_{ST} is easily estimated, but other details of the demographic history of a population typically remain more obscure. The first three demographic histories we implemented all exhibited isolation by distance, but differed in their time from equilibrium (one at approximate equilibrium vs. two nonequilibrium scenarios), and the fourth was the island model. For the first three scenarios, dispersal was determined by a discretized version of a Gaussian dispersal kernel with standard deviation $\sigma = 1$ km, (Fig. S1, Supporting information). The four scenarios are as follows, with details given in Table 2 and the Appendix S1 (Supporting information):

1. *Equilibrium isolation by distance* (IBD-eq). The landscape was started at carrying capacity and was run until equilibrium, with migration implemented via the dispersal kernel.
2. *Nonequilibrium isolation by distance due to expansion from one refugium* (1R) *or two refugia* (2R). Refugia were located in the southern portion of the landscape (Fig. S2, Supporting information). Simulations were run long enough for populations to fill the landscape and reach carrying capacity (demographic equilibrium), but not long enough for them to reach genetic equilibrium. (Note that these demographic expansion models at genetic equilibrium would simply be alternative examples of the IBD model.)
3. *Island model at equilibrium* (IM). The landscape was started at carrying capacity and was run until equilibrium, with a migration rate of 0.01 among demes. Although unrealistic, the island model should meet the assumptions of the outlier methods that we tested (see Methods: Outlier tests).

Selected loci

Selection acted on juvenile survival under a variety of different environmental spatial patterns. Our simulations assumed that demographic dynamics were independent of the strength of selection (i.e. soft selection). We attempted to simulate a set of selection coefficients that would be representative of that observed in a real genome. This is difficult since the true distribution of selection coefficients is unknown; we inferred that (i) more loci have weak effects than strong effects, (ii) some environmental axes will have more loci adapting to them than others and (iii) environments differ in their degree of spatial heterogeneity. Thus, for each replicate data set, selected loci evolved under 17 different environmental patterns (see 'Methods: Replication of data sets'). Details of the generation of the environments are presented in the Appendix S1 (Supporting information).

Table 2 A summary of parameter values used in the *IM*, *IBD-eq*, *1R* and *2R* simulations. All landscapes were of equal spatial extent (270 × 270 km). Columns show the number of demes (No. demes), the deme size, the carrying capacity per deme (K/deme), the carrying capacity per square km (K/km^2), the intrinsic growth rate of a deme (r), the number of generations the simulation was run for (No. gens), how dispersal was modelled, whether or not the simulation was at equilibrium when it was stopped ('Genetic eq.?') and the generation at which demographic equilibrium was reached (i.e. when all demes on the landscape had reached carrying capacity)

Demography	Landscape size (demes)	Deme size	K/deme	K/km^2	r	No. gens	Dispersal	Genetic eq.?	Demographic eq. at gen
Island model (*IM*)	72 × 72	3.75 km × 3.75 km	936	67	NA	5000	m = 0.01	Yes	1
Isolation by distance (*IBD-eq*)	360 × 360	0.75 km × 0.75 km	9	16	NA	10 000	σ = 1 km	Yes	1
Expansion from one refugium (*1R*)	360 × 360	0.75 km × 0.75 km	40	71	0.4	1000	σ = 1 km	No	653
Expansion from two refugia (*2R*)	360 × 360	0.75 km × 0.75 km	70	124	0.4	1000	σ = 1 km	No	512

We assumed that the strength of selection in deme i (s_i) was determined by the standardized environment in that deme: $s_i = \lambda\ x_i$ where λ describes the strength of selection on the landscape and x is the standardized environment in deme i. With this function, selection was positive in environmental values greater than the mean environment, neutral at the mean environment and negative at values less than the mean. The parameter λ equals $2s/(4\sigma_{ENV})$, where s is the landscape-level selection coefficient and σ_{ENV} is the standard deviation of the environment. This function essentially resulted in patches on the landscape with the highest (lowest) environmental values having a selection coefficient of approximately s ($-s$). We modelled four selection strengths $s \in \{0.001, 0.005, 0.01, 0.1\}$ at proportions of {40%, 30%, 20%, 10%}, respectively.

In each simulation, we quantified local adaptation as the correlation between allele frequencies and the environment across the entire landscape (Pearson's ρ). The simulations included many loci of weak effect, and weakly selected loci do not always contribute to local adaptation (because, for example, migration or drift may in some case be strong enough to prevent selective differentiation: Yeaman & Otto 2011). For all of the results in this study, therefore, we assess the power of tests to detect only the loci that have contributed most to local adaptation (loci with $\rho > 0.4$).

Replication of data sets

For each demographic history, we simulated three replicate data sets. Each data set had a set of independent, randomly generated environments (i.e. Fig. S3, Supporting information) to which selected loci adapted. We distributed the 1000 selected loci for each data set across environments unequally (Appendix S1, Supporting information).

Each replicate set of loci consisted of an independent set of 10 000 neutral loci and 1000 selected loci for each set of environments. From the loci simulated for each replicated data set, we created three nested data sets by sampling without replacement: 10 000 neutral (10 000-N, 0% under selection), 9900 neutral and 100 selected (9900-N:100-S, 1% under selection), and 9000 neutral and 1000 selected (9000-N:1000-S, 10% under selection).

Simulated sampling of genetic data

Sampling of allele frequencies occurred after migration, with samples from 75 populations randomly distributed on the landscape (Fig. S4, Supporting information). Twenty individuals were sampled from each population, for a total data set of 1500 individuals. This sampling design gave a good approximation of the F_{ST} of the entire landscape (Fig. S5, Supporting information).

Evaluation of false-positive rate, true-positive rate and FDR

The false-positive rate is the number of significant neutral loci (false positives) divided by the total number of neutral loci tested. The true-positive rate (power) is the number of loci correctly determined by the method to be under selection divided by the number of selected loci tested. The FDR is the number of false-positive neutral loci divided by the total number of positive results (Fig. S6, Supporting information).

For each case, results were transformed to q-values to correct for multiple comparisons. We used a q-value cut-off of 0.01 to define a positive result (Storey & Tibshirani 2003). The q-value of a locus can be described as the expected proportion of false positives among all loci with P-values equal to or less than the observed locus (Storey & Tibshirani 2003). Therefore,

loci that have q-values equal to 0.01 should have an expected FDR of 1%. At this significance level for a data set with 100 F_{ST} outliers, we would only expect one locus to be a false-positive neutral locus and the other 99 to be under selection. At this significance level for a data set with no truly selected loci in the data, we expect only one false-positive neutral locus out of every 100 neutral-only data sets. We also refer to the q-value cut-off as the 'stated FDR'. While we do not advocate the use of a strict cut-off for deciding whether or not a locus is worthy of further scrutiny, we use a cut-off here to simplify the comparison between demographies and methods.

Outlier tests

We only tested methods that could be used for our simulated data (i.e. we excluded the method of Excoffier *et al.* (2009b) because our data lacked obvious hierarchical structure and that of Fariello *et al.* (2013) because this method is for haplotype data). We tested BayeScan (Foll & Gaggiotti 2008), FDIST2 (Beaumont & Nichols 1996), FLK (Bonhomme *et al.* 2010) and the F_{ST} analog X^TX from BAYENV2 (Günther & Coop 2013). For BayeScan, FDIST2 and FLK, we evaluated the programs under three scenarios of available data: 'default settings', 'neutral parameterization' and 'empirical P-values' (described in more detail below). As BAYENV2 does not implement a significance test for X^TX, it could only be compared to the other programs in specific scenarios (ranking loci under the default settings, and 'empirical P-values'). For most analyses, we focus on comparing error rates in the right tail, but we also show that these methods create many false-positive loci under balancing selection in the left tail of the F_{ST} distribution. For each case, probabilities (or Bayes Factors) were transformed to q-values to test for significance (Storey & Tibshirani 2003; Muller *et al.* 2006).

BayeScan of Foll & Gaggiotti (2008) was implemented with version 2.1 of the software (provided at http://cmpg.unibe.ch/software/BayeScan/). Unlike the other methods, BayeScan cannot be parameterized with a neutral set of loci. We explored the capacities of BayeScan further by manipulating 'the prior odds of neutrality' parameter, which is the prior probability of a locus being under selection in the data set. The default value for prior odds in the program is 10 (for every 10 neutral loci in the data set, odds are that 1 locus is under selection).

We implemented the FDIST2 method in R following Beaumont & Nichols (1996) using some of the machinery of the landscape simulator, so that we could control mean F_{ST} and explore some of the simulated F_{ST} distributions [which are not output by LOSITAN (Antao *et al.* 2008)]. Details of the implementation are in the Appendix S1 (Supporting information). The P-values from our implementation were highly correlated with results from LOSITAN (Fig. S7, Supporting information). The source code for the R implementation of FDIST2 is located in the Dryad repository (doi: 10.5061/dryad.v8d05).

We implemented FLK with the R code provided at https://qgsp.jouy.inra.fr (accessed on 15 December 2013) and the main workflow shown in Figure 11 of Bonhomme *et al.* (2010). For each locus, we used the ancestral allele frequency from the simulations as the outgroup. FLK is implemented in two steps: in the first step, the F-matrix is estimated from the allelic data. In the second step, the F-matrix is then used to compute the $T_{F\text{-}LK}$ statistic and get P-values for each loci.

We calculated X^TX for each data set with the BAYENV2 software (Günther & Coop 2013). BAYENV2 is also implemented in two steps: in the first step, the variance–covariance matrix is calculated from allelic data. We used the variance–covariance matrix from the final run of the MCMC after 10^5 iterations. In the second step, the variance–covariance matrix is used to control for evolutionary history in the calculation of X^TX for each SNP (again using 10^5 iterations of the MCMC). Note that there is no current implementation for calculating P-values from X^TX: significance must be based on rankings or empirical P-values.

Default settings. First, we tested the common case for which no prior information separates neutral loci from possible selective loci. We assume that the user has a list of loci and is attempting to determine which of these are possibly under selection. We used the default settings of the programs, which means that the programs parameterize their null distributions from all of the data. In the default case, which yielded the highest false-positive rates, we also evaluated how calling significance based on ranking the top 1% of loci would affect error rates ('Ranks').

Neutral parameterization. Second, we assumed that the user has a set of putatively neutral loci, perhaps identified as markers which occur at locations outside coding or probable regulatory regions and that have a lower a priori probability of being under strong section. In this scenario, the user obtains important information about the population from the putatively neutral set of loci. For FDIST2, this gives a less biased estimate of the mean F_{ST} of neutral loci (used to parameterize simulations of the island model), and for FLK and BAYENV2, these neutral loci can be used to estimate the **F** (coancestry) matrix or Ω covariance matrix, respectively.

BayeScan cannot be parameterized with a neutral set; so, for this program, we examined the effects of varying

the prior odds of neutrality. We varied these odds, comparing 10 (the default), 100, 1000 and 10 000.

Empirical P-values. Finally, we also considered the case where data from a very large number of neutral loci were available, such that the distribution of measures of differentiation (F_{ST} from FDIST2, α from BayeScan, $T_{F\text{-}LK}$ from FLK or X^TX from BAYENV2) could be estimated from the very large set of neutral genes (>1000). We refer to this as the empirical *P*-value approach, because the *P*-value of a test for the null hypothesis of neutrality could be obtained for a locus by comparison with the null distribution provided by this set. Note that in this scenario, we first used neutral parameterization for each method and then calculated empirical *P*-values.

Results

Distributions of F_{ST} for neutral and selected loci

Despite having similar mean F_{ST}, the four demographic histories exhibited different distributions of F_{ST} across neutral loci (Fig. 1, right column). For neutral loci, the island model had the lowest variance in F_{ST}, followed by the *IBD-eq*, *1R* and *2R* models (Fig. 1, right column). Figure 1 also shows example landscapes for false-positive neutral loci (left column). In the refugia models, clines in allele frequencies on the landscape often arose by neutral processes.

The distribution of selected loci was affected by demographic history, the strength of selection and the landscape of environmental variation on which selection occurred. Demographic history affected the distribution of selected loci partially because the efficacy of selection was not equal across demographies (Figs S8 to S12, Supporting information). Note that this also affected the power to detect selected loci in each demography.

Detecting selective differentiation: default settings

We first focus on detecting highly differentiated loci with the default settings of the methods. Demographic history affected false-positive rates in BayeScan, FDIST2 and FLK. (BAYENV2 could not be compared because it does not include a significance test for X^TX.) For all programs, false-positive rates were lowest for the *IM*, higher in *IBD-eq*, higher yet in *1R* and highest for *2R* (Fig. 2A–C). In all scenarios, BayeScan gave the highest false-positive rates, and FLK had the lowest false-positive rates, with FDIST intermediate between the two programs. For all programs, false-positive rates were unacceptably high in the refugia scenarios: as high as 5–15% for BayeScan and FDIST2 and as high as 1–3% for FLK.

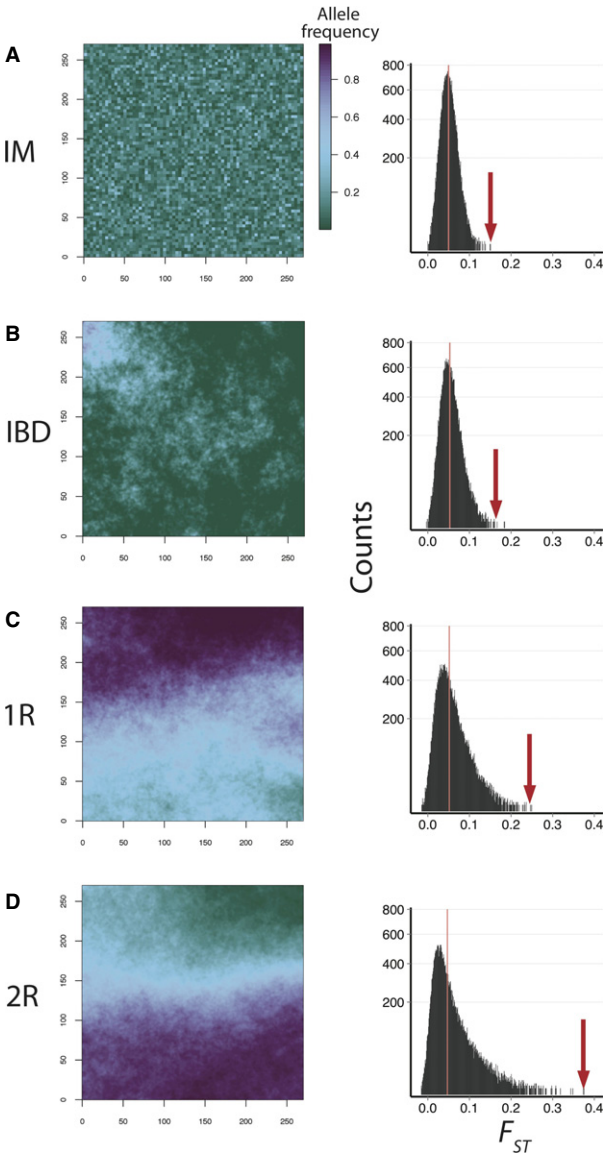

Fig. 1 Left column: Example landscapes for an outlier neutral locus at the end of the simulation for: (A) island model, (B) isolation by distance, (C) expansion from one refugium and (D) expansion from two refugia with secondary contact. Axes on the landscape represent latitude and longitude position in km. Right column: F_{ST} distributions for all neutral loci, for 75 random samples on the landscape. Each demography had approximately the same mean F_{ST}, as indicated by the vertical line. Note that the *y*-axis is on a square-root scale, so that the tails of the distribution can be more easily compared. Small arrows indicate the F_{ST} of the sample landscape shown in the left column.

True-positive rate (power) was generally similar for BayeScan and FDIST2 in all scenarios and slightly lower for FLK (Fig. 2D, E). Although FLK had slightly lower power, fewer false positives from this method led to a much lower FDR in most scenarios (Fig. 2F–H). FDIST2

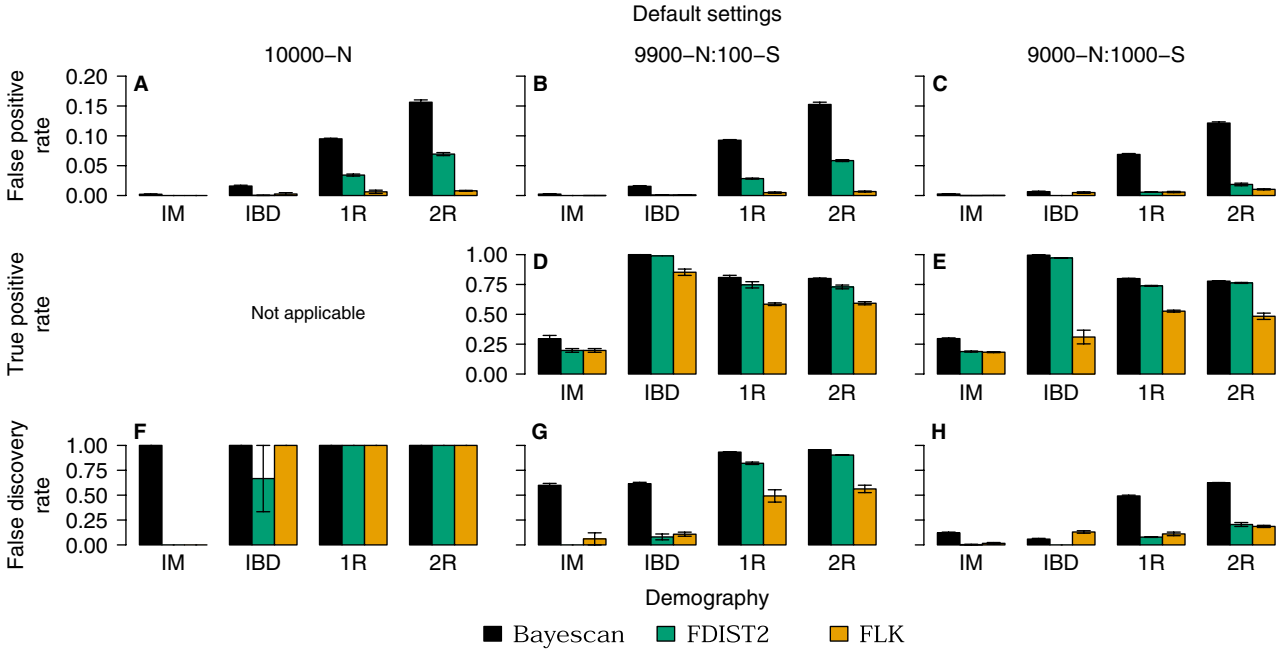

Fig. 2 False-positive rates, power and false discovery rates (FDR) for the default settings in BayeScan (prior odds = 10), FDIST2 and FLK. Rates are based on *q*-values and a stated FDR of 0.01. Note the *y*-axis scale for false-positive rate is the same in Fig. 3, but larger than in Figs 5 and 6. (A,F) 10 000 neutral loci; (B,D,G) 9900 neutral loci and 100 selected loci; (C,E,H) 9000 neutral loci and 1000 selected loci. IM, island model; IBD, isolation by distance; *1R*, expansion from one refuge; *2R*, expansion from two refugia. Error bars are standard errors. False-positive rates for FDIST2 and FLK were zero in the island model, which led to a 0% FDR.

had a similar or lower FDR than FLK in the *IM* and *IBD-eq* demographies, due to the higher power of FDIST2 in these demographies. Note that the FDRs decrease with the percentage of selected loci in the data set, because there are more true positives in the data set to be detected.

To compare these programs to the program BAYENV2, we used rank order of the X^TX or F_{ST} for each locus, as suggested by Günther & Coop (2013). By this approach, a locus was flagged as significant if its differentiation statistic was in the highest 1% of all loci. Figure S13 (Supporting information) shows that all programs give similar performance based on ranks—but that using ranks is undesirable because error rates and power are dependent on the number of selected loci in the data set.

Detecting selective differentiation: neutral parameterization

Here, we consider error rates in highly differentiated loci (the right tail of the F_{ST} distribution), for the case when a set of presumed neutral loci are used to provide parameters such as the mean F_{ST} or the pattern of coancestry (FDIST2, FLK or BAYENV2) or when the prior odds on neutrality was manipulated (BayeScan).

BayeScan. BayeScan's default value for the prior odds for numbers of neutral:selected loci is 10. Using more realistic values of the prior odds for neutrality (100–10 000) resulted in decreased false-positive rates without a large change in power. Figure 3 shows that the performance of BayeScan improved with increasing prior odds, regardless of the true odds in the data set. The prior odds of neutrality also affected the posterior distribution of F_{ST}s calculated by BayeScan, although this also depended on the extent to which demographic history violated the assumptions of BayeScan (Fig. S14, Supporting information). Prior odds had very little effect on the posterior distribution of F_{ST} for loci under strong selection, because information provided by the data was so strong that the prior had no more influence (Fig. S14, Supporting information). However, when the data violated the assumptions of BayeScan (as in the refugia cases), even neutral loci could remain as outliers and were unaffected by the prior odds (Fig. S14, Supporting information). For the comparison among methods under neutral parameterization (below), we used a prior odds of 10 000 for BayeScan, because this had the lowest error rates—regardless of true odds in the data.

FDIST2. For *IBD-eq* and the refugia models using FDIST2, neutral parameterization resulted in an increase

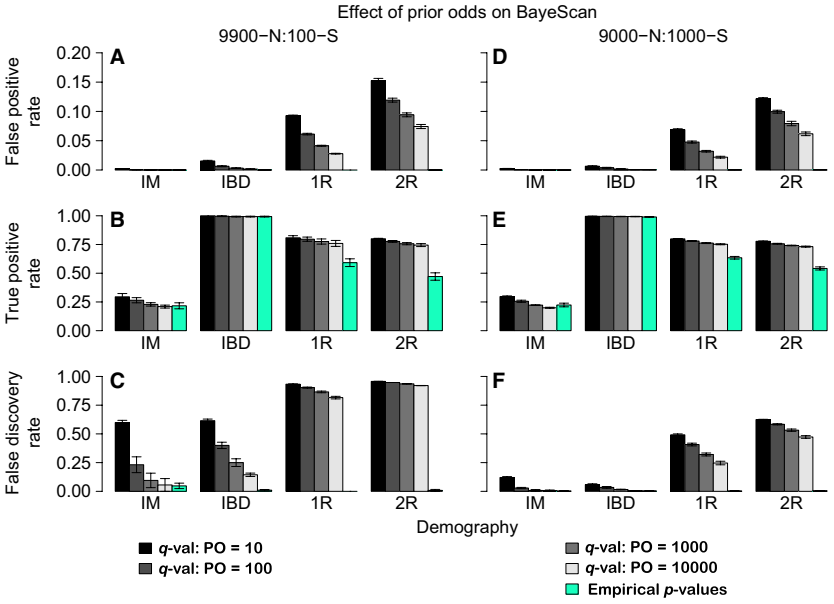

Fig. 3 Effect of prior odds on false-positive rates, power and false discovery rates (FDR) from BayeScan. Error rates were based on *q*-values (with a stated FDR of 0.01), while manipulating prior odds (PO) are based on empirical *P*-values (see Methods). False-positive rates based on empirical *P*-values were <0.002. Note the *y*-axis scale for false-positive rate is the same in Fig. 2, but larger than in Figs 5 and 6. (A–C) 9900 neutral loci and 100 selected loci; (D–F) 9000 neutral loci and 1000 selected loci. IM, island model; IBD, isolation by distance; *1R*, expansion from one refuge; *2R*, expansion from two refugia. Error bars are standard errors.

of false positives in the right tail of the F_{ST} distribution (Fig. S15, Supporting information). This counter-intuitive result occurred because the island model had a narrower distribution of F_{ST}s than the true demographic history (Compare Fig. 1A to 1B–D). Figure 1A represents the island model simulated by FDIST2 under neutral parameterization, while the island model simulated under the default settings had the same distribution, but the mean 'neutral' F_{ST} was biased upwards by accidental inclusion of selected loci. When compared to neutral parameterization, the default settings had the effect of *decreasing* false positives in the right tail (Fig. S15, Supporting information) and increasing false positives in the left tail (Results: Error rates in the left tail: false-positive rates for balancing selection).

FLK and Bayenv2. For FLK, we found that neutral parameterization had a small but positive effect on performance in the data sets with 1% selection and a larger and more significant effect on performance in the data sets with 10% selection (Fig. S16, Supporting information). This same comparison could not be made for Bayenv2, because the method does not employ a significance test for X^TX. FLK and Bayenv2 both depend on neutral genes to calculate the pattern of correlation or coancestry among populations, and this in principle can be more accurate if only truly neutral loci are used to calculate the coancestry or covariance matrices for these methods.

We evaluated the effect of unwanted inclusion of selected loci by measuring the correlation between coancestry or covariance matrices estimated with all the data to those estimated with neutral data only. A higher correlation means that the selected loci in the data set did not bias the estimation of the coancestry/covariance matrix. Bayenv2 tended to be less sensitive to bias caused by inclusion of selected loci than FLK for more complex demographic histories, while FLK only outperformed Bayenv2 for the island model (Fig. 4). The effect of selection on the coancestry matrix from FLK and the covariance matrix from Bayenv2 are shown in Fig. S17 (Supporting information).

Comparison among methods. As explained above, neutral parameterization decreased error rates for BayeScan and FLK, but increased error rates for FDIST2 (Fig. 5; note the *y*-axis on the top row of Fig. 5 is half that of Fig. 2). This resulted in FDIST2 showing similar performance to BayeScan in terms of false-positive rates,

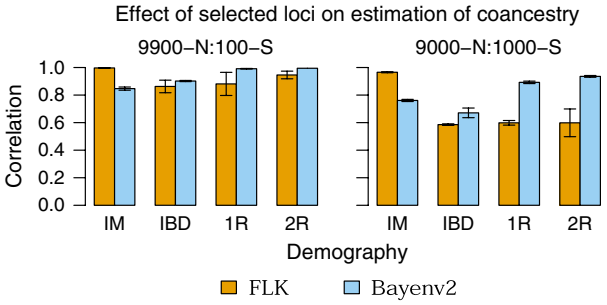

Fig. 4 The correlation between the coancestry (covariance) matrix using both neutral and selected loci, and a coancestry (covariance) matrix using only neutral loci in FLK (Bayenv2). For more complicated demographic histories (IBD, *1R* and *2R*), Bayenv2 had a higher correlations between neutral and non-neutral covariance matrices than did the coancestry matrices estimated from FLK. Error bars are standard errors.

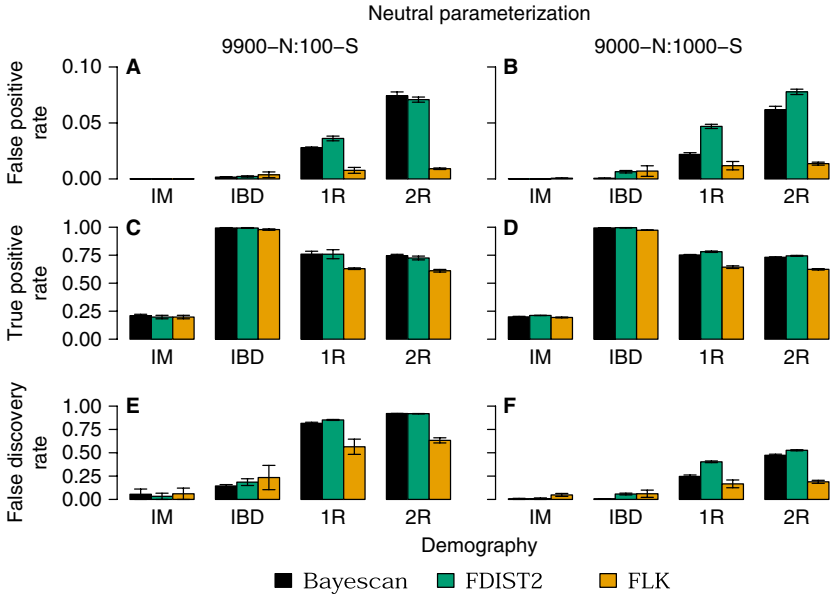

Fig. 5 False-positive rates, power and false discovery rates (FDR) for neutral parameterization in BayeScan (prior odds = 10 000), FDIST2 and FLK. Rates are based on *q*-values and a stated FDR of 0.01. Note that the *y*-axis for false-positive rate is half of that in Figs 2 and 3. (A,C,E) 9900 neutral loci and 100 selected loci; (B,D,F) 9000 neutral loci and 1000 selected loci. IM, island model; IBD, isolation by distance; *1R*, expansion from one refuge; *2R*, expansion from two refugia. Error bars are standard errors.

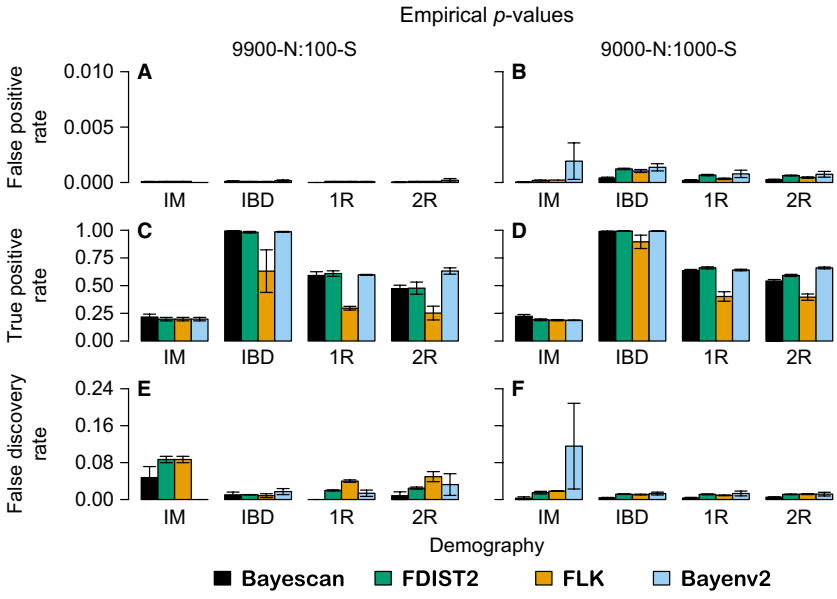

Fig. 6 (A) False-positive rates, power and false discovery rates (FDR) for the empirical *P*-values calculated from BayeScan (α), FDIST2 (F_{ST}), FLK ($T_{F\text{-}LK}$) and BAYENV2 (X^TX). Empirical *P*-values were transformed to *q*-values, and rates are based on a stated FDR of 0.01. Note that the *y*-axis for false-positive rate is an order of magnitude less, and the axis on FDR is one quarter less than similar figures (Figs 2, 3 and 5). (A–C) 9900 neutral loci and 100 selected loci; (D–F) 9000 neutral loci and 1000 selected loci. IM, island model; IBD, isolation by distance; *1R*, expansion from one refuge; *2R*, expansion from two refugia. Error bars are standard errors.

power and FDRs (Fig. 5). Again, FLK had much lower false-positive rates and FDRs, with only slightly lower power. However, false-positive rates were still undesirably high in the refugia models, even for FLK (Fig. 5A, B).

Detecting selective differentiation: empirical P-values

Empirical *P*-values are calculated by comparing the differentiation values for candidate loci to a distribution of those values for putatively neutral loci. Note that we are considering a perfect application of the empirical *P*-value approach, because the loci in our neutral set are known without error to be neutral. The true power of this approach will be lessened in real situations where some of the putatively neutral set are in fact experiencing the effects of selection.

The empirical *P*-value approach results in the lowest false-positive rates of all the approaches we have tested: less than or equal to 2 in 1000 (Fig. 6). All programs had similar false-positive rates (Fig. 6A, B). BayeScan, FDIST2 and BAYENV2 had similar power, and FLK had slightly lower power. Figure S17 (Supporting information) illustrates that the covariance matrix estimated by BAYENV2 is detecting much finer-scale population structure than the coancestry matrix estimated by FLK, which may be the source of the difference in power.

Error rates in the left tail: false-positive rates for balancing selection

As we only simulated loci under divergent selection, any positive test in the left tail of the distribution of F_{ST} or X^TX is a false positive. These are loci that are typically inferred to be under balancing selection. We evaluated false-positive rates in the left tail using default settings, neutral parameterization and empirical *P*-values (Fig. 7; BAYENV2 is only shown for the third scenario).

With default settings, false-positive rates in the left tail were undesirably high for all methods in *IBD-eq*, *1R* and *2R* scenarios (Fig. 7A, B). For data sets with 1% selection (9900-N:100-S), BayeScan typically had the highest false-positive rates in the left tail for all demographies (Fig. 7A). For data sets with 10% selection (9000-N:1000-S), false positives in FDIST were off the charts—as high as 0.72 in *IBD-eq* (Fig. 7B). This occurred because the estimate of the mean neutral F_{ST} was biased upwards, creating many false positives in the left tail (Results: Detecting selective differentiation: neutral parameterization: FDIST2). For data sets with 10% selection, false positives generally increased for BayeScan and FLK with more complex demographic histories (Fig. 7B).

Neutral parameterization almost eliminated false positives in the left tail for BayeScan and FLK and greatly reduced them for FDIST2 (Fig. 7C, D). Empirical *P*-values resulted in very low false-positive rates in the left tail for all programs (<0.001; Fig. 7E, F).

Discussion

In this study, we investigated how the most common approaches for F_{ST} outlier tests would perform under equilibrium and nonequilibrium demographic scenarios. We found that all methods were sensitive to demographic history, to neutral parameterization and to the percentage of selected loci in the data set. Under *IBD-eq* and range expansion, the older, widely used methods (FDIST2 and BayeScan) showed very high false-positive rates for loci apparently under both divergent and balancing selection, and these were the most sensitive to neutral parameterization. In species with these demographic histories, many of the published F_{ST} outliers inferred from FDIST2 and BayeScan are probably false positives. Newer methods (FLK and BAYENV2) show great promise.

Effect of demographic history

The increase in error rates corresponded to the increase in the variance of the F_{ST} distribution (compared to the assumed island model) in the demographies we tested. Why was the variance in F_{ST} so much larger in the refugia scenarios? As a result of nearby samples sharing a recent evolutionary history, there were effectively fewer independent populations in the data, which causes the variance of the distribution of F_{ST} to be higher than predicted.

The increase in variance of the F_{ST} distribution due to population structure and correlated evolutionary history has been recognized previously (Excoffier *et al.* 2009b; Eckert *et al.* 2010; Hofer *et al.* 2012; Fourcade *et al.* 2013), but the magnitude of the effect is more pronounced with the realistic cases newly considered here. FDIST2 and BayeScan found many false positives for both divergent and balancing selection, because the assumption of evolutionary independence among samples was violated in the *IBD-eq*, *1R* and *2R* scenarios. In the nonequilibrium scenarios, FLK had much lower error rates than FDIST2 or BayeScan without a significant loss in power, probably because the method estimates and adjusts for coancestry among samples. False-positive rates for FLK were still unacceptably high (~1%) for the nonequilibrium scenarios, even with neutral parameterization. (For every 10 000 neutral tests, a false-positive rate of 1% equates to 100 false positives,

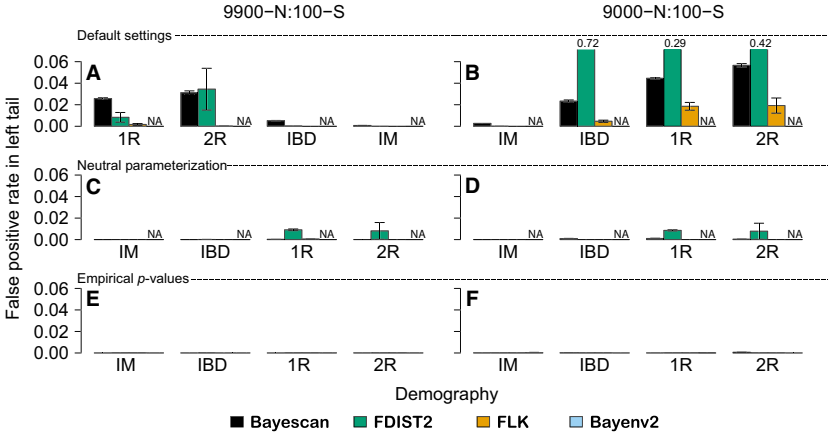

Fig. 7 False-positive rates in the left tail of the F_{ST} distribution. As we did not simulate any loci under global balancing selection, any positive test in the left tail was a false positive. (A,B) Default settings; (C,D) neutral parameterization, (E,F) empirical *P*-values. Note that BAYENV2 is only evaluated for empirical *P*-values. In (B), error rates for FDIST2 are indicated by text. All *y*-axes are on the same scale. IM, island model; IBD, isolation by distance; *1R*, expansion from one refuge; *2R*, expansion from two refugia. Error bars are standard errors. NA, not applicable.

which may be more than the number of truly selected loci and confuse later analyses.)

Shifting ranges and nonequilibrium populations are expected to be common in nature. Therefore, landscape genomic studies for F_{ST} outliers should also include some analysis or demonstration of whether the population is close to equilibrium. For example, with range expansions, genetic diversity may decline with distance from the refuge (Austerlitz *et al.* 1997), and we observed a decline in H_e with latitude in the *1R* and *2R* data sets (not shown). Additionally, a new statistic for detecting range expansions from genetic data (Peter & Slatkin 2013) may be useful in determining the location of the origin of the expansion. If a system shows evidence of being out of equilibrium, all the F_{ST} outlier tests we evaluated (when significance was not based on empirical *P*-values) were likely to have false positives in >1% of tests, and these programs should be applied with caution to genome-scale data.

We found that the observed FDR (the number of false positives divided by the total number of positive tests) was over 90% in some cases, despite a stated FDR of 1%. The FDR is high because there are too many false positives in these analyses, as discussed above. A very large fraction of loci detected by FDIST2 or BayeScan, and an uncomfortable number of loci detected by FLK, may in fact be false positives in species that have undergone recent range expansion.

Defaults and the effect of neutral parameterization

Under default values, BayeScan had the highest false positives under directional and balancing selection of all programs tested. Note that this result is the opposite of that from all previous simulation studies, which concluded BayeScan had low rates of false positives (Pérez-Figueroa *et al.* 2010; Narum & Hess 2011; Vilas *et al.* 2012; De Mita *et al.* 2013). As noted in the introduction, none of these studies compared methods on common statistical grounds, and these studies either simulated conditions that would meet the assumptions of BayeScan (as in Pérez-Figueroa *et al.* 2010 and Vilas *et al.* 2012) or simulated more realistic conditions but with small data sets (100 loci in Narum & Hess 2011; 750 loci in De Mita *et al.* 2013). Increasing the prior odds in BayeScan vastly reduced the false-positive rate without affecting power, but in the refugia models, false-positive rates were much higher than FLK.

Using neutral loci to parameterize the null models in FDIST2, FLK and Bayenv2 had significant effects on error rates from all programs. Neutral parameterization will be particularly important to supplement RNA-Seq data or a candidate gene data set, because we expect these data to be enriched for loci that have experienced selection in their evolutionary history.

For FDIST2, neutral parameterization exacerbated error rates in nonequilibrium scenarios: this occurred because the island model erroneously predicted narrower variance than the true distribution of neutral F_{ST}. With spatial autocorrelation of allele frequencies, the null island-model-distribution of F_{ST} can underestimate the probability that a neutral allele may be an outlier. Procedures that essentially 'cull' outlier F_{ST}s to better estimate neutral mean F_{ST} (such as the 'Use "neutral" mean F_{ST}' option implemented in LOSITAN: Antao *et al.* 2008) suffer similarly if the underlying simulated model does not perfectly match the empirical data set (Fig. S18, Supporting information shows that the effect of the 'Use "neutral" mean F_{ST}' option in LOSITAN is qualitatively similar to neutral parameterization for the *2R* model). This problem can be improved using a more realistic null model (such as the hierarchical model: Excoffier *et al.* 2009b; Hofer *et al.* 2012; or a two-refugia model: Eckert *et al.* 2010); however, it is unlikely that the empirical F_{ST} distribution can be simulated perfectly, especially for samples exhibiting isolation by distance on a landscape.

For FLK, neutral parameterization became increasingly important for controlling error rates as the percentage of selected loci in the data set increased. FLK also requires an outgroup that can be sequenced at the same loci, which might hinder its application in some systems where no obvious outgroup is available. The covariance matrix from Bayenv2 seems to be relatively less sensitive to the percentage of selected loci than FLK, but Bayenv2 does not implement a significance test for X^TX, and so, it can only be compared to FLK in specific cases (discussed below).

Ranks and empirical P-values

Given the false positives in the refugia scenarios, one might be tempted to ignore the test results and simply follow up on these loci that are most extreme in terms of F_{ST} or other differentiation statistic. This is a good approach in some cases, but it is also subject to a potentially high rate of false positives. However, we caution against this approach, because false-positive rates and FDRs based on rank orders are too dependent on the number of true positives in the data set.

The best implementation of the F_{ST} outlier test for all programs was to calculate empirical *P*-values. Note that the empirical *P*-value approach requires 1000s of putatively neutral loci—determined a priori—to create a null distribution on which all loci in the data set may be tested. For example, a set of putatively neutral SNPs could be obtained by restriction enzyme digests, which

are less specific in targeting coding regions (Davey *et al.* 2011; Elshire *et al.* 2011). In this study, we implemented the empirical *P*-value approach with perfect knowledge of which loci were neutral. In practice, this approach would probably have slightly lower power due to the inclusion of some selected loci. In our results, BAYENV2 had slightly higher power than the other programs under the more realistic demographies.

Power of F_{ST} outlier tests may depend on genetic architecture

We only simulated single loci under selection, and power may be lower for loci affecting polygenic traits under selection. Polygenic adaptation may involve small allele frequency changes at many loci, and power to detect these may be low with outlier tests (Le Corre & Kremer 2003; Mackay *et al.* 2009; Pritchard & Di Rienzo 2010; Le Corre & Kremer 2012). On the other hand, if genetic architectures evolve with fewer, larger and more tightly linked divergent alleles (Yeaman & Whitlock 2011), these large-effect loci would be F_{ST} outliers because their allele frequencies are divergent among environments.

Conclusions and summary of recommendations

Given that it can be easy to associate a false-positive outlier with a putative metabolic or developmental function (Pavlidis *et al.* 2012), it can be costly for follow-up studies to examine a list of loci that contain many false positives. F_{ST} outlier tests have low error rates when samples can be considered to have diverged independently from an ancestral population or have diverged with equal dispersal connections among all populations. This assumption will usually be met when there are only two sampled populations in the data set (corresponding to an island model with two demes). For most types of landscape genomic data, however, significance of outliers should be viewed with caution, because genetic correlations among samples may result in a violation of the assumptions of model-based F_{ST} outlier tests. Cases of nonindependence in gene frequencies may be most extreme in populations that have expanded from a refugium but have not yet reached equilibrium. Thus, F_{ST} outlier analyses should be accompanied by tests for equilibrium.

When possible, having a set of loci from genomic regions likely to be neutral can be a powerful tool. A neutral data set can significantly improve the accuracy of the null models used by these methods, especially when there are a large number of genes experiencing selection. We recommend the use of FLK when no neutral loci or a small set of neutral loci are available, although it still has an undesirably high false-positive rate (greater than ~1%) in nonequilibrium scenarios. With a large neutral data set, if possible, empirical *P*-values can be extremely accurate, and BAYENV2 had the highest power under IBD and nonequilibrium histories. When implementing the empirical *P*-value approach, it is imperative to decide a priori which loci will be used to create the null distribution.

We advocate that test statistics (F_{ST}, $T_{F\text{-}LK}$, X^TX) and significance statistics (*P*-values, Bayes factors and/or *q*-values) for all loci in an empirical or simulation study—not just the significant ones—be archived in a data repository or published as Supporting information. This will facilitate comparisons across studies and reanalysis of data sets when better methods become available.

Acknowledgements

We would like to thank Sam Yeaman, Loren Rieseberg, Sally Aitken, Armando Moreno Geraldes, Jeremy Draghi, Florence Débarre, Holly Kindsvater, Kimberly Gilbert, Kaylee Byers and the reviewers for many useful comments on this manuscript and helpful discussions. This research is part of the AdapTree Project, funded by Genome Canada, Genome BC, Alberta Innovates Bio Solutions, the Forest Genetics Council of British Columbia, the BC Ministry of Forests, Lands and Natural Resources Operations, Virginia Tech, the University of British Columbia and the University of California, Davis, and it has been partially supported by a Discovery Grant from the National Sciences and Engineering Research Council (Canada).

References

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a F_{ST}-outlier method. *BMC Bioinformatics*, **9**, 323.

Austerlitz F, Jung-Muller B, Godelle B, Gouyon P-H (1997) Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical Population Biology*, **51**, 148–164.

Beaumont MA (2005) Adaptation and speciation: what can F_{ST} tell us? *Trends in Ecology and Evolution*, **20**, 435–440.

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 B: Biological Sciences*, **263**, 1619–1626.

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

Charlesworth B, Charlesworth D (2010) *Elements of Evolutionary Genetics*. Roberts and Company, Greenwood Village, Colorado, USA.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

De Mita S, Thuillet AC, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

Eckert AJ, van Heerwaarden J, Wegrzyn JL *et al.* (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*, **185**, 969–982.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Excoffier L, Foll M, Petit RJ (2009a) Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 481–501.

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

Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013) Detecting signatures of selection through haplotype differentiation among hierarchically-structured populations. *Genetics*, **193**, 929–941.

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.

Fourcade Y, Chaput-Bardy A, Secondi J, Fleurant C, Lemaire C (2013) Is local selection so widespread in river organisms? Fractal geometry of river networks leads to high bias in outlier detection. *Molecular Ecology*, **22**, 2065–2073.

Guillaume F, Rougemont J (2006) Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, **22**, 2556–2557.

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

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hofer T, Foll M, Excoffier L (2012) Evolutionary forces shaping genomic islands of population differentiation in humans. *BMC Genomics*, **13**, 107.

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482–490.

Le Corre V, Kremer A (2003) Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics*, **164**, 1205–1219.

Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, **21**, 1548–1566.

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

Lewontin RC, Krakauer J (1975) Testing the heterogeneity of F values. *Genetics*, **80**, 397.

Lotterhos KE, Whitlock MC (2014) Data from: evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology*, doi: 10.5061/dryad.v8d05.

Mackay TF, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, **10**, 565–577.

Muller P, Parmigiani G, Rice K (2006) FDR and Bayesian multiple comparisons rules. Johns Hopkins University, Dept. of Biostatistics Working Papers.

Narum SR, Hess JE (2011) Comparison of F_{ST} outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**(Suppl 1), 184–194.

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

Pavlidis P, Jensen JD, Stephan W, Stamatakis A (2012) A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, **29**, 3237–3248.

Peng B, Kimmel M (2005) simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, **21**, 3686–3687.

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.

Peter BM, Slatkin M (2013) Detecting range expansions from genetic data. arXiv:1303.7475.

Pritchard JK, Di Rienzo A (2010) Adaptation–not by sweeps alone. *Nature Reviews Genetics*, **11**, 665–667.

R Core Team (2013) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Robertson A (1975) Remarks on the Lewontin-Krakauer test. *Genetics*, **80**, 396.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences USA*, **100**, 9440–9445.

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

Wright S (1949) The genetical structure of populations. *Annals of Human Genetics*, **15**, 323–354.

Yeaman S, Otto SP (2011) Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution*, **65**, 2123–2129.

Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. *Evolution*, **65**, 1897–1911.

Both authors worked together to establish the experimental design and to develop the landscape simulator. K.E.L. wrote the code and did the analyses. Both authors contributed to writing the manuscript.

Data accessibility

The code for the landscape simulator and the FDIST2 implementation are stored in the Dryad digital repository: doi: 10.5061/dryad.v8d05.

Supporting information

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

Appendix S1 Methods.

Fig. S1 The probability of dispersal as a function of distance from the focal deme in the centre was based on a discretized Gaussian distribution with a standard deviation of $\alpha = 1$ km and deme size = 0.75 × 0.75 km.

Fig. S2 The starting location of populations on the landscape for the one-refugium (*1R*) and two-refugia (*2R*) demographies.

Fig. S3 Selective environments for the first replicate of data sets (a new set was generated for each replicate of data sets).

Fig. S4 Sampling scheme on the landscape used for all data sets.

Fig. S5 F_{ST} for an infinite sample from all demes on the landscape vs. F_{ST} for the sample of 20 individuals from each of 75 locations.

Fig. S6 Definition of false-positive rate, true-positive rate and false discovery rate.

Fig. S7 *P*-values from the FDIST2 implementation in R vs. *P*-values from LOSITAN.

Fig. S8 F_{ST} distributions with selected loci.

Fig. S9 The correlation between allele frequencies and the environment across the landscape vs. F_{ST} for the island model.

Fig. S10 The correlation between allele frequencies and the environment across the landscape vs. F_{ST} for isolation by distance.

Fig. S11 The correlation between allele frequencies and the environment across the landscape vs. F_{ST} for the one-refuge model.

Fig. S12 The correlation between allele frequencies and the environment across the landscape vs. F_{ST} for the two-refugia model.

Fig. S13 Error rates based on ranking the top 1% of loci in each data set (i.e. top 100 loci).

Fig. S14 Effect of prior odds parameter on the distribution of F_{ST} from BayeScan for the island model (IM) and expansion from two refugia (*2R*).

Fig. S15 False-positive rates, power and false discovery rates based on different parameter-izations in FDIST2: default settings (using the entire data set to estimate mean F_{ST}); neutral parameterization (using only neutral loci to estimate mean F_{ST}); and empirical *P*-values (using a large neutral set to calculate empirical *P*-values of all loci in the data set based on F_{ST}).

Fig. S16 False-positive rates, power and false discovery rates based on different parameterizations in FLK: default settings (using the entire data set to estimate the coancestry matrix F_{ij}); neutral parameterization (using only neutral loci to estimate the coancestry matrix F_{ij}); and empirical *P*-values (using a large neutral set to calculate empirical *P*-values of all loci in the data set based on the statistic T_{F-LK}).

Fig. S17 Examples of coancestry and covariance matrices estimated by FLK and Bayenv2, respectively, for the 10 000-N and 9000-N:1000-S data sets.

Fig. S18 Effect of using the 'Use "neutral" mean F_{ST}' option in LOSITAN on power (TPR: true-positive rate) and false-positive rates (FPR) in both tails of the F_{ST} distribution.