MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 2301–2314 doi: 10.1111/j.1365-294X.2010.04641.x

When can ecological speciation be detected with neutral loci?

XAVIER THIBERT-PLANTE and ANDREW P. HENDRY

Redpath Museum and Department of Biology, 859 Sherbrooke St. West, McGill University, Montréal, QC, Canada H3A 2K6

Abstract

It is not yet clear under what conditions empirical studies can reliably detect progress toward ecological speciation through the analysis of allelic variation at neutral loci. We use a simulation approach to investigate the range of parameter space under which such detection is, and is not, likely. We specifically test for the conditions under which divergent natural selection can cause a 'generalized barrier to gene flow' that is present across the genome. Our individual-based numerical simulations focus on how population divergence at neutral loci varies in relation to recombination rate with a selected locus, divergent selection on that locus, migration rate and population size. We specifically test whether genetic differences at neutral markers are greater between populations in *different* environments than between populations in *similar* environments. We find that this expected signature of ecological speciation can be detected under part of the parameter space, most consistently when divergent selection is strong and migration is intermediate. By contrast, the expected signature of ecological speciation is not reliably detected when divergent selection is weak or migration is low or high. These findings provide insights into the strengths and weaknesses of using neutral markers to infer ecological speciation in natural systems.

Keywords: adaptive radiation, divergent selection, ecological speciation, F_{ST}, gene flow, individual-based simulation, neutral markers, porous genome

Received 29 September 2009; revision received 22 March 2010; accepted 26 March 2010

Introduction

Ecological speciation is a process whereby divergent selection causes the evolution of reproductive barriers between populations adapting to different environments (Schluter 2000). This process has been confirmed in a number of natural systems (Rundle & Nosil 2005), but recent work has sometimes failed to detect its presence (for a review, see Hendry 2009). This apparent variation in progress towards ecological speciation could be the result of biological factors that promote or constrain adaptive divergence and reproductive isolation (Berner *et al.* 2009; Hendry 2009; Nosil *et al.* 2009a), or it could be the result of ethodological limitations (Räsänen & Hendry 2008; Hendry 2009). Here, we use numerical simulations to consider limitations that might attend one common method for inferring progress towards ecological speciation. Our main motivation is to provide information relevant to empiricists studying ecological speciation in natural systems.

We specifically evaluate the use of neutral genetic markers to test the prediction that gene flow is lower between populations in *different* environments than between populations in *similar* environments (e.g. Smith *et al.* 1997; Gíslasan et al. 1999; Lu & Bernatchez 1999; Ogden & Thorpe 2002; Crispo *et al.* 2006; Nosil *et al.* 2008; Berner *et al.* 2009). This prediction has its origin in expectations that stronger ecological reproductive barriers should reduce gene flow between populations in different environments and thereby allow greater genetic divergence at neutral markers (Barton & Bengtsson 1986). This basis for inference has, however, been brought into question by the realization that alleles at neutral markers unlinked to selected loci might flow almost freely between populations even in different environments (Emelianov *et al.* 2004; Gavrilets &

Correspondence: Xavier Thibert-Plante, Fax: +1 865 974 9461; E-mail: xavier@nimbios.org

Vose 2005; Thibert-Plante & Hendry 2009). And yet, ambiguity remains because a 'generalized barrier to gene flow' across the genome might be possible (Gavrilets 2004; Grahame *et al.* 2006; Via & West 2008; Nosil *et al.* 2009a; Via 2009; Feder & Nosil 2010) if selection acts against the whole genome of migrants and hybrids. That is, linkage disequilibrium before recombination between genomes of individuals from different populations will potentially reduce gene flow even at unlinked neutral markers.

Here, we must make a distinction between the above 'generalized barrier' approach and the use of genome scans to identify outlier loci associated with ecological differences between populations (Emelianov *et al.* 2004; Grahame *et al.* 2006; Via & West 2008; Via 2009; Nosil *et al.* 2009a; Feder & Nosil 2010). This latter method is useful for identifying regions of the genome that show low gene flow between populations in different environments, and are therefore likely under divergent selection. By contrast, the generalized barrier approach asks whether or not reduced gene flow between populations in different environments can be detected at non-outlier loci that are presumably 'far' from the selected loci: i.e. neutral loci not physically linked to loci under divergent selection. Insights into divergence at these unlinked neutral loci will help the interpretation of non-outlier loci in genome scans, as well as inferences from classic population genetic studies that focus exclusively on those loci.

Our specific goal was to evaluate the conditions under which environmental differences between populations can cause detectable divergence at neutral loci. We examine the influence of several potentially important parameters, including migration rate, the strength of divergent selection, recombination rate and population size. First, we expect that intermediate migration rates will allow the greatest in neutral genetic divergence between populations in different environments compared with that between populations in similar environments (henceforth, the 'expected signature' of ecological speciation). The reasons for this prediction are that: (i) under *low* migration rates, all populations will diverge approximately equally owing to drift (because gene flow is too low to constrain divergence), whereas (ii) under *high* migration rates, divergent selection will not be very effective at reducing divergence in neutral markers (as gene flow can remain high even after selection against migrants and hybrids). Second, greater divergent selection will cause greater ecologically based reproductive barriers (Schluter 2000; Nosil *et al.* 2008, 2009a,b; Thibert-Plante & Hendry 2009), and so we expect that it will also allow greater divergence at neutral loci. Third, we expect that lower recombination rates (e.g. closer physical linkage) between neutral and selected loci will cause greater divergence in the former through the process of genetic hitchhiking (Maynard Smith & Haigh 1974; Charlesworth *et al.* 1997). In this respect, we also expect that selection needs to be much stronger than recombination if it is to reduce gene flow (Spirito *et al.* 1983; Bengtsson 1985). Fourth, we expect that larger population sizes will more reliably generate the expected signature of ecological speciation because selection tends to overwhelm drift in such populations (Whitlock & Phillips 2000).

The above theoretical expectations have not been thoroughly and systematically evaluated in the context of ecological speciation. Previously, Thibert-Plante & Hendry (2009) used a simulation model to consider some of the above issues, but the present analysis is much more comprehensive and more closely tied to empirical situations (e.g. neutral markers in the present study are similar to microsatellites). Here we also more comprehensively explore the parameter ranges under which the expected signature can be most reliably detected using common statistical methods. This latter topic is important because, even when a generalized barrier exists, it might be so weak and variable as to elude detection in a typical empirical study.

Model

We simulate diploid individuals in a multilocus, multiallele model, in which one locus is under selection and the other 40 loci are neutral. The locus under selection has two possible alleles, and the allele that is selectively favoured differs between the two environments. Four populations exchanging migrants are modelled, with two populations per environment type (Fig. 1). Other types of population structure could have been modelled, but this is the simplest for examining patterns across a large range of parameter space. A single selected locus with only two alleles was chosen for simplicity; future work might profitably examine the effects of distributing selection across multiple loci (Feder & Nosil 2010).

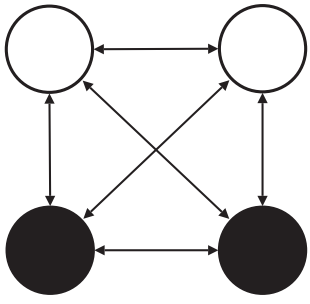

Fig. 1 Population structure in our simulations: four different populations in similar and different environments (colours) exchanging migrants (arrows).

Life cycle

The simulated life cycle begins with migration and then proceeds to reproduction. Selection occurs during reproduction (but before recombination) and is manifest as differences in reproductive output according to the relative fitness of individuals in a soft selection situation (see below). Explicit viability selection is thus absent from our model, but the lack of reproduction by an individual is evolutionarily equivalent to its death. Note that, as in nature, selection can act on migrant genes both before (on migrants) and after (on hybrids) recombination. Migration occurs evenly across all populations (the same number of individuals migrating from each population is distributed among each of the other populations). A given individual can move only once during the life cycle, and the individuals that move from each population are chosen randomly.

Mating proceeds by randomly selecting two individuals from a population with probabilities according to their relative fitness within that population. Fitness is dependent on the interaction between an individual's genotype and its environment ($G\times E$), according to the rules stated in Table 1. The probability of an individual i being selected for reproduction at any time is

$$p_i = \frac{f_i}{\sum_j f_j} \quad (1)$$

where f_i is the fitness of individual i and $\sum_j f_j$ is the sum of the fitnesses of all individuals in the population. A selected pair of individuals produces one offspring and then returns to the potential pool of parents. This procedure is repeated until the offspring population reaches the size of the parental population (i.e. soft selection, Whitlock 2002). Reproduction then stops and the life cycle starts over again with migration. Generations are thus non-overlapping.

Neutral genes

The neutral loci act like microsatellites, with high mutation rates (Weber & Wong 1993) (in the order of 10^{-3}) that change the number of repeats in a sequence (Valdes *et al*. 1993; Di Rienzo *et al*. 1994). Alleles are identified by their number of repeats. The number of repeats is limited to be between 1 and 999 (as in Balloux & Goudet 2002), which is larger than the number of alleles that populations should carry at equilibrium (Kimura & Ohta 1975). Mutations are stepwise and consist of an increase, or a decrease, of one in the number of repeats (Kimura & Ohta 1975). Stepwise mutations at the boundaries (1 and 999) always produce the only permitted adjacent value (2 or 998 respectively), but these are unlikely to ever be reached. We also ran the full set of simulations with some multistep mutations (Di Rienzo *et al*. 1994). Results were similar and so these additional simulations are not shown.

We vary the recombination rate of neutral loci with the selected locus. For a recombination rate of 0.5, the loci are unlinked to the selected locus and can be considered to be on different chromosomes. Within each of the 10 replicates of a given set of simulation conditions (defined by a given migration rate, selection and population size), we include 400 neutral loci – 40 with each of the 10 recombination rates to the selected locus. This inclusion of 400 loci in a single simulation reduced computation time by a factor of 10 relative to whether we had run a separate simulation for each recombination rate. In all cases, all neutral loci within a simulation are unlinked to each other, which makes them appropriate replicates of a given recombination rate within a given simulation set. Note that number of loci per analysis of a given parameter set was always 40 (not 400) because this was the number of loci per recombination rate. Although natural genomes include neutral loci of varying degrees of recombination with each other, we did not implement this variation in our simulations. The reason was that varying recombination rates among neutral loci would reduce their value as independent replicates of divergence within a given simulation. Moreover, the most important question is what happens when neutral loci are unlinked to selected loci, in which case they might also be unlinked to each other. Indeed, empirical studies typically discard neutral loci that are in linkage disequilibrium so as to consider the loci that are independent of each other.

Table 1 The absolute fitness of each genotype in the two different environments, where $0 \le \sigma \le s \le 1$

Selected locus	Fitness in environment 1	Fitness in environment 2
aa	1	$1-s$
aA or *Aa*	$1-\sigma$	$1-\sigma$
AA	$1-s$	1

Simulation set-up

At the start of a simulation, two random alleles are allocated to each individual at the selected locus. This leads to Hardy–Weinberg equilibrium with expected allele frequencies of 0.5 at the selected locus in each population. The strength of natural selection on this locus is set at s, the proportional decrease in fitness for the disfavoured allele (relative to unity for the favoured allele) in each environment. Allelic effects are additive, such

that heterozygotes have intermediate fitness ($1-\sigma$, where $\sigma = s/2$) (Table 1). In each life cycle iteration thereafter, migration between the populations occurs as a fixed proportion m of the population. That is, each population contributes (and receives) an expected Nm individuals split equally among (and from) the other populations. Population size (N) is the same in each population and is maintained at a constant level (as consistent with soft selection). We use a binomial random number generator to obtain the actual Nm in each generation (Kachitvichyanukul & Schmeiser 1988). With this technique, we achieve a non-null number of migrants, on average, even at low migration rates, while retaining the expected Nm.

Simulations are run for 5000 iterations (i.e. generations) and F_{ST}, according to Weir (1996), is tracked for each locus at each step of the simulation. We used 10 replicate simulations for each parameter set: strength of selection (s = 0–1.0, in increments of 0.1), migration rate ($m = \{0.5, 0.3, 0.2, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$), recombination rate ($r = \{0, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5\}$) and population size ($N = \{100, 1000\}$). All of the above parameter combinations are explored and are independent, except for the above-noted inclusion of all recombination rates in a given simulation.

Different neutral loci are unlinked to each other and potentially linked to the same selected locus but with different recombination rates, as described above. Initially, the 400 loci (10 different recombination rates with 40 loci each) are all set at 500 repeats, the middle point in the possible number of repeats (as in Balloux & Goudet 2002). We did not explore the effect of varying numbers of loci owing to computational limitations and because these effects should be relatively small (see Discussion).

As noted above, the mutation rate for neutral loci is 10^{-3}. The mutation rate of the locus under selection is 10^{-5}. These mutation rates are the same as those used in other numerical studies (Gavrilets & Vose 2005, 2007; Gavrilets *et al.* 2007) and are similar to some estimates from real organisms (Dallas 1992; Brinkmann *et al.* 1998). We did not explore other mutation rates because this allowed us more computational resources for exploring the other key parameters. We chose per-patch population sizes of 100 and 1000 to bracket those used in related numerical simulations (Gavrilets & Vose 2005, 2009).

Genetic measures were taken, and statistical tests (see below) performed, after reproduction but before migration. F_{ST} values were calculated for each set of parameter combinations based on the entire population. In addition, empiricists generally subsample from larger populations, and so we also calculated F_{ST} values (for simulations with the larger population size) after subsampling only 20 individuals. This particular sub-sampling level was chosen to be the same as that of Balloux & Goudet (2002). Results were generally similar between the full and subsampled estimates and so the former are reported in the present paper. An exception is an explicit comparison of the two estimates that serves to illustrate their general similarity and the slight differences (see below).

Statistical tests

Many statistical analyses of genetic differences are computationally intensive, such as bootstrapping F_{ST} 10 000 times to get confidence limits or implementing analyses in STRUCTURE 2.2 (Pritchard *et al.* 2000). We estimated that applying such analyses to all of our simulations would take more than a century of computation on a 2-GHz computer. We therefore implemented several short cuts.

Our main short cut was simply to use only 100 bootstrap replicates for comparisons of F_{ST}. Specifically, we calculate F_{ST} 100 times based on independent random samples with replacement of individuals (N = population size) from each population. We then conclude that F_{ST} is greater between populations in similar environments than between populations in different environments if the 95% confidence intervals do not overlap (one-tailed test). Given our four populations, this generated six comparisons per simulation: two for populations in similar environments and four for populations in different environments. We treated these as independent estimates to calculate the proportion of times, in each replicate simulation, where comparisons of populations in different environments yield a greater F_{ST} than comparisons of populations in similar environments.

We also evaluated whether our short cut of using 100 replicate F_{ST} bootstraps (rather than the more typical 10 000 replicates) caused any bias in interpretation. For this, we compare confidence intervals from our 100 bootstraps to confidence intervals from 10 000 bootstraps in three separate simulation conditions. The three conditions are chosen for their different range of F_{ST}; their parameters are: ($s = 0.1$, m=10^{-3}), ($s = 0.5$, $m = 10^{-3}$) and ($s = 0.5$, $m = 10^{-6}$). These comparisons are made over 50 consecutive generations after the first generation, after 1000 generations, and after 4950 generations. The results of these different iterations and simulations are combined for analysis.

Another short cut was to use STRUCTURE (Pritchard *et al.* 2000) for a subset of the simulations. Here, we used the *ad hoc* criteria of Evanno *et al.* (2005) to test for how many discrete populations were inferred in a given simulation. Specifically, we use the admixture model with the degree of admixture (alpha) inferred from the

data and with the option of correlated allele frequencies between populations. The distribution of allele frequencies (lambda) is set to unity and the length of the burn-in period is 10 000. Twenty replicates are run for each data set and the number of populations evaluated are from one to 10. For this analysis, we use the same simulations as those used for the above 100 vs. 10 000 bootstrap comparisons. A clear signature of ecological speciation in STRUCTURE implies finding two populations with individuals within a population coming from a single environment. Failure of these conditions implies the absence of, or a failure to detect, progress towards ecological speciation.

Results

To illustrate general patterns seen at the end of our simulations (4900–5000 generations), we focus first on results for larger populations: 1000 individuals in each of the four populations. We later discuss any differences seen in the simulations with 100 individuals per population.

The selected locus

Dynamics at the selected locus are straightforward, showing a migration–selection balance (Fig. 2). At low selection and high migration, fixation of the positively selected allele is rare and the negatively selected allele can sometimes fix (Fig. 2a). As selection increases and migration decreases, fixation of the positively selected allele becomes more common, and this essentially always occurs at very low migration rates. When fixation of the positively selected allele occurs, the fitness reduction in the new environment will be s for migrants and σ for hybrids.

Divergence at neutral markers

At low migration rates ($m = 10^{-4}$), genetic divergence is strong ($F_{ST} \approx 0.22$) between all populations regardless of whether they are in similar or different environments (Fig. 3a,b). This level of divergence is very close to the expected value at equilibrium (F_{ST}=0.18) from Rousset (1996, eq. 6). In this situation of high divergence the strength of selection has no apparent influence, but in recombination decreases the variation in F_{ST} for a given set of simulation conditions.

At intermediate migration rates (e.g. $m = 10^{-3}$), genetic divergence generally decreases and other parameters become influential. (Note that for a population size of N=1000 a value of m=10^{-3} corresponds to Nm=1). For instance, populations in different environments (Fig. 3b,d,f,h) here generally show higher neutral genetic divergence than do populations in similar environments (Fig. 3a,c,e,g). This is true across all non-null selection levels when recombination with the selected

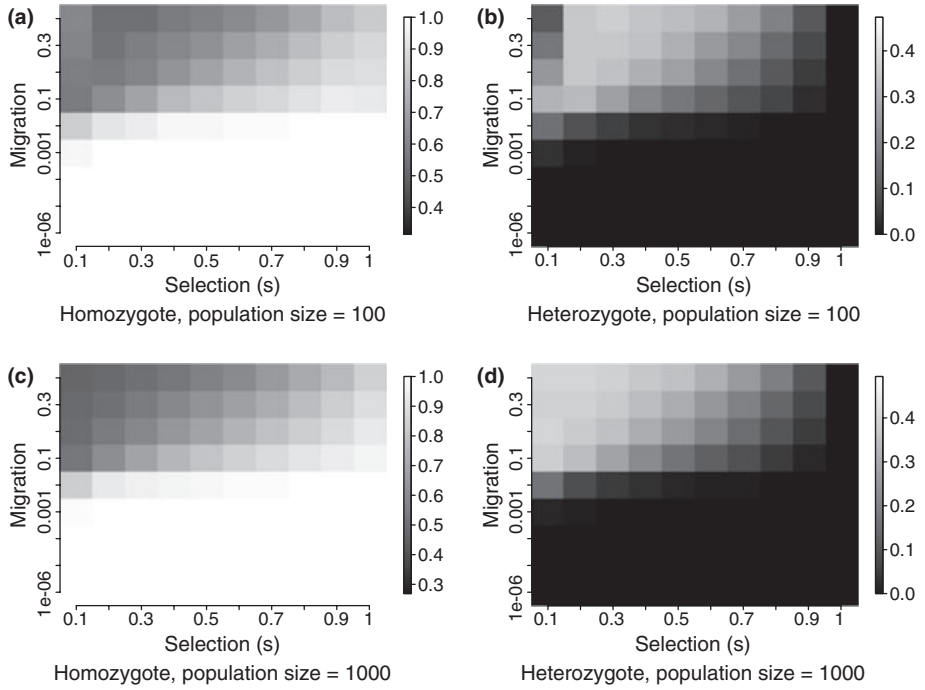

Fig. 2 The average frequency of individuals which are homozygotes for the positively selected locus (a,c) and the average frequency of individuals which are heterozygous at the selected locus (b,d). Results are for the last iteration averaged across all simulations.

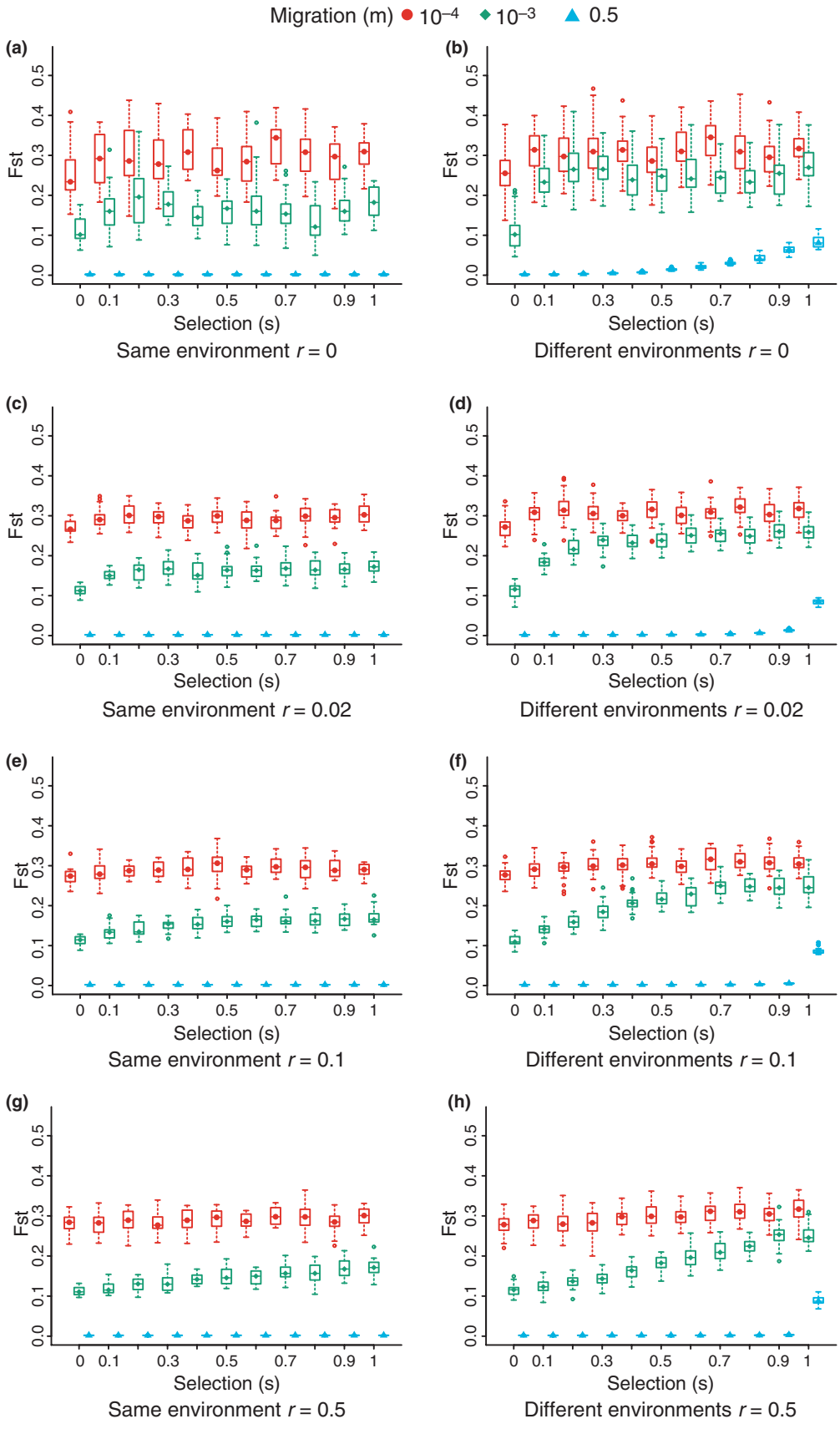

Fig. 3 Qualitative comparisons of F_{ST} between populations in similar environments (left panels) and different environments (right panels). These results are for the last iteration of each simulation with the larger population size (1000). Each box is bounded by the first and third quartiles, the line inside the box is the second quartile (median) and the whiskers extend to 1.5 times the interquartile range (third quartile minus first quartile) or to the maximum or minimum value, as appropriate. All data outside the whisker range are considered outliers and are represented by open circles.

locus is absent (Fig. 3a,b) and for stronger levels of selection when recombination is higher (Fig. 3c–h).

At the highest migration rate (m=0.5), neutral genetic divergence is present in only two instances. The first occurs when recombination with the selected locus is absent and selection is at least moderately strong (Fig. 3a,b). The second occurs when recombination is present and selection is exceptionally strong (right-hand side of Fig. 3d,f and h).

Statistical tests for signatures of ecological speciation

We now consider situations in which the above patterns might be statistically detectable in a typical empirical study. We will here use the phrase 'positive results' when the expected signature of ecological speciation is detected: greater genetic differences between populations in different environments than between populations in similar environments. In all cases, the proportion of comparisons for a given parameter set that yield a positive result is roughly equivalent to the expected statistical power of the test.

At low migration rates ($m = 10^{-4}$), positive results are obtained approximately half of the time (Fig. 4). Positive results become more common when recombination rates are non-null or when selection is reasonably strong. This result does not, however, represent a reliable cue for ecological speciation. It instead simply

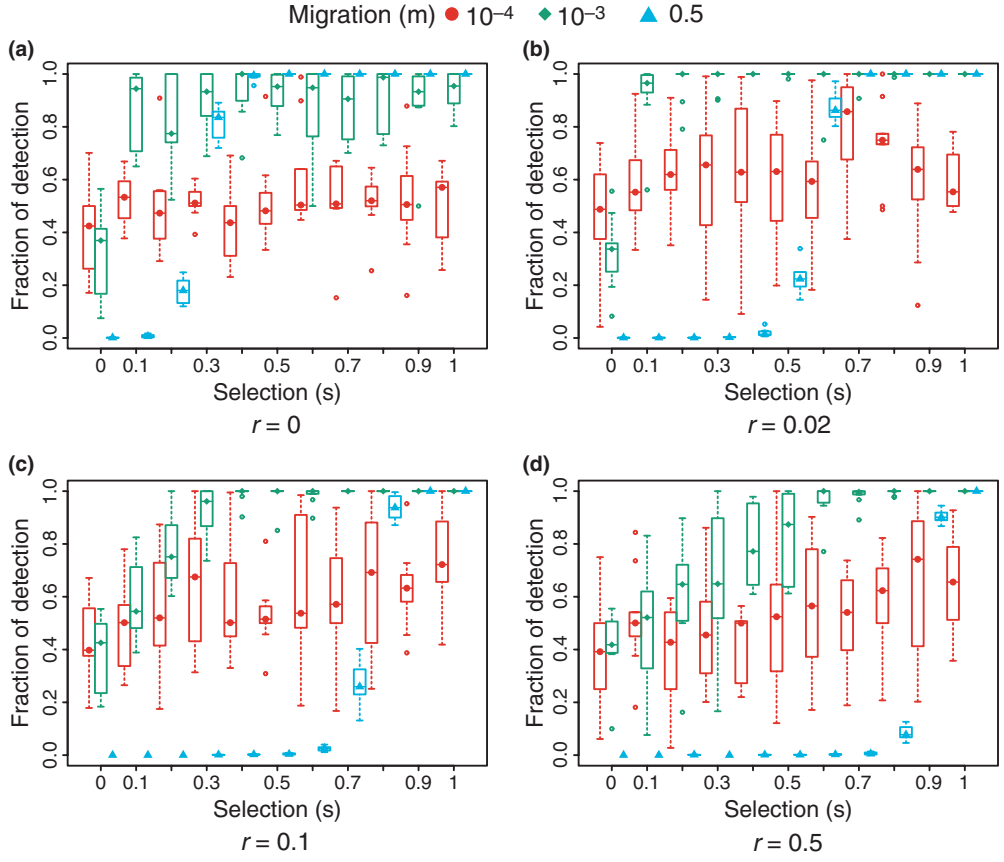

Fig. 4 Results of statistical analyses (100 bootstrap comparisons of 95% confidence intervals) for when F_{ST} comparisons reveal the expected signature of ecological speciation. The y-axis shows the fraction of simulations in which F_{ST} was significantly greater between populations in different environments than between population in similar environments. These results are cumulative of the last 100 iterations of each simulation with the larger population size (1000). For an explanation of the boxplot conventions, see caption to Fig. 3.

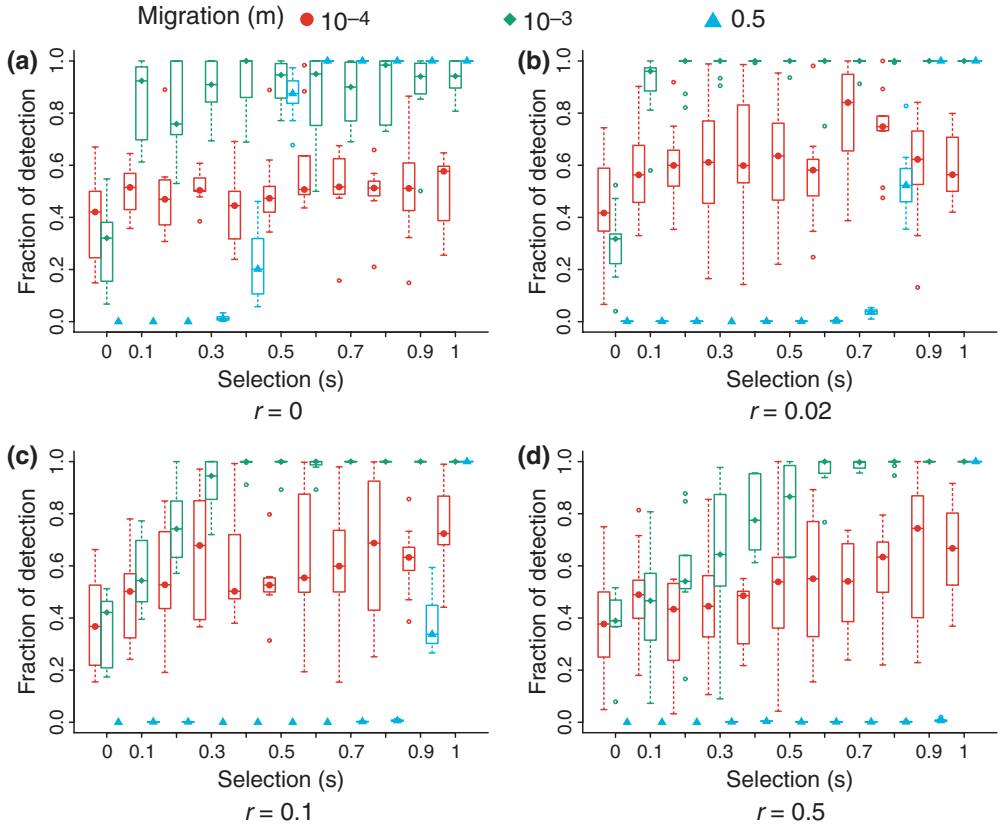

Fig. 5 Results of subsampling statistical analyses (100 bootstrap comparisons of 95% confidence intervals) for when F_{ST} comparisons reveal the expected signature of ecological speciation. The *y*-axis shows the fraction of simulations in which F_{ST} was significantly greater between populations in different environments than between population in similar environments. These results are cumulative of the last 100 iterations of each simulation with 20 individuals sampled in each population of size (1000). For an explanation of the boxplot conventions, see caption to Fig. 3.

reflects random divergence among populations that sometimes by chance leads to the expected signature.

At intermediate migration rates (e.g. $m=10^{-3}$), and in the presence of natural selection ($s>0$), positive results are nearly always obtained when recombination is absent or low (Fig. 4a). When recombination rates are high, positive results are nearly always obtained when selection is also high (Fig. 4d). When recombination rates are high and selection is weak to moderate, positive results are highly variable within a given parameter set.

At the highest migration rate ($m=0.5$), a sharp transition in the frequency of positive results is seen: from no positive results at low selection to all positive results at high selection. The location of this transition is influenced by the recombination rate, with a lower level of selection required to obtain positive results when recombination rates are lower (Fig. 4).

All of the above results were based on the entire population. Generally, similar qualitative results were obtained when sampling only 20 individuals from each population (Fig. 5). Quantitative results differ in the following ways. First, for a given set of simulation conditions, more variable results were sometimes obtained with subsampling. Second, for a given level of migration and recombination, stronger selection is sometimes necessary to detect the expected signature of ecological speciation. This is most obvious when migration rates are very high (Fig. 5).

Earliest divergence

All of the above results are those recorded after 4900 generations, at which time an equilibrium is generally evident in our simulations. Many empiricists, however, may be dealing with non-equilibrium conditions. We therefore considered the time frame after which the expected signature of ecological speciation can be consistently detected following colonization of new environments. We specifically recorded, for a given simulation, the first time (number of generations) after which positive results are obtained for 10 successive

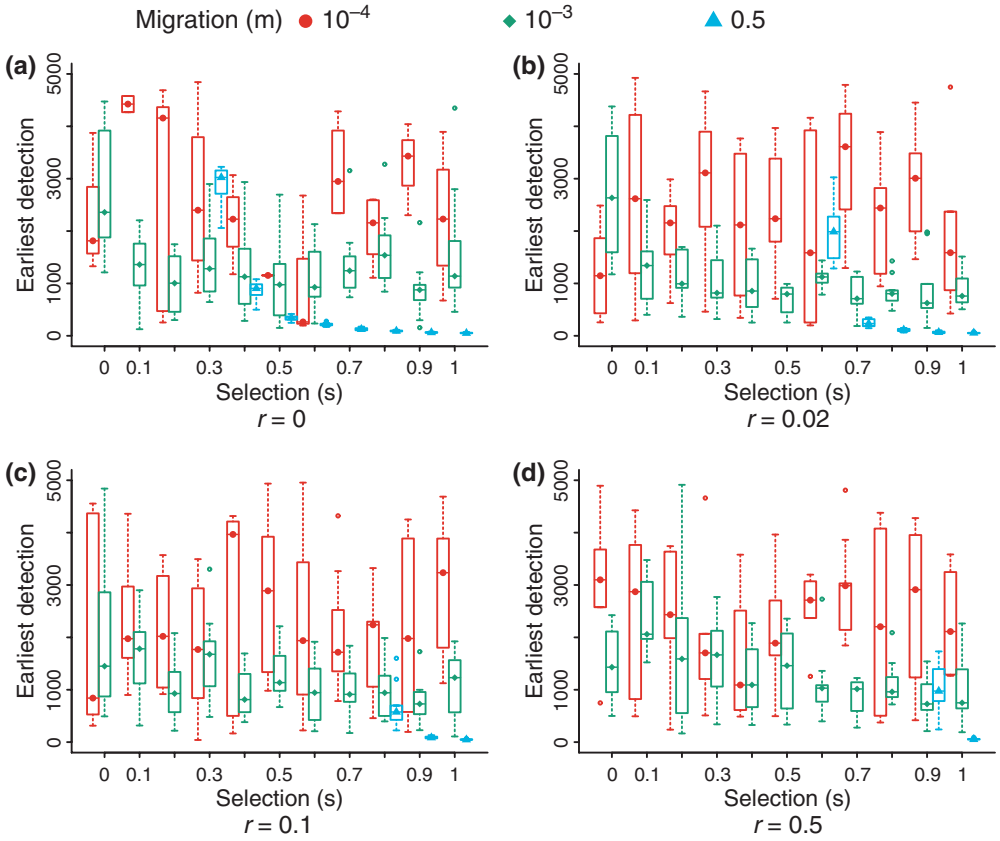

Fig. 6 Time in generations of the first consistent statistical detection (earliest detection) of the expected signature of ecological speciation based on bootstrap F_{ST} comparison. These results show the first generation after colonization at which consistent detection (i.e. for at least 10 consecutive generations) was achieved in simulations with the larger population size (1000). For an explanation of the boxplot conventions, see caption to Fig. 3.

generations. For intermediate migration rates (the situation where detection was most reliable at equilibrium; see above), consistent positive results are usually obtained after a few hundreds to a thousand generations (Fig. 6). Consistent detection occurs the earliest when selection is the strongest (Fig. 6). Detection is sometimes possible after only a few generations – but more generations are required for this result to be consistent across 10 successive generations. For example, when gene flow is high (m=0.1), selection is high (s=0.8) and recombination is free (r=0.5); 30% of the tests yield a positive result only after 15 generations, a time frame similar to the empirical study of Hendry *et al.* (2000).

Population size

We now consider how the above results hold, or are modified, when population sizes are small (100) instead of large (1000). The main change here is a general increase in F_{ST} (Fig. S1, Supporting Information) over the simulations described above. Apart from this numerical difference, trends with respect to the other parameter values are generally not altered, either qualitatively (Fig. S1) or statistically (Fig. S2). However, when positive results are rare for large populations, they are even rarer for smaller populations. In short, it can sometimes be more difficult, and never easier, to detect the expected signature of ecological speciation when populations are small than when they are large.

Evaluating our statistical short cuts

In one-tailed comparisons of the lower F_{ST} confidence limit for populations in different environments to the upper F_{ST} confidence limit for populations in similar environments, it makes little difference whether bootstrapping is based on either 100 or 10 000 replicates (Fig. S3). Specifically, conclusions regarding whether populations in different environments show greater genetic divergence than populations in similar environments are equivalent in 98.84% of the comparisons tested. Of the 1.16% differences, 0.89% repre-

Table 2 Comparison between the results of STRUCTURE and the results of the 100 bootstrap F_{ST} comparison method

		s=0.1; m=10^{-3}		s=0.5; m=10^{-6}		s=0.8; m=10^{-3}	
Iteration	Recombination	Structure	Bootstrap	Structure	Bootstrap	Structure	Bootstrap
100	0	2*	25	6	0	2*	50
	0.08	4	25	4	62.5	4	0.5
	0.5	4	12.5	4	12.5	4	62.5
3000	0	2	50	4	37.5	2	100
	0.08	4	37.5	4	37.5	2	100
	0.5	4	87.5	3	25	2	100
5000	0	2	100	4	50	2	100
	0.08	2*	50	4	50	2	100
	0.5	4	37.5	4	0	2	100

The strength of selection is s and the migration rate is m. STRUCTURE gives the number of inferred populations and the bootstrap method gives the per cent of comparisons in which F_{ST} is greater for populations in different environments than for populations in similar environments. When STRUCTURE finds only two populations, and these are in different environments, then it has detected the expected signature of ecological speciation.

sented false positives for the expected signature when using 100 replicates and 0.27% represented false negatives under those conditions. Moreover, no trend of those false positives or false negatives was evident with migration rate, strength of selection, number of generations or recombination rate. In short, our bootstrap short cut was of no consequence to our conclusions.

For the analyses with STRUCTURE, we find results very similar to those obtained using the above bootstrap F_{ST} comparisons (Table 2). That is, when the F_{ST} bootstrap method consistently found positive results, STRUCTURE always found two populations corresponding to the two environments. When the F_{ST} bootstrap method did not find consistent results, neither did STRUCTURE. In these latter cases, STRUCTURE sometimes grouped populations from different environments and sometimes split populations from similar environments. Only one of the 27 tests in STRUCTURE identified the signature of ecological speciation when F_{ST} comparisons did not.

Discussion

Does divergent selection generate a detectable generalized barrier to gene flow? That is, can we use neutral genetic markers to reliably detect progress toward ecological speciation? The short answer is that it depends on a variety of parameters, which is the same general conclusion obtained by Thibert-Plante & Hendry (2009). The quantitative results of these two studies should not, however, be directly compared because the two models are very different. Results of the present model are much more comprehensive and appropriate when considering implications for empirical studies. In the following discussion, remember that the 'expected signature' of ecological speciation is a reduction in neutral gene flow between populations in different environments relative to that between populations in similar environments.

One important parameter determining detection success (i.e. statistical confirmation of the expected signature) is migration rate. Under our specific simulation conditions, reliable detection is really possible only at intermediate levels of migration. If migration is instead high, detection is difficult simply because divergent selection is often not powerful enough to reduce gene flow to the point where neutral genetic divergence can proceed. If migration is instead low, detection is inconsistent. That is, genetic divergence at neutral markers is detectable about half of the time, but this occurs simply by chance; i.e. gene flow is so low that all populations drift apart to a similar degree. In the absence of divergent selection (s=0), detecting greater differences between populations in different environments is a false positive clearly due to drift. In the presence of divergent selection (s>0), drift may also be important, but its effects cannot be conclusively separated from those of selection. Finding the expected signature of ecological speciation in cases of very low gene flow is therefore not a reliable indication that ecological speciation is actually occurring.

A second important parameter is the magnitude of the environmental difference between populations, i.e. the strength of divergent selection. Genetic divergence at (even unlinked) neutral markers is generally higher when environmental differences are greater, as long as migration rates are intermediate (as above). This effect of selection occurs because greater environmental differ-

ences lead to greater natural selection against maladapted migrants and first-generation hybrids. In this case, linkage disequilibrium before recombination of genes between individuals from different environments causes reduced gene flow at neutral loci even if they are not physically linked to loci under divergent selection. Several other parameters influence the level of divergent selection at which the expected signature is apparent. As migration increases, for instance, a larger environmental difference (stronger divergent selection) is required for a generalized barrier to gene flow. One might therefore wonder how these critical levels of divergent selection correspond to those observed in nature. The advantage of local individuals over foreign individuals in reciprocal transplant experiments has an average of 45% (Hereford 2009). The advantage of local individuals as defined by Hereford (2009) (s_{Hereford}) can be compared with our selection strength by using equation 2.

$$s_{\text{Hereford}} = \frac{2s}{2 - s} \qquad \text{(eqn 2)}$$

Most of the levels of divergent selection (s) in our model that first led to consistent detection of the expected signature were well within this range. For instance, the critical selection level above which detection occurs 80% of the time is s=0.4 (s_{Hereford}=0.5) for the larger population size at an intermediate migration rate (m=10^{-3}) with free recombination (r=0.5). Natural populations therefore might often be in the selection range where variation in other parameters really does influence the inferential potential of neutral genetic markers.

A third important parameter is recombination rate. As expected, the closer a neutral locus is to a selected locus, the greater the effect of divergent selection on reducing gene flow at the neutral locus (Gavrilets 2004; Grahame *et al.* 2006; Nosil *et al.* 2008, 2009a). That is, genetic hitchhiking can increase the chance of detecting the expected signature of ecological speciation. This effect can be most clearly seen in the case of high gene flow (m=0.5): as recombination increases, greater environmental differences are required to detect the expected signature (Fig. 4). The same pattern is found for lower migration rates, but the transitional strength of selection is less abrupt. But how does one interpret divergence at neutral loci linked to selected loci? Certainly, this is no longer necessarily indicative of a 'generalized barrier' to gene flow. We will return to this question below but our main concern here is when ecological speciation can be detected with unlinked neutral genetic markers.

A fourth important parameter is population size. When the expected signature of ecological speciation is either always or never detected for large populations under a given parameter set, the same is generally true for small populations. However, for parameter combinations where variable outcomes are obtained for large populations, smaller populations generally made it even more difficult to detect the expected signature of ecological speciation. The reason appears to be that smaller populations are more sensitive to drift and less responsive to selection (Falconer & Mackay 1996). In other words, when populations are smaller, the signal from selection is more difficult to separate from the noise due to drift.

It is important to note that we did not, for computational reasons, simulate the effects of other parameters of potential interest: numbers of neutral loci, mutation rates and numbers of selected loci. First, we suspect that variation in numbers of loci will not be a major concern. The reason is that an increase in the number of neutral loci beyond 12 does not change the variance in F_{ST}, and, below that, results only in a slight increase in variance (Balloux & Goudet 2002). Moreover, increasing efficiency and declining cost of genetic analyses dictate that large numbers of loci will attend most future studies. Second, increasing the mutation rate will reduce the likelihood that two alleles of the same size are identical by descent, thus potentially reducing our statistical power. On the other hand, decreasing mutation rates may decrease rates of genetic divergence given that all populations started from a monomorphic situation. Third, increasing the number of selected loci will increase the chances that a given neutral locus is linked to a locus under selection. In this case, neutral loci might be more likely to hitchhike to higher frequency. However, the result is not certain because the effect of a given selected locus might well decrease as more loci come under selection. Moreover, results might not change for those neutral loci that remain unlinked. This would be profitable to simulate in future work.

It is also important to remember that our particular model focused on a specific type of population structure (two populations of two types all exchanging migrants at similar rates) and a specific type of comparison among those populations (divergence between vs. within environment types). Many other possibilities could have been simulated. Examples include clines across habitat transitions (e.g. Ogden & Thorpe 2002; Berner *et al.* 2009), comparisons of the rate of dispersal with the rate of gene flow (e.g. Hendry *et al.* 2000), comparisons of isolation by distance for populations that are or are not in similar environments (e.g. Smith *et al.* 1997; Crispo *et al.* 2006), and correlations between adaptive divergence and neutral genetic divergence (e.g. Gíslason *et al.* 1999; Lu & Bernatchez 1999). In addition, our model focused on soft selection, whereas a model of hard selection might have meant that populations facing particularly strong selection

would go extinct (Holt & Gomulkiewicz 1997; Thibert-Plante & Hendry 2009). We are confident that our *general* conclusion applies to these situations as well: unlinked neutral markers are only sometimes reliable in detecting the expected signature of ecological speciation. It is quite likely, however, that the specific results, such as the conditions under which the method is most useful, will vary among different types of population structure and different types of statistical comparison.

General implications

We have confirmed, at least in our simulations, that ecological differences can sometimes cause reduced gene flow at unlinked neutral markers (see also Gavrilets & Vose 2007; Gavrilets *et al.* 2007; Nosil *et al.* 2008), and that this effect can be statistically detectable within a certain range of parameter space. Consistent and appropriate statistical detection in our model was most likely when divergent selection was strong, migration rates were intermediate and population sizes were large. Detection was also easier when recombination rates between neutral and selected loci were lower, but, in this case, we are no longer confirming a generalized barrier to gene flow. In this case, a possible generalized barrier becomes confounded with what amounts to selection on the neutral locus acting through genetic hitchhiking. This issue poses major practical and conceptual issues for empiricists because the use of truly unlinked neutral markers greatly reduces the range of parameter space under which the expected signature of ecological speciation can be reliably detected. On the other hand, the use of neutral markers that are linked to selected loci means that analyses are sensitive to the effects of selection on regions of the genome, rather than just the indirect effects of ecologically driven reproductive barriers. This is obviously an area where genome scans (Emelianov *et al.* 2004; Grahame *et al.* 2006; Via & West 2008; Nosil *et al.* 2009a; Via 2009; Feder & Nosil 2010) become particularly informative.

When unlinked neutral loci are used, several additional points of caution are warranted. First, we uncovered a wide range of parameter space where divergence at selected loci is present but not detectable at neutral genetic markers. This means that a failure to detect the expected signature of ecological speciation does not necessarily mean that divergent selection is absent and ecological speciation is not proceeding (i.e. a false negative). Second, we found a range of parameter space where the expected signature is found but is driven by chance, not divergent selection. Most notably, these false positives are common when gene flow is low because divergence among all populations is high and largely independent of divergent selection. Third, we found large ranges of parameter space where results are highly variable among replicate simulations. That is, for a given parameter set, the expected signature is sometimes found and sometimes not, presumably owing to the stochastic nature of divergence at neutral markers. As expected, more reliable results are then obtained when populations are larger.

How then can our results be of assistance to empiricists seeking to detect progress towards ecological speciation: i.e. the ecologically driven evolution of reproductive barriers? One encouraging short answer is that, except when migration is very low, finding the expected signature of ecological speciation might often indicate that ecological speciation really is proceeding. In these cases, unlinked neutral markers can be quite useful. This is somewhat of a relief (at least to us) because we have previously used positive results in related assays to infer the presence of ecological speciation (Berner *et al.* 2009). One discouraging short answer is that when the expected signature is not detected, this does not necessarily mean that ecological speciation is not proceeding. This is not a relief (at least to us) because we have used negative results in related assays to infer minimal progress towards ecological speciation (Crispo *et al.* 2006). Taken together, these two answers add up to the conclusion that neutral genetic markers can be a valuable part of studies of ecological speciation, but that confidence in interpreting a given result often requires additional information. For instance, our analysis confirms the importance of obtaining information about particular neutral loci (possible linkage to selected loci) and particular population parameters (migration, selection and population sizes). These parameters can tell the investigator whether or not they should worry about false positives or false negatives in neutral-marker assays. In addition, the use of neutral genetic markers should be coupled to other methods for inferring ecological speciation, such as the testing of an ecological basis for specific reproductive barriers. Ecological speciation is clearly out there (Rundle & Nosil 2005), but just as clearly not everywhere (Hendry 2009), and so continued improvements to inferential methods are necessary.

Acknowledgements

XTP and APH were sponsored by the Natural Sciences and Engineering Research Council (NSERC) of Canada. Thanks to E. Crispo for help with the statistics. We also thank Erika Crispo, Ben Haller and three anonymous referees for their comments and valuable suggestions on the manuscript. XTP is grateful to the Biology Graduate Student Association (BGSA) of McGill for a travel grant to present this work at the Canadian Society for Ecology and Evolution in Halifax and to The Society for the Study of Evolution (SSE) for an international

travel grant to present this work in Moscow, Idaho. Thanks to McGill University (Department of Biology) and S. Bunnell for help using the bioinformatics cluster for the simulations.

References

Balloux F, Goudet J (2002) Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Molecular Ecology*, **11**, 771–783.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridizing populations. *Heredity*, **57**, 357–376.

Bengtsson BO (1985) The flow of genes through a genetic barrier. In: *Evolution: Essays in honour of John Maynard Smith* (eds Greenwood PJ, Harvey PH, Slatkin M), pp. 31–42. Cambridge University Press, Cambridge.

Berner D, Grandchamp A-C, Hendry AP (2009) Variable progress toward ecological speciation in parapatry: stickleback across eight lake–stream transitions. *Evolution*, **63**, 1740–1753.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics*, **62**, 1408–1415.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research*, **70**, 155–174.

Crispo E, Bentzen P, Reznick DN, Kinnison MT, Hendry AP (2006) The relative influence of natural selection and geography on gene flow in guppies. *Molecular Ecology*, **15**, 49–62.

Dallas JF (1992) Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. *Mammalian Genome*, **3**, 452–456.

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 3166–3170.

Emelianov I, Marec F, Mallet J (2004) Genomic evidence for divergence with gene flow in host races of the larch budmoth. *Proceedings of the Royal Society Biological Sciences Series B*, **271**, 97–105.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics*. Longman, Essex, UK.

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, in press, doi: 10.1111/j.1558-5646.2010.00943.x.

Gavrilets S (2004) *Fitness Landscapes and the Origin of Species*. Princeton University Press, Princeton, New Jersey.

Gavrilets S, Vose A (2005) Dynamic patterns of adaptive radiation. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 18040–18045.

Gavrilets S, Vose A (2007) Case studies and mathematical models of ecological speciation. 2. Palms on an oceanic island. *Molecular Ecology*, **16**, 2910–2921.

Gavrilets S, Vose A (2009) Dynamic patterns of adaptive radiation: evolution of mating preferences. In: *Speciation and Patterns of Diversity* (eds Butlin R, Bridle J, Schluter D), pp. 102–126. Cambridge University Press, Cambridge.

Gavrilets S, Vose A, Barluenga M, Salzburger W, Meyer A (2007) Case studies and mathematical models of ecological speciation. 1. Cichlids in a crater lake. *Molecular Ecology*, **16**, 2893–2909.

Gíslason D, Ferguson MM, Skúlason S, Snorrason S (1999) Rapid and coupled phenotypic and genetic divergence in icelandic arctic char (*Salvelinus alpinus*). *Canadian Journal of Fisheries and Aquatic Sciences*, **56**, 2229–2234.

Grahame JW, Wilding CS, Butlin RK (2006) Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. *Evolution*, **60**, 268–278.

Hendry AP (2009) Ecological speciation! or the lack thereof? *Canadian Journal of Fisheries and Aquatic Sciences*, **66**, 1383–1398.

Hendry AP, Wenburg JK, Bentzen P, Volk EC, Quinn TP (2000) Rapid evolution of reproductive isolation in the wild: evidence from introduces salmon. *Science*, **290**, 516–518.

Hereford J (2009) A quantitative survey of local adaptation and fitness trade-offs. *The American Naturalist*, **173**, 579–588.

Holt RD, Gomulkiewicz R (1997) How does immigration influence local adaptation? A reexamination of a familiar paradigm. *The American Naturalist*, **149**, 563–572.

Kachitvichyanukul V, Schmeiser BW (1988) Binomial random variate generation. *Communications of the ACM*, **31**, 216–222.

Kimura M, Ohta T (1975) Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proceedings of the National Academy of Sciences of the United States of America*, **72**, 2761–2764.

Lu G, Bernatchez L (1999) Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): support for the ecological speciation hypothesis. *Evolution*, **53**, 1491–1505.

Maynard Smith J, Haigh J (1974) Hitch-hiking effect of a favorable gene. *Genetical Research*, **23**, 23–35.

Nosil P, Egan SP, Funk DJ (2008) Heterogeneous genomic differentiation between walking-stick ecotypes: 'isolation by adaptation' and multiple roles for divergent selection. *Evolution*, **62**, 316–336.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009a) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Nosil P, Harmon LJ, Seehausen O (2009b) Ecological explanations for (incomplete) speciation. *Trends in Ecology and Evolution*, **24**, 145–156.

Ogden R, Thorpe RS (2002) Molecular evidence for ecological speciation in tropical habitats. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 13612–13615.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Räsänen K, Hendry AP (2008) Disentangling interactions between adaptive divergence and gene flow when ecology drives diversification. *Ecology Letters*, **11**, 624–636.

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

Rundle H, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.

Schluter D (2000) *The Ecology of Adaptive Radiation*. Oxford University Press, New York.

Smith TB, Wayne RK, Girman DJ, Bruford MW (1997) A role for ecotones in generating rainforest biodiversity. *Science*, **276**, 1855–1857.

Spirito F, Rossi C, Rizzoni M (1983) Reduction of gene flow due to the partial sterility of heterozygotes for a chromosome mutation 1. Studies on a neutral gene not linked to the chromosome mutation in a 2 population-model. *Evolution*, **37**, 785–797.

Thibert-Plante X, Hendry AP (2009) Five questions on ecological speciation addressed with individual-based simulations. *Journal of Evolutionary Biology*, **22**, 109–123.

Valdes AM, Slatkin M, Freimer NB (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics*, **133**, 737–749.

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9939–9946.

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**, 4334–4345.

Weber JL, Wong C (1993) Mutation of human short tandem repeats. *Human Molecular Genetics*, **2**, 1123–1128.

Weir BS (1996) Intraspecific differentiation. In: *Molecular Systematics* (eds Hills DM, Moritz C, Mable BK), pp. 385–405. Sinauer Associates, Inc., Sunderland, MA.

Whitlock MC (2002) Selection, load and inbreeding depression in a large metapopulation. *Genetics*, **160**, 1191–1202.

Whitlock MC, Phillips PC (2000) The exquisite corpse: a shifting view of the shifting balance. *Trends in Ecology and Evolution*, **15**, 347–348.

This study forms part of Xavier Thibert-Plante's PhD thesis on ecological speciation. Xavier Thibert-Plante is a computer guru interested in addressing ecological questions on speciation using individual-based modelling. Andrew Hendry investigates factors that influence the evolution of biological diversity, including natural selection, gene flow, adaptation, and reproductive isolation. He conducts research in a number of study systems, including the Galápagos islands (Darwin's Finches), Trinidad and Tobago (Guppies) and British Columbia (Sticklebacks).

Supporting information

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

Fig. S1 Qualitative comparisons of F_{ST} between populations in similar environments (left panels) and different environments (right panels). These results are for the last iteration of each simulation with the smaller population size (100). For an explanation of the boxplot conventions, see caption to Fig. 3.

Fig. S2 Results of statistical analyses (100 bootstrap comparisons of 95% confidence intervals) for when F_{ST} comparisons reveal the expected signature of ecological speciation. The *y*-axis shows the fraction of simulations in which F_{ST} was significantly greater between populations in different environments than between populations in similar environments. These results are cumulative of the last 100 iterations of each simulation with the smaller population size (100). For an explanation of the boxplot conventions, see caption to Fig. 3.

Fig. S3 Comparisons of confidence limits obtained with 100 vs. 10 000 bootstraps (larger minus smaller bootstraps, *y*-axis) relative to the estimated F_{ST} with 10 000 bootstraps (*x*-axis) for populations in different (top panel) or similar (bottom panel) environments. The key observation is that the two bootstrap levels show consistent upper confidence limits for similar environment and consistent lower confidence limits for different environments. The result is no bias when using 100 bootstrap confidence limits to infer the expected signature of ecological speciation.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.