Genetics: Published Articles Ahead of Print, published on August 17, 2009 as 10.1534/genetics.108.099812

TESTING FOR SPATIALLY-DIVERGENT SELECTION: COMPARING Q_{ST} TO F_{ST}

MICHAEL C. WHITLOCK and FREDERIC GUILLAUME *

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

Corresponding author: Michael Whitlock, whitlock@zoology.ubc.ca, *Department of Zoology, University of British Columbia, 6270 University Blvd., Vancouver, BC V6T 1Z4 Canada*

ABSTRACT

Q_{ST} is a standardized measure of the genetic differentiation of a quantitative trait among populations. The distribution of Q_{ST}s for neutral traits can be predicted from the F_{ST} for neutral marker loci. To test for the neutral differentiation of a quantitative trait among populations, it is necessary to ask whether the Q_{ST} of that trait is in the tail of the probability distribution of neutral traits. This neutral distribution can be estimated using the LEWONTIN-KRAKAUER distribution and the F_{ST} from a relatively small number of marker loci. We develop a simulation method to test whether the Q_{ST} of a given trait is consistent with the null hypothesis of selective neutrality over space. The method is most powerful with small mean F_{ST}, strong selection, and a large number (>10) of measured populations. The power and Type I error rate of the new method are far superior to the traditional method of comparing Q_{ST} and F_{ST}.

---

key words: Q_{ST}, F_{ST}, divergent selection, local adaptation, spatial heterogeneity

Introduction

In 1993, SPITZE (1993) and PROUT and BARKER (1993) introduced Q_{ST}, a quantitative genetic analog of WRIGHT's F_{ST}. Just as F_{ST} gives a standardized measure of the genetic differentiation among populations for a genetic locus, Q_{ST} measures the amount of genetic variance among populations relative to the total genetic variance. In the years since, Q_{ST} has been frequently used to test for the effects of spatially divergent (or less commonly, spatially uniform) selection (see reviews in LYNCH *et al.* 1999, MERILÄ and CRNOKRAK 2001, MCKAY and LATTA 2002, HOWE *et al.* 2003, LEINONEN *et al*. 2008, WHITLOCK 2008). In principle, the average Q_{ST} of a neutral additive quantitative trait is expected to be equal to the mean value of F_{ST} for neutral genetic loci. F_{ST} can be readily measured on commonly available genetic markers, and Q_{ST} can be measured as well with an appropriate breeding design in a common-garden setting. As a result, Q_{ST} promises to be an index of the effect of selection on the quantitative trait. If Q_{ST} is higher than F_{ST}, then this is taken as evidence of spatially divergent selection on the trait. If Q_{ST} is much smaller than F_{ST}, then this has been taken as evidence of spatially uniform stabilizing selection, which makes the trait diverge less than expected by chance.

The comparison with F_{ST} is essential to rule out genetic drift as an alternative mechanism for phenotypic divergence among populations. Because finite populations may diverge genetically in the absence of selection, divergence must be greater than expected by drift alone if we are to conclusively demonstrate that divergent selection has played a role in genetic differentiation among populations. Therefore it has become common practice to use F_{ST} of putatively neutral markers as a control for the effects of genetic drift and to compare observed Q_{ST} values for traits to these neutral F_{ST} values.

These comparisons follow two separate methods, to address related but distinct questions. First, many studies of quantitative genetic differentiation measure the Q_{ST} of many traits

and the F_{ST} of many loci, followed by a comparison of the mean Q_{ST} to the mean F_{ST}. Such a comparison may judge whether the conditions are suitable in that species for local adaptation, that is, whether selective differences between populations are large enough relative to gene flow to allow adaptive differentiation (WHITLOCK 2008). We do not consider this sort of comparison in this paper.

The other type of comparison asks whether the Q_{ST} of a single trait is greater than expected by drift, as measured by F_{ST}. This type of comparison is most common, but it is statistically difficult. Unfortunately, as emphasized in a recent review by WHITLOCK (2008), there is great variation in the expected F_{ST} among neutral loci and among the Q_{ST} of different neutral traits. (See figure 1.) The majority of this variation results from evolutionary differences between loci and not sampling error in the observations. ROGERS and HARPENDING (1983) imply that the distribution of Q_{ST} of a single neutral trait should be approximately equivalent to that for F_{ST} of a single neutral locus, and this has been confirmed by simulation for traits determined by additive loci compared to biallelic marker loci (WHITLOCK 2008). The two distributions are similar, but there is great heterogeneity among traits or loci. As a result, in order to show that selection is acting on a trait, it is necessary to show that the value of Q_{ST} has a low probability of being observed given the distribution of neutral Q_{ST}.

Comparing Q_{ST} to the distribution inferred from F_{ST} is difficult for two reasons. First, typical data sets rarely include enough loci to directly infer the distribution of F_{ST} without extra inferential steps. In our approach, we use the distribution of Q_{ST} predicted from the mean F_{ST} and the χ^2 distribution by LEWONTIN and KRAKAUER (1973) to bridge this gap. WHITLOCK (2008) has shown that this distribution is appropriate for nearly all realistic situations for traits determined by additive genetic effects. Second, Q_{ST} for a trait is rarely measured with high precision, so the position of a given estimated Q_{ST} value in the distribution cannot be known without error.

To test the null hypothesis that the spatial distribution of a particular trait is not affected by selection, we wish to compare the observed $\hat{Q}_{ST}$ of that trait (marked with a hat to indicate it is an estimate) to the distribution of Q_{ST} expected for neutral traits.

Unfortunately, calculating the distribution of Q_{ST} for neutral traits is not straightforward, because the estimate of Q_{ST} for a particular trait is variable for several reasons. The estimate of Q_{ST} is subject to measurement error, caused by the finite samples of families and individuals in the quantitative genetic experiment. These cause error in the estimate of the additive genetic variance within populations ($V_{A,within}$) and the genetic variance among populations ($V_{G,among}$), which translate into error of the estimate of Q_{ST}. In addition, there is another source of variation in Q_{ST} among neutral traits, caused by the idiosyncrasies of the evolutionary process in each local population in the study. The true value of Q_{ST} for the set of populations being studied can vary tremendously around its expectation, even for neutral traits, because by chance a finite set of populations may drift in a similar direction (WHITLOCK 2008). As a result, measurements of Q_{ST} can vary both because of statistical and evolutionary variation.

Fortunately, these two sources of variation are fairly well understood individually. The sampling error for the estimates of the variance components can be estimated from standard approaches, and this variation can be well approximated using information from the mean squares of the analysis of the breeding experiment (O'HARA and MERILÄ 2005). The variation in neutral Q_{ST} that results from heterogeneity of evolutionary history can be approximated by the LEWONTIN-KRAKAUER distribution (LEWONTIN and KRAKAUER 1973), if information is available on the mean Q_{ST} of neutral traits (WHITLOCK 2008). This approximation does not depend on the demographic details of the populations in question (WHITLOCK 2008), but the effects of deviations from assumptions of additive gene effect have not yet been tested. The mean of the distribution of values of Q_{ST} for neutral traits is usually not known, but fortunately the mean of the distribution of F_{ST} of neutral loci is expected to be approximately equal to the mean Q_{ST} of neutral traits (SPITZE 1993), and this does not depend on demographic details (WHITLOCK 1999). Therefore the mean F_{ST} measured from a series of genetic markers thought to be selectively neutral can be combined with the LEWONTIN-KRAKAUER distribution to predict the distribution of true neutral Q_{ST} across the range of possible evolutionary trajectories.

Given that the mean value of $\hat{Q}_{ST}$ of neutral traits is expected to equal the mean F_{ST} of neutral markers under certain assumptions (discussed later), we will use $\hat{Q}_{ST} - \overline{F}_{ST}$ as a test statistic and compare the observed quantity to the zero value proposed by the null hypothesis. We will use a traditional hypothesis testing approach, which means that we need to specify the sampling distribution of $\hat{Q}_{ST} - \overline{F}_{ST}$ under the assumption of neutrality. Traditionally, the sampling distribution of $\hat{Q}_{ST}$ is inferred from the data on the trait itself, for example, using bootstrapping to infer the sampling distribution. This is appropriate when calculating a confidence interval for Q_{ST} but is a biased measure of the sampling variance of neutral Q_{ST}. The variance of the sampling distribution of $\hat{Q}_{ST}$ varies with its expected value; larger values of true Q_{ST} have more variable sampling distributions than traits with smaller true Q_{ST}. This association between Q_{ST} and its sampling error is quite strong, as shown in Figure 2. As a result, if the sampling properties of neutral $\hat{Q}_{ST}$ are inferred from a trait with high Q_{ST}, the estimate of the variance of the null distribution will be too high, and the hypothesis test comparing $\hat{Q}_{ST}$ to F_{ST} will be conservative. On the other hand, if a low Q_{ST} is used to estimate the variance of the null distribution, the estimated error will be too small, and the test will reject true null hypotheses too often.

We address this problem by using F_{ST} from putatively neutral maker loci in combination with estimates of the additive genetic variance within populations to predict the sampling variance that would be expected for the Q_{ST} of a neutral trait. We show that the power and Type I error rate of this test are greatly superior to traditional methods.

Method

Testing neutrality

To generate the null distribution of $\hat{Q}_{ST} - \overline{F}_{ST}$, we use a parametric simulation approach. To calculate a $\hat{Q}_{ST} - \overline{F}_{ST}$ value from data, we need estimates of three quantities: $\overline{F}_{ST}$, $V_{A,within}$, and $V_{\mathbf{G},among}$. To calculate the null distribution, we simulate random sampling for each of these quantities under the assumption that the null hypothesis that Q_{ST} equals $\overline{F}_{ST}$

is true. We calculate $\hat{Q}_{ST} - \overline{F}_{ST}$ from the simulated values, and after repeating this 1000 times, we generate the sampling distribution of $\hat{Q}_{ST} - \overline{F}_{ST}$ assuming the null hypothesis.

$\overline{F}_{ST}$ is calculated from marker loci; we use the WEIR and COCKERHAM (1984) method in our test calculations. To simulate the sampling error in estimates of $\overline{F}_{ST}$, for each replicate simulation we randomly sample with replacement from the marker loci until the number of loci in the simulated data set equals the number of loci in the real data set. Mean F_{ST} is calculated from these sampled loci using the method of WEIR and COCKERHAM (1984), and the observed value of their θ is used as the simulated $\overline{F}_{ST}$ value.

$V_{A,within}$ is calculated from a quantitative genetic breeding design. There are several suitable experimental designs for such estimates. In this paper we assume that the additive genetic variance is estimated by a half-sib design, but the approach could easily be modified for other designs. $V_{A,within}$ can be estimated from four times the variance among sires; and to estimate the variance among sires we need the mean squares of sires (MS_{sires}) and the mean squares of dams (MS_{dams}). To simulate estimates of $V_{A,within}$, we use an approach analogous to a parametric bootstrap (O'HARA and MERILÄ 2005). As tested by O'HARA and MERILÄ (2005), $df_{sires} MS_{sires} / \overline{MS}_{sires}$ and $df_{dams} MS_{dams} / \overline{MS}_{dams}$ should be χ^2 distributed, where *df* represents the degrees of freedom associated with a particular level and the overbar indicates the true value of the mean square. Therefore by multiplying the estimated MS/df times a random number from a χ^2 distribution for each of sires and dams we can simulate the sampling distribution of these quantities and therefore of $V_{A,within}$. This procedure is implemented exactly as the parametric bootstrap in O'HARA and MERILÄ (2005), except to avoid a strong source of bias we do not constrain variance component estimates to be positive.

$V_{\mathbf{G},among}$ is calculated from the variance among populations in the mean value of the trait when the organisms are grown in a common environment. The novel aspect of our design comes from how the sampling of $V_{G.among}$ is simulated. As mentioned in the introduction, the sampling variance for $V_{\mathbf{G},among}$ is correlated with the true value of $V_{\mathbf{G},among}$, and therefore if the null hypothesis is true but $V_{\mathbf{G},among}$ incorrectly appears high by sampling error, the estimate of its sampling distribution will also be estimated poorly. If we were

only estimating the value of Q_{ST} itself, this would pose no real problems, but because we are trying to compare Q_{ST} to the neutral expectation, it can be a real source of bias in the calculations. Our solution is to simulate the sampling distribution of $V_{\mathbf{G},among}$ assuming that the null hypothesis is true. We therefore calculate the value of $V_{\mathbf{G},among}$ that would be expected given the observed $\overline{F}_{ST}$ and $V_{A,within}$. Given that Q_{ST} is defined as $Q_{ST} = V_{G,among}/(V_{G,among} + 2V_{A,within})$ and that for neutral traits and neutral loci the average values of Q_{ST} and F_{ST} are approximately equal, we can find the expected value of $V_{\mathbf{G},among}$ under neutrality to be:

$$\hat{V}_{G,among} \cong \frac{2\overline{F}_{St}V_{A,within}}{1-\overline{F}_{ST}}.$$

To simulate the sampling distribution around this expectation, we again assumed that the distribution of trait means among populations follows a normal distribution and multiply $\hat{V}_{G,among}/df_{populations}$ times a random number drawn from a χ^2 distribution with degrees of freedom equal to the number of populations (num_{pops}) minus one. This sampling procedure is the same as assumed by the LEWONTIN-KRAKAUER distribution shown to work well to approximate the distribution of Q_{ST} under a variety of demographic circumstances (WHITLOCK 2008). Simulating the sampling error in this way is identical to the approach taken by O'HARA and MERILÄ (2005) in their parametric bootstrapping, except for using the expected value of $V_{\mathbf{G},among}$ calculated from F_{ST} instead of the observed $V_{G,\,among}$.

For a given hypothesis test using a specific data set, we generate 1000 simulated estimates of $\hat{Q}_{ST} - \overline{F}_{ST}$. For each simulation, $\overline{F}_{ST}$, $V_{A,within}$, and $V_{\mathbf{G},among}$ are randomly drawn as specified above, and $\hat{Q}_{ST} - \overline{F}_{ST}$ is calculated from these simulated values. The distribution of these 1000 simulated values is the null distribution of the hypothesis test. Therefore by comparing the quantile of the observed value of $\hat{Q}_{ST} - \overline{F}_{ST}$ to the simulated distribution, we may determine the *P*-value of the hypothesis test of neutrality.

The supplemental material includes an R program to implement this procedure.

Simulations

We tested the method using simulations conducted with the population genetics simulation software *Nemo* (GUILLAUME & ROUGEMONT, 2006) updated to include quantitative traits. Neutral marker loci were simulated with 100 biallelic loci, with mutation rates of 10^{-5} in either direction. One hundred loci potentially affected the quantitative traits. Mutation was based on an infinite allele model, where the allelic effect of an allele was, if mutated, changed by a factor randomly selected from a Gaussian distribution with genomic mutational variance equal to 0.001. Mutation rates for the quantitative trait loci were set at 10^{-5}. Each of twenty local populations had an effective population size of 500 diploid individuals, and the migration rate among populations varied from $m = 0.05$ to $m = 0.001$ to produce different F_{ST} values, ranging from approximately $F_{ST} = 0.01$ to $F_{ST} = 0.3$. Measurements were taken on the populations after 50,000 generations (or 25,000 generations for the neutral cases), allowing the populations to reach an approximate equilibrium before sampling. The $\hat{Q}_{ST}$ of 10,000 traits were simulated for the neutral traits and 100 for each set of parameters with selection.

In addition to the island model calculations that make the bulk of the simulation tests, we also simulated a one-dimensional, circular stepping-stone model with 60 local populations. Simulations with $F_{ST} = 0.04$ were performed, corresponding to a migration rate of 0.12. Migration occurred only between adjacent (left and right) populations in the stepping stone model, and at most every third population was sampled for F_{ST} and the Q_{ST} calculations, as suggested by BEAUMONT and NICHOLS (1996) and WHITLOCK (2008). For the heterogeneous selection cases, the populations were alternatively assigned to habitats in groups of five.

In some simulations, the quantitative trait was selectively neutral, to allow tests of the Type I error rates of the method. In other simulations, the quantitative trait was subjected to either uniform stabilizing selection (for which all local populations had the same optimum with Gaussian selection with $V_S = 5$) or heterogeneous selection (for which the

selective optimum for half of the local populations was different from the optimum in the other half of the populations.) The strength of selection for the heterogeneous environment case was calculated such that a perfectly adapted individual on one environment would have a 5% or 50% reduction in fitness in the other selective environment in the island or stepping stone model, respectively. The parameters of the selection functions were $V_S = 5$, and the difference between the habitat optimum phenotypes was 0.716 in the island model, and 2.63 in the stepping stone model. There was no environmental effect added to the genotypic values of the quantitative trait loci ($V_E = 0$).

For each simulation, $\hat{Q}_{ST}$ was calculated from a simulated half-sib breeding design. In the default configuration, samples were taken from 20 populations, and for each population five sires were mated to five dams each. These numbers were varied to better understand the power of the approach. Five offspring from each dam were measured, and from the results $\hat{Q}_{ST}$ was calculated from the population and sire effects using an analysis of variance.

For all parameter combinations, we tested the null hypothesis of neutrality using the new method and with the best method previously available, the parametric bootstrap approach from O'HARA and MERILÄ (2005). We refer to this latter approach as the "traditional approach" throughout.

Simulation results

The simulations show that the new method has a more accurate Type I error rate and more power than the traditional method. There is sufficient power to detect high Q_{ST} when the $\hat{Q}_{ST}$ of a trait is several-fold greater than the mean F_{ST} and when large numbers of populations (10 or more) are included in the analysis. However, large numbers of marker loci are not necessary. On the other hand, it is difficult to reliably detect the signal of homogeneous selection; the power to discriminate significantly small Q_{ST} values is low, even when the mean F_{ST} value is much higher than expected for most intraspecific comparisons.

First, examine the cases where the null hypothesis is true; that is, when the trait is evolving without the influence of selection. The traditional method has an overall Type I error rate that is a bit high overall (Table 1), but it is seen to be particularly poor when the Type I errors are divided into the two tails. The Type I error rate for the traditional method with low Q_{ST} values is 7.0-7.8% (in contrast to the expected 2.5%), whereas the Type I error rate is far too low for high values of Q_{ST} compared to mean F_{ST} (0.41-0.44%). In all cases, the one-tailed error rates are different from the stated 2.5% with extremely small *P*-values (the largest being $P = 4 \times 10^{-59}$). In contrast, the new method has a much better Type I error rate. The total error rate for the new method is always within the 95% confidence interval of the expected value of 5%, and the errors are more evenly divided into the two tails.

With heterogeneous selection in the island model, the mean Q_{ST} ranged from 0.026 to 0.564, depending on the amount of migration among populations (see table 2). The power of the method depends in part on the relative value of the typical Q_{ST} value in comparison to the mean F_{ST}. When Q_{ST} is expected to be much greater than the mean F_{ST}, the method has substantial power (Figure 3). Importantly, the new method has much higher power to detect heterogeneous selection than the traditional method (Figure 3). With small sample sizes and low true differences between Q_{ST} and F_{ST}, neither method is able to detect the effects of selection, and with extremely large samples both methods have high power. But for intermediate (and realistic) sample sizes with moderate Q_{ST} values, the new method has substantially more power to detect heterogeneous selection than the traditional method. We also ran simulations of stronger selection (where an individual perfectly adapted to the other environment would have a 10% fitness reduction), where Q_{ST} is higher. In these cases the power was very high for both methods, except for the cases when there were only two populations in the study. There again, the new method greatly outperformed the traditional method (results not shown).

In contrast, under only rare circumstances was there much power to detect that the Q_{ST} value of a trait was significantly smaller than expected under neutral differentiation (figure 4). Even when the mean neutral F_{ST} is relatively high, the left tail of the

distribution of neutral Q_{ST} is still relatively dense for small values, making it difficult to separate a low Q_{ST} from neutral expectations.

These preceding calculations are based on moderately large sample sizes for the quantitative genetic measurements but not very many (10) marker loci for the calculation of F_{ST} . Increasing the number of marker loci increases power, but not dramatically (Figure 5a). On the other hand, using more families per population to estimate $\hat{Q}_{ST}$ better has a beneficial effect (figure 5b). However, the power of the analysis is critically dependent on the number of populations surveyed (figure 3). The variance of the expected $\hat{Q}_{ST}$ distribution reduces in proportion to the number of demes measured (WHITLOCK 2008), and the reliability of $\hat{Q}_{ST}$ estimates increases strongly with number of demes (GOUDET and BÜCHI 2007). Reliable inference about the neutrality of quantitative traits requires sampling of large numbers of populations. The estimation of both Q_{ST} and F_{ST} depends critically on the estimate of the variance among populations, and the power of the estimate of this variance depends on the number of populations sampled. In studies with small numbers of populations, the $\hat{Q}_{ST}$ estimates were also quite biased for both methods (results not shown), explaining the apparently higher power for the smallest sample sizes.

Results under the stepping stone model are quite similar. The mean Q_{ST} for the stepping stone simulations was 0.638 with selection and 0.0488 for the neutral case. The power of the analysis is largely dependent on the number of populations sampled (Figure 6) and varies in an equivalent way with the number of families and neutral loci sampled (results not shown).

Discussion

The Q_{ST} of neutral traits is potentially extremely variable from trait to trait, especially when the number of populations in the system (or in the study) is small. This distribution is approximately predictable with knowledge of the mean F_{ST} of neutral marker loci for the same populations (WHITLOCK 2008). A simple function of Q_{ST} (equal to

$\frac{\left(num_{pops}-1\right)Q_{ST}}{\overline{Q}_{ST}}$) is approximately distributed by a χ^2 distribution with num_{pops} -1 degrees of freedom; this derives from the LEWONTIN-KRAKAUER distribution (LEWONTIN and KRAKAUER 1973). Given that for traits determined by additively acting alleles the mean Q_{ST} is approximately equal to the mean F_{ST}, the sampling distribution of neutral Q_{ST} can be predicted.

Most studies of Q_{ST} explicitly compare $\hat{Q}_{ST}$ of a trait to F_{ST}, as a test of whether spatially heterogeneous or homogeneous selection affects the distribution of the trait. These studies use the observed properties of $\hat{Q}_{ST}$ to predict its sampling distribution. However, when testing the null hypothesis of neutrality, we need to infer the sampling properties of $\hat{Q}_{ST}$ for neutral traits, not of traits with high or low expected Q_{ST}s. The difference matters because the width of the sampling distribution of $\hat{Q}_{ST}$ depends on its mean value (Figure 2).

We have developed a new method to test for selective neutrality using the difference between $\hat{Q}_{ST}$ and mean F_{ST}. We account for the expected distribution of Q_{ST} under neutrality using a distribution inferred from the mean F_{ST}. Compared to the traditional method, the new approach works extremely well. The traditional method, which infers the distribution of $\hat{Q}_{ST}$ from the observed $\hat{Q}_{ST}$, has very poor false positive rates (Type I error). High Q_{ST} reject the null hypothesis far too rarely, and low Q_{ST} rejects the null hypothesis too often (Table 1). This is because the error variance is overestimated for high Q_{ST} and underestimated for low Q_{ST} (Figure 2). The Type I error rate for our new method are close to the stated values, and they are symmetric in the upper and lower tails as is desirable.

The new method is also more powerful than the traditional method for detecting spatially heterogeneous selection. Both the new and traditional methods work well when Q_{ST} is much greater than F_{ST} and with data from many populations, and both fail with too few data (e.g. when the number of populations is two). However, in intermediate cases with moderate Q_{ST} and moderately large sample sizes, the new method has much more power

than the traditional approach. With homogeneous selection, the traditional method appears to have more power, but this is largely due to its inflated Type I error rate. Positive results are not reliable for homogeneous selection and small numbers of populations.

Unfortunately, in some biologically interesting circumstances, there are a limited number of populations that exist in nature, and in these circumstances it is simply not possible to reliably show that even a large $\hat{Q}_{ST}$ is different from the neutral expectation. This is especially true when the mean F_{ST} of neutral markers is also high. For example, some applications of the Q_{ST} approach have been made comparing a pair of sub-species. In these cases, the mean F_{ST} is typically high (or the two populations would not have been given sub-specific status) and the total number of such populations in nature is just two. In this case, there is little hope of finding significant evidence of selective differentiation via the Q_{ST} approach. For example, when there are only two populations, the 97.5 percentile of the distribution of F_{ST} or Q_{ST} is approximately five times the mean of the distribution, according to the LEWONTIN-KRAKAUER distribution. Even with no error in estimating Q_{ST}, a trait would have to have a Q_{ST} value five times as large as the mean F_{ST} to be significantly in the tail of the distribution, for the two population case. Q_{ST} is never estimated with such small error, so in practice the $\hat{Q}_{ST}$ of the trait would have to be much larger than five times the mean F_{ST} to find statistical evidence of selection.

There is little power in typical data sets to test for spatially-uniform stabilizing selection using $\hat{Q}_{ST}$ - F_{ST} comparisons. It has been suggested that small values of Q_{ST} relative to F_{ST} may indicate strong stabilizing selection with the same optimum in all populations, because such selection would oppose genetic drift and maintain approximately the same mean in each local population. However, the distribution of neutral $\hat{Q}_{ST}$ includes a dense left-hand tail in most intraspecific comparisons, because, with a small mean F_{ST} and a few populations sampled, a large number of loci with small F_{ST} (or neutral traits with small Q_{ST}) are expected just by chance. Only with very strong selection and levels of F_{ST} that verge on interspecific values (F_{ST} = 0.2) have we found even moderate power to detect spatially uniform selection (Fig.4).

There are a few other caveats that need to be kept in mind when applying this method, in common with all interpretations of Q_{ST}. It is crucial that F_{ST} and Q_{ST} are both estimated without bias, and there are many sources of bias that affect most $\hat{Q}_{ST}$ measures (WHITLOCK 2008). In particular, it is important that $\hat{Q}_{ST}$ is estimated from a breeding design and not just from phenotypic data. Furthermore, it is essential that the study organisms are grown in a common garden to avoid conflating phenotypic plasticity with local adaptation.

Importantly, the simulations conducted here all assumed that traits are determined by alleles that interact additively, both between and within loci. Dominance variance can under some circumstances cause mean Q_{ST} to be greater than mean F_{ST}, even for neutral traits. There is controversy over whether the effects of dominance will typically lead to increased values of Q_{ST} (GOUDET and BÜCHI 2006, GOUDET and MARTIN 2007, LOPEZ-FANJUL *et al*. 2003, 2007), but importantly the distribution of Q_{ST} among neutral traits has not been investigated for traits affected by dominance or epistasis. Our ability to use the distribution predicted from the F_{ST} of marker loci depends on the distribution being similar for Q_{ST}, and this has not been investigated for traits with dominance. This method, and indeed any comparison of Q_{ST} and F_{ST}, requires stringent assumptions about the additive basis of the quantitative trait.

The method also relies on the assumption that we are able to identify neutral markers to use for F_{ST} to generate the null distribution. With a large number of marker loci, the chances may be high that at least some of the loci are affected by spatially heterogeneous selection. If such loci can be identified by a procedure such as *fdist2* (BEAUMONT and NICHOLS 1996), then removing them from the analysis is probably best, although this may make the test less conservative. Alternatively, all marker loci could be left in the analysis, on the assumption that the loci affecting quantitative traits may sometimes differentiate by pleiotropic effects or by linkage to other selected loci. Keeping the full spectrum of marker loci potentially would control for these extraneous effects.

Finally, there are some specific issues with the new simulation method that limit its breadth of application. The method given here uses the LEWONTIN-KRAKAUER

distribution to infer the distribution of neutral Q_{ST} from mean F_{ST}. According to simulation results this should work fine for typical values of mean F_{ST} (less than about 0.2). However, the LEWONTIN-KRAKAUER distribution is based on a χ^2 distribution, and its right tail extends to positive infinity and is not constrained to be less than one. As a result, for large values of mean F_{ST} the probability of the right tail of this LEWONTIN-KRAKAUER distribution becomes an inaccurate representation of the true tail probability.

In order to use Q_{ST} to test for selection, we have to compare an individual trait's $\hat{Q}_{ST}$ to the distribution of possible values of Q_{ST} under neutrality. By doing so, we have developed a method that has much better Type I error rates and higher power for detecting spatially heterogeneous selection than traditional approaches.

Acknowledgments:

We thank BOB O'HARA for providing the R code for the parametric bootstrap, and SALLY OTTO, JÉRÔME GOUDET, and an anonymous reviewer for extremely helpful comments on a previous version of this paper. JÉRÔME GOUDET pointed out that F_{ST} estimated from multiallelic loci have a different distribution, which helped us to clarify the use of the Lewontin-Krakauer distribution for Q_{ST}. This research was supported by a Discovery Grant from the Natural Science and Engineering Research Council (Canada) to M.C.W. and a Swiss National Science Foundation grant PA00A3-115383 to F.G.

LITERATURE CITED

BEAUMONT, M. A. 2005. Adaptation and speciation: What can F_{ST} tell us? TREE 20:435-440.

BEAUMONT, M. A., and R. NICHOLS. 1996. Evaluating loci for use in the genetic analysis of populations structure. Proc. Roy. Soc. Lond. B 263:1619-1626.

CHARLESWORTH, B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. Molecular Biology and Evolution 15:538-543.

GOUDET, J. and G. MARTIN. 2007. Under neutrality, $Q_{ST} \leq F_{ST}$ when there is dominance in an island model. Genetics, in press.

GOUDET, J., and L. BÜCHI. 2006.The effects of dominance, regular inbreeding and sampling design on Q_{ST}, an estimator of population differentiation for quantitative traits. Genetics 172:1337-1347.

GUILLAUME, F., and J. ROUGEMONT. 2006. Nemo: an evolutionary and population genetics programming framework. Bioinformatics 22:2556–2557.

HOWE, G. T., S. N. AITKEN, D. B. NEALE, K. D. JERMSTAD, N. C. WHEELER, and T. H. H. CHEN. 2003. From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. Can. J. Bot. 81:1247–1266.

JOHANSSON, M., C. R. PRIMMER, and J. MERILÄ. 2007. Does habitat fragmentation reduce fitness and adaptability? A case study of the common frog(*Rana temporaria*). Mol. Ecol. 16:2693–2700.

LEINONEN, T., R. O'HARA, J. M. CANO, and J. MERILÄ. 2008. Comparative studies of quantitative trait and neutral marker divergence: A meta-analysis. J. Evol. Biol. 21:1–17.

LEWONTIN, R. C., and J. KRAKAUER. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175-195.

LÓPEZ-FANJUL, C., A. FERNÁNDEZ, and M. A. TORO. 2003. The effects of neutral nonadditive gene action on the quantitative index of population divergence. Genetics 164:1627-1633.

LÓPEZ-FANJUL, C., A. FERNÁNDEZ, and M. A. TORO. 2007. The effect of dominance on the use of the $Q_{ST} - F_{ST}$ contrast to detect natural selection on quantitative traits. Genetics, in press.

LYNCH, M. et al. 1999. The quantitative and molecular genetic architecture of a subdivided species. Evolution 53:100–110.

MCKAY, J. K. and R. G. LATTA. 2002. Adaptive population divergence: markers, QTL and traits. Trends Ecol. Evol. 17:285–291.

MERILÄ. J., and P. CRNOKRAK. 2001. Comparison of genetic differentiation at marker loci and quantitative traits. J. Evol. Biol. 14: 892-903.

O'HARA, R. B., and J. MERILÄ. 2005. Bias and precision in Q_{ST} estimates: Problems and some solutions. Genetics 171:1331-1339.

PROUT, T., and J. S. F. BARKER. 1993. F statistics in *Drosophila buzzatii*: selection, population size and inbreeding. Genetics 134:369 – 375.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria.

ROGERS, A. R., and H. C. HARPENDING. 1983. Population structure and quantitative characters. Genetics 105:985-1002.

SPITZE, K., 1993 Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation. Genetics 135:367–374.

WEIR, B. S., and C. C. COCKERHAM. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

WHITLOCK, M. C. 2008. Evolutionary inference from Q_{ST}. *Molecular Ecology* 17:1885–1896.

Tables

Table 1. Type I error rates for the island model simulations based on the island model with twenty populations and twenty sires in the sample, for a two-sided test with $\alpha = 0.05$.

Migration rate	Traditional method		New method	
	Left Tail (low Q_{ST})	Right tail (high Q_{ST})	Left tail (low Q_{ST})	Right tail (high Q_{ST})
0.001	0.0706	0.0042	0.0244	0.024
0.01	0.0700	0.0044	0.0257	0.026
0.05	0.0784	0.0041	0.0245	0.0293

Table 2. Mean Q_{ST} and F_{ST} values for different island model parameters.

Migration rate	Mean F_{ST} (neutral)	Mean Q_{ST} (heterogeneous selection)	Mean Q_{ST} (homogeneous selection)
0.001	0.318	0.564	0.044
0.01	0.045	0.232	0.015
0.05	0.009	0.026	0.005

Figures

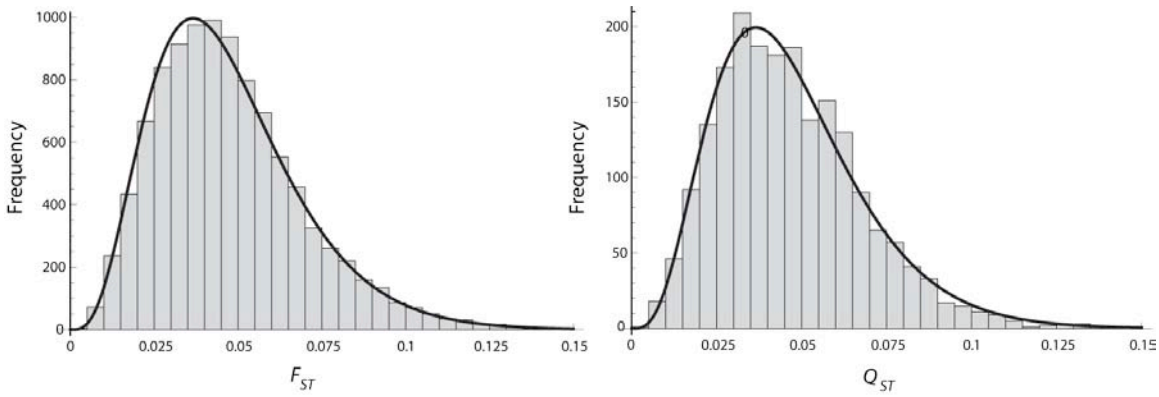

Figure 1. The distribution of F_{ST} for neutral loci and Q_{ST} for neutral quantitative traits. The histograms show the results of simulations of a set of ten local populations each of 100 individuals, connected by 5% migration following island model assumptions. The solid line shows the distribution predicted by the LEWONTIN-KRAKAUER (1973) distribution. The distribution of Q_{ST} for neutral traits is very similar to the distribution of F_{ST} for single neutral loci, as can be seen by their mutual good fit to the LEWONTIN-KRAKAUER distribution. (Figure modified from WHITLOCK (2008).)

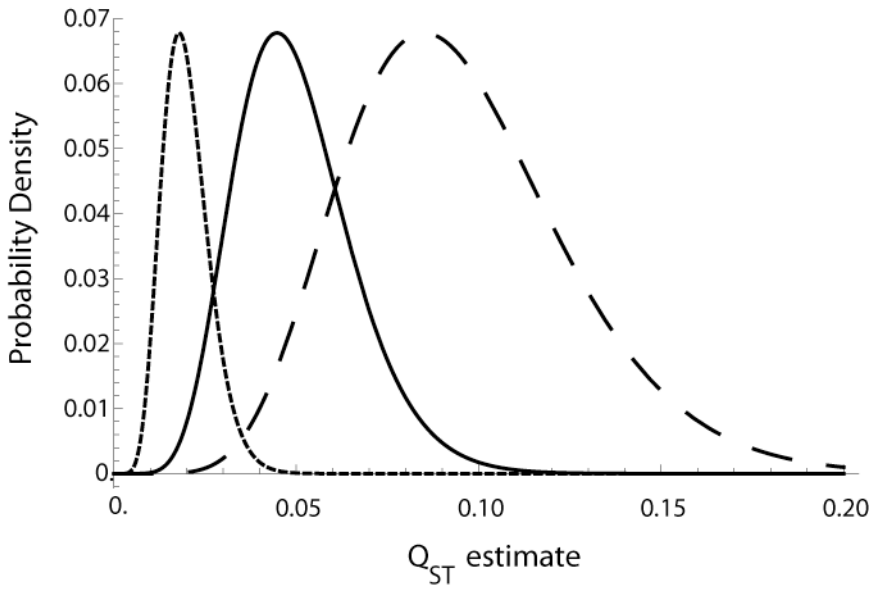

Figure 2. The width of the estimated sampling distribution of $\hat{Q}_{ST}$ varies with mean Q_{ST}. The solid line shows the sampling distribution of Q_{ST} when the true mean Q_{ST} value is 0.05. The dotted line shows the sampling distribution that would be estimated for Q_{ST} from a trait that by chance was at the first percentile of this distribution, and the dashed line shows the sampling distribution that would be inferred from a value taken at the 99th percentile. If the Q_{ST} of a trait differs from the expectation by chance, then the width of the sampling distribution will also be estimated with substantial error. In particular, the error variance of $\hat{Q}_{ST}$ is overestimated with Q_{ST} estimates that are too high and underestimated for small Q_{ST} values.

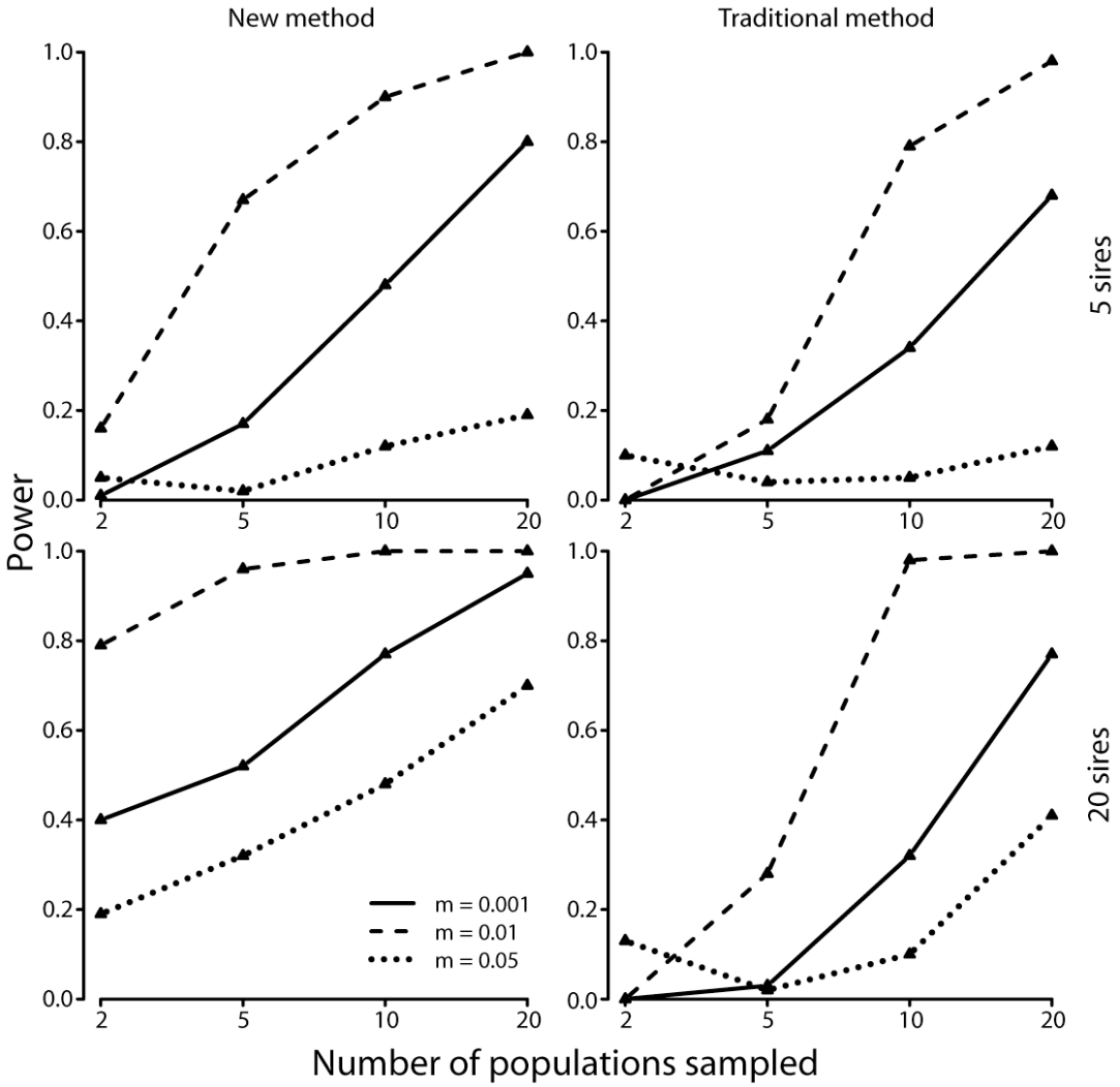

Figure 3. The power of the new approach (left graphs) compared to the traditional approach (right graphs), as a function of the number of populations included in the sample. Results are shown for the island model for three different migration rates. The populations experienced spatially heterogeneous selection; an individual that is perfectly adapted to one habitat will have a 5% reduction in fitness in the other habitat. Each habitat contains half of the populations. Each population was measured for five (top graphs) or twenty (bottom graphs) sires, each mated to five dams, with five offspring per dam for the $\hat{Q}_{ST}$ estimates, and F_{ST} was calculated from ten loci. When F_{ST} is high (with low migration rates), it is more difficult to distinguish a high Q_{ST} value caused by heterogeneous selection, and the power of the test is very weak if a small number of

populations are measured in the study. The new simulation method has much better power than the traditional comparison of Q_{ST} and F_{ST}.

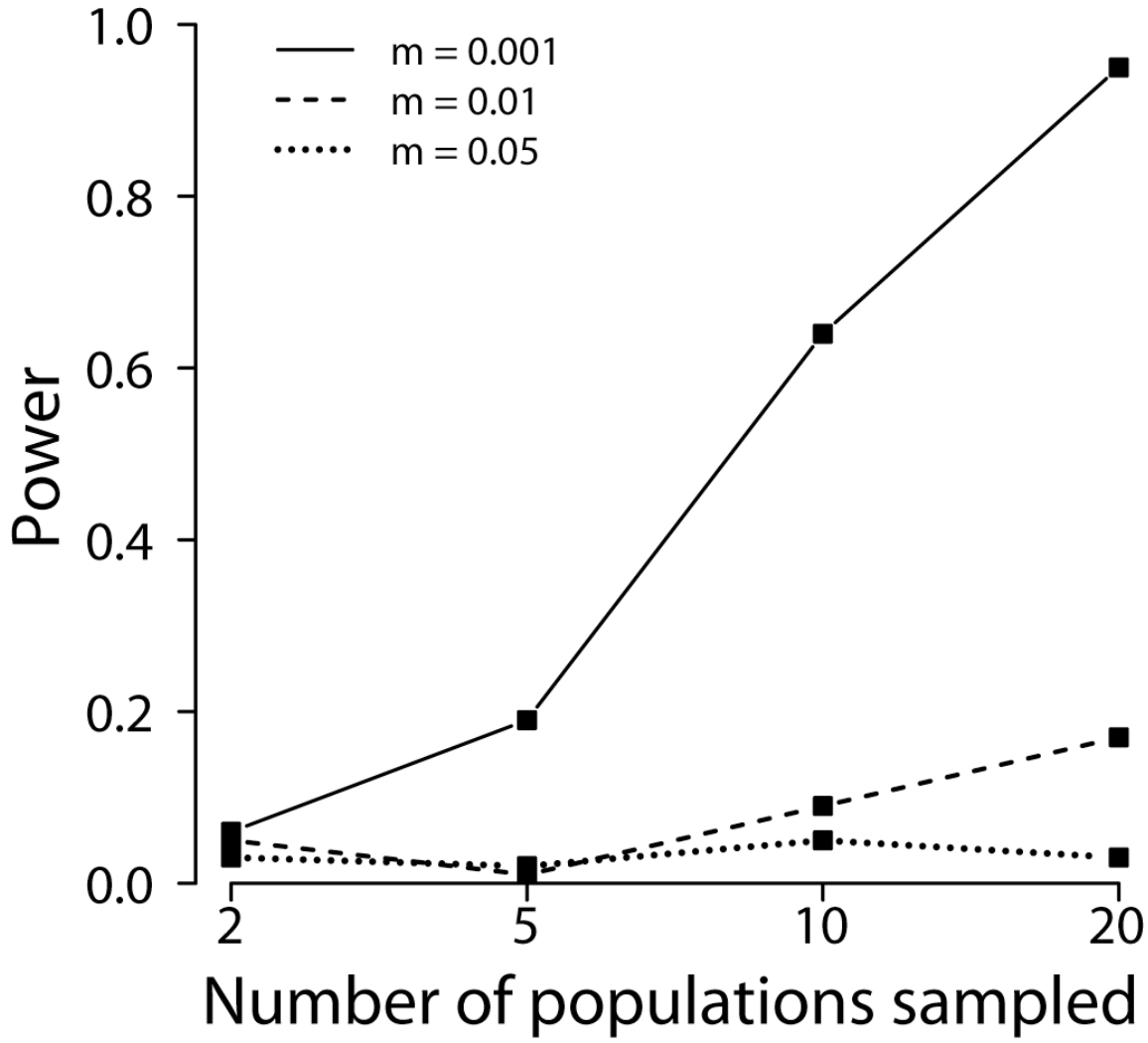

Figure 4. Power of Q_{ST} to detect homogeneous selection. The trait experienced stabilizing selection in each population with a uniform optimum. Stabilizing selection was strong, with $V_S = 5$. Sample sizes are the same as in Figure 3a.

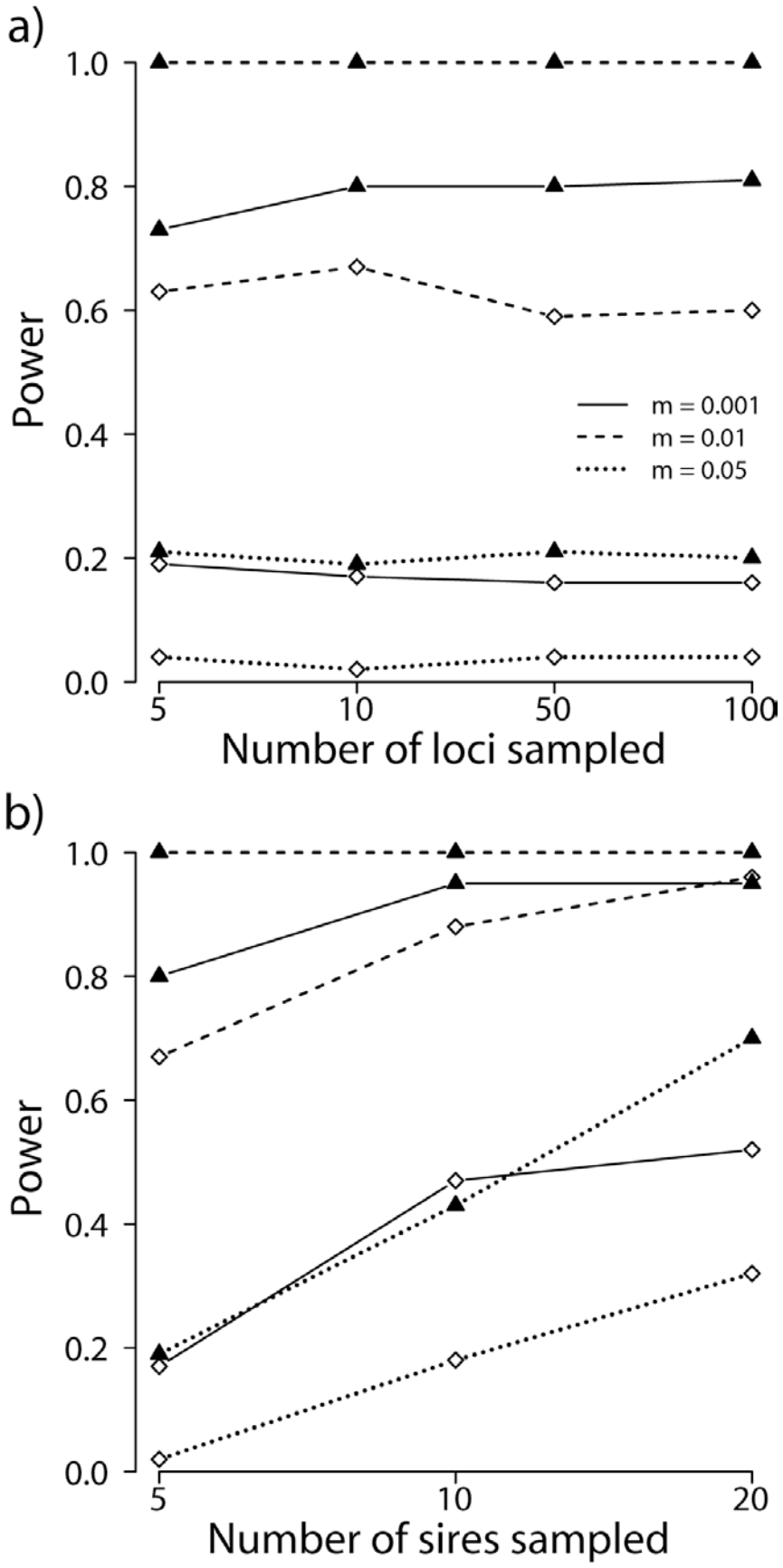
a)
1.0
0.8
0.6
0.4
0.2
0.0
Power
m = 0.001
m = 0.01
m = 0.05
5
10
50
100
Number of loci sampled
b)
1.0
0.8
0.6
0.4
0.2
0.0
Power
5
10
20
Number of sires sampled

Figure 5. Power to detect heterogeneous selection as a function of (a) the number of marker loci examined and (b) the number of sires per population. All other sample sizes and parameters are the same as in Figure 3 and 4, with 20 (filled symbols) or 5 (open symbols) populations sampled. The power of the analysis is not much affected by the number of marker loci examined, but increasing the number of families per population can increase power.

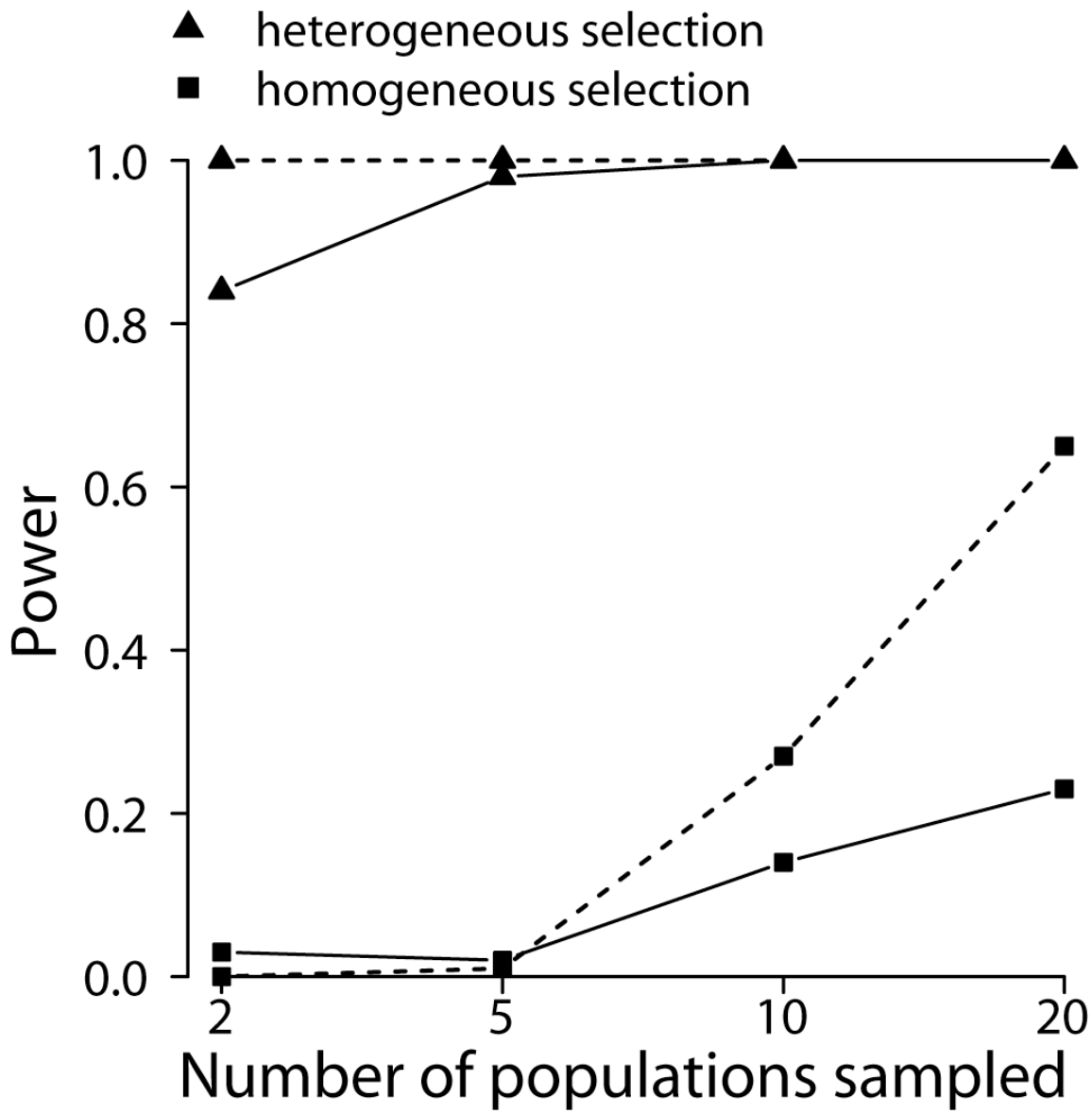

Figure 6. The power of the simulation method applied to simulated data from a stepping-stone model. Sixty populations on a linear stepping-strong were simulated with $N = 500$ and $m = 0.12$. F_{ST} averaged 0.04. In the heterogeneous selection case, each population experienced one of two selective environments, chosen at random for each population with equal probability. The resulting Q_{ST} was approximately 0.6 on average. In the

homogeneous selection case, the Q_{ST} was approximately 0.008. The method was applied using data from populations separated by at least two intervening populations, sampling five (solid lines) or twenty (dashed lines) populations.