HacDivSel: Two new methods (haplotype-based and outlier-based) for the detection of divergent selection in pairs of populations of non-model species.

A. Carvajal-Rodríguez

Departamento de Bioquímica, Genética e Inmunología. Universidad de Vigo, 36310 Vigo, Spain.

Keywords: haplotype allelic class, F_{ST}, G_{ST}, outlier test, divergent selection, genome scan, ecological genetics, non-model species.

**:* A. Carvajal-Rodríguez. Departamento de Bioquímica, Genética e Inmunología. Universidad de Vigo, 36310 Vigo, Spain. Phone: +34 986813828

email: acraaj@uvigo.es

Running title: HacDivSel: detection of divergent selection

Abstract

In this work two new methods for detection of divergent selection in populations connected by migration are introduced. The new statistics are robust to false positives and do not need knowledge on the ancestral or derived allelic state. There is no requirement for performing neutral simulations to obtain critical cut-off values for the identification of candidates. The first method, called $nvdF_{ST}$, combines information from the haplotype patterns with inter-population differences in allelic frequency. Remarkably, this is not a F_{ST} outlier test because it does not look at the highest F_{ST}s to identify loci. On the contrary, candidate loci are chosen based on a haplotypic allelic class metric and then the F_{ST} for these loci are estimated and compared to the overall F_{ST}. Evidence of divergent selection is concluded only when both the haplotype pattern and the F_{ST} value support it. It is shown that power ranging from 79-94% are achieved in many of the scenarios assayed while the false positive rate is controlled below the desired nominal level (γ = 0.05). Additionally, the method is also robust to demographic scenarios including population bottleneck and expansion. The second method, called EOS, is developed for data with independently segregating markers. In this case, the power to detect selection is attained by developing a new G_{ST} extreme-outlier set test (EOS) based on heuristic problem solving via a *k*-means clustering algorithm. The utility of the methods is demonstrated through simulations and the analysis of real data. Both algorithms have been implemented in the program HacDivSel.

Introduction

Current population genetics has a main focus in the detection of the signatures of selection at the molecular level. In previous years, the main effort was focused in human and other model organisms but, now, the increasing amount of information on genomes of non-model species also permits to focus the search for selection in many other situations of interest. One of these includes the sought for local adaptation and selection in structured populations. By non-model species we mean a species for which we lack a priori information on candidate loci with known function that could be potentially adaptive. As well, the allelic state, ancestral or derived, is unknown. There are several methods aiming to detect selection in genomic regions undergoing local adaptation (Whitlock 2015). Some of them are based on finding outlier loci when measuring genetic differentiation between populations. From its original formulation (LK test, Lewontin & Krakauer 1973) this technique has been both questioned and improved in many different ways (Akey 2009; Akey *et al.* 2002; Bonhomme *et al.* 2010; Chen *et al.* 2010; Duforet-Frebourg *et al.* 2014; Excoffier *et al.* 2009; Fariello *et al.* 2013; Foll & Gaggiotti 2008; Whitlock & Lotterhos 2015). For example, under the infinite island model, the effect of gene flow and the corresponding correlations in the gene frequencies among local subpopulations could inflate the neutral variance in F_{ST}, leading to high rate of false positives (Akey 2009; Akey *et al.* 2002; Bonhomme *et al.* 2010; Chen *et al.* 2010; Duforet-Frebourg *et al.* 2014; Excoffier *et al.* 2009; Fariello *et al.* 2013; Foll & Gaggiotti 2008). Moreover, several processes not related with local adaptation, as background selection, species wide selective sweeps or bottleneck and population expansions scenarios, can also produce F_{ST} outliers (Bierne *et al.* 2013; Maruki *et al.* 2012). Finally, F_{ST} methods are not designed for detecting polygenic selection (Bierne *et al.* 2013; Li

et al. 2012). All of the above provokes that, still under optimal conditions, the F_{ST}-outlier methods tend to produce many false positives (Lotterhos & Whitlock 2014; Perez-Figueroa *et al.* 2010).

Recently, some promising alternatives have appeared to deal with these issues (Bonhomme *et al.* 2010; Duforet-Frebourg *et al.* 2014; Fariello *et al.* 2013; Frichot *et al.* 2013; Whitlock & Lotterhos 2015) allowing for more accurate identification of loci under local divergent selection (De Villemereuil *et al.* 2014; Lotterhos & Whitlock 2014). However, besides the fact that these methods come with stronger computational cost, they still present some caveats that difficult their use in exploratory studies with non-model organisms. For example, it is necessary to known the allelic state, ancestral or derived, or having information on the diploid genotype, or the methods are dependent on some a priori assumptions on the outlier model and/or need to account for the population structure.

In the present work we develop two new methods specialized to detect divergent selection in pairs of populations with gene flow. This should be done at genomic or sub-genomic level, working without information on the structure of the population tree and ignoring the state, ancestral or derived, of the alleles. It would be also desirable that selection could be detected without simulating any neutral demographic scenario. The first method combines haplotype information with a diversity-based F_{ST} measure. It is a window-size based method that performs automatic decision-making to fix the adequate window. The second method is more useful when the haplotype information is not at hand and performs a two step G_{ST} outlier test. The first step of the algorithm consists in a heuristic search for different outlier

clusters, the second step is just an LK test that will be performed only if more than one outlier cluster was found and in this case the test is applied through the cluster with the higher G_{ST} values.

Both methods are robust to false positives in the several scenarios assayed, proving to be powerful for detecting single or polygenic selection when two populations connected by migration undergo divergent selection.

The design of the work is as follows:

Firstly, the model for the haplotype-based method is developed, which includes the computation of a normalized variance difference for detecting specific haplotype patterns under selection. A high variance difference value allows the choice of candidate sites which are then used to perform an F_{ST} index measure by comparing the F_{ST} of each candidate with the mean across the genome. The significance of the F_{ST} index is evaluated by resampling the sequence sites using inter-population mean frequency as expectation. Secondly, the algorithm for a conservative, extreme outlier set test (EOS) is developed to deal with the case of fully unlinked SNPs. After that, a brief explanation of the implementation of both methods in the program HacDivSel is given. Then, in the results section we evaluate the methods by simulation of a two population scenario under divergent selection. Finally the EOS test is applied to a recently published real data set and the results are discussed.

The *nvdF$_{ST}$* model

Generalized HAC variance difference

For a given sample, let the major-allele-reference haplotype (MARH) be the one carrying only major frequency alleles of its constituting SNPs (Hussin *et al.* 2010). Define the mutational distance between any haplotype and MARH as the number of sites (SNPs) in the haplotype carrying a non-major allele. Each group of haplotypes having the same mutational distance will constitute a haplotype allelic class. Therefore (with some abuse of notation) we call HAC to the distance corresponding to each haplotype class and the HAC of a given haplotype corresponds to the number of non-major (i.e. minor) alleles it carries, so that every haplotype having the same number of minor alleles belongs to the same HAC class.

Given the definitions above, consider a sample of n haplotypes of length L SNPs. For each evaluated SNP i ($i \in [1,L]$) we can perform a partition of the HAC classes into P_1, the subset of HACs for the haplotypes carrying the most frequent (major) allele at the SNP i under evaluation and P_2 the subset with the remaining haplotypes carrying the minor allele at i. That is, let '0' to be the major allele for the SNP i and accordingly is '1' the minor. Then, P_1 includes every haplotype carrying the allele '0' for the SNP i and P_2 the remaining haplotypes carrying '1' for that SNP. In P_1 we have different HAC values depending on the distance of each haplotype from MARH and similarly in P_2. In each subset we can compute the variance of the HACs. That is, in P_1 we have the variance v_{1i} and correspondingly variance v_{2i} in P_2. The rationale of the HAC-based methods is that if the SNP i is under ongoing selection then the variance in the partition 1 will tend to be zero because the allele at higher frequency (i.e. in the partition 1) should be the favored one and the sweeping effect will make the HAC values

in this partition to be lower (because of sweeping of other major frequency alleles) consequently provoking lower variance values (Hussin *et al.* 2010). The variance in the second partition should not be affected by the sweeping effect because it does not carry the favored allele. So, the difference v_{2i} - v_{1i} would be highly positive in the presence of selection and not so otherwise. For a window size of *L* SNPs, the variance difference between P_2 and P_1 can be computed to obtain a summary statistic called Svd (Hussin *et al.* 2010) that can be generalized to

$$gSvd_i = \frac{V_{2i}-V_{1i}}{L} \times f_i(1-f_i)^a \times b.$$

Where f_i is the frequency of the derived allele of the SNP *i*, and the parameters *a* and *b* permit to give different weights depending on if it is desired to detect higher frequencies (*a* = 0) or more intermediate ones (*a* > 0) of the derived allele. If *a* = 0 and *b* = 1 the statistic corresponds to the original Svd and if *a* =1 and *b* = 4 it corresponds to the variant called SvdM (Rivas *et al.* 2015). Note that when taking *a* = 1 it is not necessary to distinguish between ancestral and derived alleles because f_i and 1- f_i are interchangeable.

A drawback in the gSvd statistic is its dependence on the window size as has already been reported for the original Svd (Hussin *et al.* 2010; Rivas *et al.* 2015). Although gSvd is normalized by *L*, the effect of the window size on the computation of variances is quadratic (see Appendix A-1 for details) which explains why the normalization is not effective in avoiding a systematic increase of the statistic under larger window sizes. This impact of the window size is important because the two different partitions may experience different scaling effects, which would increase the noise in the estimation. Additionally, the change in the scale due to the window size will be dependent on the recombination and selection

rates. Thus, it is desirable to develop a HAC-based statistic not dependent on the window size. In what follows, the between-partition variance difference is reworked in order to develop a new normalized HAC-based statistic, specially focused on detecting divergent selection in local adaptation scenarios with migration.

Note that, for a sample of size n, the corresponding means and variances at each partition are related via the general mean and variance in that sample. Consider m, m_1, m_2 the mean HAC distances in the sample and in the partitions P_1 and P_2 respectively, for any candidate SNP i. We have the following relationships for the mean m and sample variance S^2 values (the subscripts 1 or 2 identify the partition, see Appendix A-2 for details)

$$m = \frac{n_1 m_1 + n_2 m_2}{n}; \qquad S^2 - \bar{S} = \frac{n}{n-1}\Delta \qquad (1)$$

with $\bar{S} = \frac{(n_{1i}-1)S_{1i}^2 + (n_{2i}-1)S_{2i}^2}{n-1}$; n_1 and n_2 are the sample sizes ($n_1 \geq n_2$ by definition) and $\Delta = \frac{n_1 n_2}{n^2}(m_1 - m_2)^2$.

Using the relationships in (1), the between partitions variance difference can be recomputed and some non-informative term discarded (see details in the Appendix) to finally obtain a new statistic for the variance difference of the candidate i

$$vd_i = \frac{(n-1)S^2 - (n-2)S_{1i}^2}{n_{2i}-1} \times 4f_i(1-f_i) \qquad \text{with } n_{2i} > 1 \qquad (2)$$

Note that (2) will augment with decreasing S_1 and increasing S_2 (because the latter increases S). Therefore, if selection is favoring the major allele of the SNP i, then the variance S_1^2 will tend to zero, the sample variance S^2 will be a function of the variance S_2^2 and the value in (2) will be positive. Because we are interested in detecting intermediate allele frequencies (see

below), the parameters *a* and *b* from gSvd have been substituted by *a* = 1 and *b* = 4 as these are the values that permit to ignore the allelic state while maximizing (2) for intermediate frequencies.

Variance upper bound and normalized variance difference

Note that HAC values vary in the range [0, *L*] which provokes that the sample variance S^2 has an upper bound at nL^2 / [4(*n*-1)]. Then the maximum variance difference occurs when f_i = 0.5, S_1^2 = 0, n_2 = *n*/2 and by substituting in (2) we get an upper bound

$$vd_i \leq \frac{nL^2}{2(n-2)} \quad (3)$$

If we divide (2) by the right side in (3) we have a normalized variance difference

$$nvd_i = \frac{2(\mathrm{n}-2)[(n-1)S^2-(n-2)S_{\mathrm{1i}}^2]}{(n_{2i}-1)\mathrm{nL}^2} \times 4f_i(1-f_i) \qquad (4)$$

The quantity from (4) can be computed for each SNP in a sample of sequences of a given length *L* and the SNP giving the maximum *nvd* considered as a candidate for selection. Furthermore, it is possible to compute (4) for each population or to combine the two populations in a unique sample. The latter is better for our purpose of looking for divergent selection in populations undergoing gene flow. When pooling both populations the frequencies tend to be intermediate in the divergent selective sites. Therefore, we compute the normalized variance difference in (4) for the data obtained by merging the shared SNPs from the two population samples. Note however that the reference haplotype (MARH) is computed just from one of the populations (the population 1, by default).

Recall that (4) is already normalized by the square of the window size *L*. However, the problem of choosing an optimal window size remains. A solution to this problem is to automate the choice by selecting the size which gives the maximum value for the statistic (Rivas *et al.* 2015). Therefore, we focus on the candidate having maximum *nvd* from every SNP and window size.

At this point we already have a HAC-based statistic, *nvd*, that is independent of the window size and that should produce higher positive values for pairs of populations undergoing divergent selection. However, if there is no selection, the maximum *nvd* value would be a false positive. Unfortunately, we ignore the distribution of the statistic and cannot decide if a given value is supporting the hypothesis of selection or not. As well we might not have enough information on the species to simulate its evolution under a given neutral demography. Therefore, we still need to identify if the value obtained for a given sample is due to the effect of selection. By doing so, we will compute two more measures before giving a diagnostic about the presence of divergent selection. The first is a sign test based on the lower bound of (4), the second is the F_{ST} of the SNP having the maximum *nvd* compared with the global F_{ST}.

Sign test

From a lower bound of (4) we derive the quantity called divergent selection sign (*dss*, see Appendix A-2 for details)

$$dss = \frac{4(n-1)S^2 - 2\sum_i \mathrm{hac}_{1i}^2}{nL^2} \qquad (5)$$

where hac_{1i} are the HAC values measured at each haplotype *i* in the partition 1 and the sum is over the n_1 sequences in that partition. A negative sign in (5) suggests that the value of *nvd* is not the result of divergent selection. Indeed, we require (5) to be positive to count a given candidate as significant.

Combined method: nvdF_{ST}

The sign test defined above is a good strategy for discarding some false candidates. However, we still lack a method for obtaining *p*-values associated to the sites chosen by the *nvd* algorithm. We can add a second measure to diagnose divergent selection by combining the information on candidate SNPs as given by *nvd* with the interpopulation differentiation measure at that site. The significance of the obtained quantity is far easier to assess. The joint use of these methods produces the combined measure *nvdF_{ST}*. The rationale of the approach is that if divergent selection acts on a specific site then the F_{ST} at that site will be higher compared to the overall F_{ST}. To obtain the *p*-value we do not perform an LK test (Lewontin & Krakauer 1973) because first, the candidate was not chosen for being an outlier and second we are assuming that the sites are not independent given that we are considering more or less linked haplotypes. Instead, we proceed as follows, let *i* be the candidate site chosen because it has the maximum *nvd* value, then we measure the index $I_i = F_{STi} - F_{ST}$ comparing the candidate with the overall F_{ST}. To get the *p*-value for a given I_i, the data is resampled several times to generate an empirical distribution. By doing so, the mean

frequency for every SNP in the pooled populations is considered as the expectation under the homogenizing effect of migration provided that $Nm > 1$ (Crow & Kimura 1970). Then, for any iteration, the probability of a given allele at each population is obtained from a binomial $B(p,n)$, where p is the mean allelic frequency at that site and n the local population sample size. The p-values correspond to the proportion of times that the resampled indexes were larger than I_i. Note that, for each site, the resampling procedure has variance pqn which will be larger at intermediate frequencies. For candidates with more or less similar frequencies at both populations we expect low index I_i values and correspondingly high p-values. When the pooled frequency is intermediate two situations are possible, first, each population has similar intermediate frequencies which again imply high p-values or alternatively, the frequencies can be extreme and opposite at each population. In the latter, I_i is high and its p-value low. Recall that we are looking for selection in populations connected by migration and working only with SNPs shared between them. Thus, the SNPs that are fixed in one of the populations are not considered.

The F_{ST} values were computed following the algorithm in Ferretti *et al* (Ferretti *et al.* 2013). The number of resamplings for each site was set to 500 times.

Effective number of independent SNPs, significance and q-value estimation

We have computed *nvd* and the F_{ST} index and got a candidate site with its p-value. Since *nvd* was obtained after testing a number of positions on a given window size, it is desirable to apply a multiple test correction for the number of independent SNPs in the window. To roughly estimate the number of independent SNPs, we calculate the linkage disequilibrium

measure D' (Devlin & Risch 1995; Lewontin 1988) at each pair of consecutive sites and then store the quantity $r' = 1 - |D'|$ for each pair. The effective number of independent SNPs (M_{effs}) between site w_{ini} and w_{end} is then obtained as one plus the summation of the r' values in the interval $[w_{\text{ini}}, w_{\text{end}})$. The Šidák correction (Cheverud 2001; Sidak 1967) can now be applied to get the corrected significance level $c = 1 - (1 - \gamma)^{1/M_{\text{effs}}}$ with nominal level γ (= 0.05 by default). Thus, the algorithm $nvdF_{ST}$ will finally suggest a candidate as significant only when its p-value (computed as explained in the previous section) is lower than c and the sign in (5) is positive.

False discovery rate and q-values (Storey 2003) has been proposed as unified approach for evaluating method performance in terms of false discoveries (De Villemereuil *et al.* 2014) and we estimate the q-value corresponding to each significant p-value (see Appendix A-3 for details on the calculation).

The *k*-means extreme outlier set test (*EOS*)

The $nvdF_{ST}$ method assumes the existence of a dense map of linked genetic markers. If the data consists mostly in independent markers this would provoke the failure to detect selection by the $nvdF_{ST}$ method because the HAC-based information will not exists. To deal with this situation, a second method was implemented consisting in a heuristic two-step procedure that performs a conservative test for identifying extreme outliers.

We intend our method to be conservative because, as mentioned, the variance of the F_{ST} distribution is quite unpredictable under a variety of scenarios. This provokes high rates of

false positives associated with the F_{ST} outlier tests. Therefore, our heuristic strategy takes advantage of the fact that, independently of the demographic scenario, the involved regions under divergent selection may produce extreme outliers that would be clustered apart from the neutral ones. Only when this kind of outliers is detected the subsequent LK test is performed.

The rationale of the algorithm is as follows:

The first step consists in computing the extreme positive outliers in the sense of Tukey i.e. those sites having a F_{ST} value higher than 3 times the interquartile range (Tukey 1977). The second step identifies different classes inside the extreme outlier set (EOS). This is done by a *k*-means algorithm (Schubert *et al.* 2012; Vattani 2011). Here, a *k*-modal distribution is assumed and all the elements of the set are classified in one of the *k* classes. The class with lower values is discarded and only the elements, if any, in the upper classes having values higher than a cutoff point are maintained in the set. By default $k = 2$ and two modes $\{0, F_{STu}\}$ were used corresponding to lower and upper bound for the F_{ST} estimator (see Appendix A-5). The cutoff is defined as the overall F_{ST} plus F_{STu} / 3 i.e. the mean plus the square root of the variance upper-bound under an asymmetric unimodal distribution (Dharmadhikari & Joag-Dev 1989). Finally, for each of the candidates remaining in the EOS the LK test (Lewontin & Krakauer 1973) is performed to compute its *p*-value. The Šidák correction (Cheverud 2001; Sidak 1967) for the number of remaining outliers in the set is applied to get the significance level.

Software description

Both *nvdF*$_{ST}$ and the EOS test have been implemented in the program HacDivSel. Complete details of the software can be found in the accompanying manual. We here just mention that the input program files may be in *MS* (Hudson 2002) of Fasta formats for the haplotype-based test or in Genepop (Rousset 2008) or BayeScan (Foll & Gaggiotti 2008) formats if the data do not include haplotype information. In any case the data should contain sequence samples from two populations. A typical command line for calling the program to analyze a file named *sel.txt* containing 50 sequences from each population would be

HacDivSel -input sel.txt -sample 50 -candidates 10 -SL 0.05 -output anyname -format ms

Where -*sample* is the sample size for each population and the label -*candidates* 10, indicates that the ten highest *nvd* values should be included in the output. The program would analyze the file and produce as output the highest 10 values and its significance at the 0.05 level for different window sizes after the *nvdF*$_{ST}$ test. It also performs the EOS test and gives the candidate outliers, if any, and their significance. Only the SNPs shared by the two populations are considered. Which imply that there are at least 4 copies of each SNP in the metapopulation.

Results

In what follows, the power of a test is measured as the % of runs where selection was detected from simulated selective scenarios and the false positive rate is measured as the % of runs where selection was detected from simulated neutral scenarios (see Appendix A-6

for details on the simulations setting). In addition, the accompanying *q*-value (Storey 2003) is an estimate from the data (see Appendix A-4 for details on the estimation process).

Combined Method ($nvdF_{ST}$)

Under a single locus architecture, the power of $nvdF_{ST}$ vary between 79-94% for both medium (60 SNPs/Mb) and high density (250 SNPs/Mb) maps (Table 1). These results can be compared with published analysis (Rivas *et al.* 2015). In the previous study the methods Svd, SvdM and OmegaPlus (Alachiotis *et al.* 2012) were evaluated with best powers of 63-79% obtained by Svd and SvdM in cases with high mutation and recombination (Rivas *et al.* 2015). When these methods were applied considering the pooled data from the merged populations then the best powers were attained by SvdM ranging from 42 to 94% (Rivas *et al.* 2015). Recall that the methods Svd, SvdM and OmegaPlus oblige the user to perform simulations of a neutral demography to obtain the *p*-values for the tests. As it can be appreciated from rows 1 to 6 in Table 1, that coincides with the scenarios of (Rivas *et al.* 2015), $nvdF_{ST}$ performs quite well (from 79 to 94%) without the need of performing additional neutral simulations. The given results are for 10,000 generations; the results for 5,000 generations were quite similar and are therefore omitted.

Under the polygenic architecture (n = 5 in Table 1) at least one candidate was found 99% of the times and more than one were found 80% of the time. However, the number of correctly identified sites was quite variable ranging between 1 and 3.

The last row in Table 1 corresponds to the case when all SNPs segregate independently. In this case, the method fails to detect selection which is not surprising because the information from the haplotype allelic classes is absent under linkage equilibrium; the adequate patterns are not found which provokes both a negative sign in the nvd test and a candidate with low F_{ST} index measure.

Table 1. Performance of the combined method ($nvdF_{ST}$) with n = 1 selective site located at the center of the chromosome or n = 5 (see Appendix A-6). Selection was α = 600 and Nm = 10. Mean localization is given in distance kb from the real selective position.

∑	θ	ρ	n	%Power	%FPR (γ = 5%)	q-value	Localization (kb)
65	12	0	1	87	2.1	0.0058	±458
63	12	4	1	94	2.7	0.0008	±200
60	12	12	1	90	1.0	0.0003	±33
251	60	0	1	79	1.8	0.0048	±60
232	60	4	1	84	6.2	0.0011	±17
249	60	60	1	86	2.4	0.0002	<±1
282	60	60	5	99	2.4	0.0002	<±1
318	60	∞	1	0	0	-	-

∑: Mean number of shared SNPs per Mb. θ: Mutation rate. ρ: Recombination rate. FPR: false positive rate. q-value: mean estimated q-value for the significant tests. ∞: Independently segregating sites.

Short-term Strong and Long-term Weak Selection Scenarios

The performance of *nvdF*$_{ST}$ under the strong selection scenario (α = 6000) is presented in Table 2. Not surprisingly, the number of segregating sites is considerably reduced. In fact the minimum window size allowed by the program had to be shortened from 51 to 25 to perform the analyses. The power of detection range between 44-67% with 0 false positive rate. These results can be compared with Svd and SvdM methods from Rivas *et al*. (Table 6, t = 500 generations Rivas *et al.* 2015). Those results were more dependent on the recombination rate having low powers (14-28%) under full linkage and great power (70-96%) with high recombination. Recall however that to assess significance with these methods the exact neutral demography was simulated by Rivas and coworkers.

Concerning very weak selection in long-term scenarios (Table 2, α = 140) the power varied between 49-52% with false positive rate between 2.2 and 5.7%.

Table 2. Performance of the combined method (*nvdF*$_{ST}$) with a single selective site in the short-term strong (α = 6000) and the long-term weak (α = 140) selection scenarios. *Nm* was 10. Mean localization is given in distance kb from the real selective position.

∑	θ	ρ	α	*t*	%Power	%FPR (γ = 5%)	*q*-value	Localization (kb)
112	60	0	6000	500	44	0	0	±66
32*	60	4	6000	500	63	0	0.0014	±5
62	60	60	6000	500	67	0	0.0008	±93

165	60	0	140	5,000	49	3.6	0.0280	±33
156	60	4	140	5,000	52	5.7	0.0219	±14
135	60	60	140	5,000	49	2.2	0.0054	±6

∑: Mean number of shared SNPs per Mb. θ: Mutation rate. ρ: Recombination rate. *t*: number of generations. FPR: false positive rate. *q*-value: mean estimated *q*-value for the significant tests. *: only 40 runs having a minimum of 25 SNPs.

Extreme Outlier Set Test (EOS)

The EOS test is quite conservative as can be observed in Table 3 where the false positive rate is below the nominal 0.05 in every case. Its power increases with the density and the independence of the markers reaching 61% of detection in the case of independent SNPs and maps with 250-300 SNPs/Mb. As expected for an outlier test, the power undergoes a breakdown under a polygenic setting (row with *n* = 5 in Table 3). Therefore, the EOS test is complementary to $nvdF_{ST}$ having its maximum power when the latter has its minimum and viceversa.

Table 3. Performance of the extreme outlier test (EOS) with *n* = 1 selective site located at the center of the chromosome or *n* = 5 (see Simulations section above). Selection was α = 600 and *Nm* = 10. Mean localization is given in distance kb from the real selective position.

∑	θ	ρ	n	%Power EOS	%FPR (γ = 5%)	q'-value	Localization (kb)
65	12	0	1	0	0	-	-

63	12	4	1	0.2	0	0.46	±3
60	12	12	1	1.1	0	0.45	±77
251	60	0	1	0.7	0	0.10	0
232	60	4	1	1.3	0	0.20	±150
249	60	60	1	58	0.4	0.5	<±1
282	60	60	5	1.6	0.4	0.49	±5
318	60	∞	1	61	1.2	3×10^{-6}	±7

∑: Mean number of shared SNPs per Mb. θ: Mutation rate. ρ: Recombination rate. FPR: false positive rate. q'-value: mean corrected (see appendix A-4) estimated *q*-value as estimated for the significant tests. ∞: independently segregating sites.

In the last three rows of Table 3, note that the false positive rate is indicating the percentage of outliers detected as selective after EOS test in a given neutral scenario. The *q*-value, however, refers to the minimum estimated false discovery rate (FDR) that can be committed when calling significant one test at a given threshold. It can be appreciated that for independently segregating sites the *q*-value is very low (3×10^{-6}) but rise up to 0.5 for the same scenario when markers are linked. In the case of unlinked markers the EOS is quite efficient and just detects the single true selective SNP. On the contrary, with linked markers the F_{ST} estimates are less reliable; more outliers are detected, which inflates the FDR and the corresponding *q*-values.

Position Effect

The ability to locate the position of the selective site increased with the marker density and the recombination rate (Table 1). The localization is given in kilobases away from the correct position. The values are averages through the runs. Standard errors are omitted since they were low, in the order of hundreds of bases or few kilobases (below 5) in the worst case (fully linked markers). Thus, when the target site is located at the centre of the studied region (Table 1) and the overall recombination rate is at least 0.3 cM/Mb ($\rho \geq 12$), the $nvdF_{ST}$ method performs acceptably well under weak selection ($\alpha \leq 600$), with the inferred location within 33 kb of distance from the true location in the worst case. However, under strong selection (Table 2, $\alpha = 600$), the localization is worst, 93 kb, but this could be due to the lower number of segregating sites (only 62 in Table 2).

The localization is also dependent of where the selective site is placed within the chromosome. The farther from the center the worse the ability to correctly localize the selective positions (Table 4). In this case, with recombination of 1.5 cM/Mb, the inferred location changes from an almost perfect localization (<1 kb from Table 1) to distances of 10-122 kb as the target is shifted away. This issue has already been shown for other HAC-based methods (Rivas *et al.* 2015). The problem is partially solved under high recombination using the EOS test because in such cases the selective sites are localized at distances ranging from few base pairs to 40 kb from its real position about 67-93% of the times (cases with $\rho = 60$ in Table 4). In the case of independent markers with the selective site located at the center (Table 3) the localization was perfect in 98% of the replicates but the average appears as ±7

kb because of two runs where the localization failed by almost 400 kb. These cases coincide with lower F_{ST} values that are marginally considered given the cutoff for selecting the SNPs thus making a bit more astringent the EOS classification in the upper class will already discard them. For example if we change the cutoff from $F_{ST} + F_{STu} / 3$ to $F_{ST} + 1.2F_{STu} / 3$ we decrease the power from 61 to 59% and get already perfect localization of the selective SNPs in every run.

Table 4. Performance of $nvdF_{ST}$ and EOS with a single selective site located at different positions. Selection was α = 600 and *Nm* = 10. Mean localization is given in distance kb from the real selective position. FPRs are the same as in Table 1. *q*-value refers to the mean *q*-value for the significant $nvdF_{ST}$ tests.

∑	θ	ρ	%Power $nvdF_{ST}$, EOS	Position (kb)	$nvdF_{ST}$ *q*-value	Localization (kb) $nvdF_{ST}$, EOS
259	60	0	81, 1	0	0.0044	+483, +457
255	60	0	81, 1.5	10	0.0049	+433, +496
256	60	0	82, 0.9	100	0.0041	+350, +413
255	60	0	78, 0.6	250	0.0039	±194, ±185
230	60	4	75, 2.5	0	0.0014	+324, +127
226	60	4	77, 3.5	10	0.0016	+326, +142
233	60	4	80, 1.8	100	0.0017	+227, +140

229	60	4	83, 1.6	250	0.0009	±123, ±20
262	60	60	63, 93	0	0.0014	+122, +40
261	60	60	68, 91	10	0.0014	+113, +34
257	60	60	81, 84	100	0.0006	±44, ±6
252	60	60	87, 67	250	0.0004	±10, ±0.06

∑: Mean number of shared SNPs per Mb. θ: Mutation rate. ρ: Recombination rate. Position: real position of the selective site.

Bottleneck-expansion Scenarios

Bottleneck-expansion scenarios are known to leave signatures that mimic the effect of positive selection. Thus, we tested the robustness of the methods by applying them to this kind of situation under a neutral setting (a reduction of the population to 1% of the original size, see details in Appendix A-6). Both algorithms performed well, $nvdF_{ST}$ false positive rate is maintained below the nominal level (4.6%) and for EOS test is only 1%.

High Migration Nm = 50 Scenarios

For the short-term (500 generations) scenario with *Nm* = 50, $nvdF_{ST}$ is still able to detect the effect of selection in spite of the homogenizing effect of migration. The detection power ranges between 34-59% with a false positive rate of 0-0.1% (Table 5). Therefore, the test is very conservative under this setting. Noteworthy the power diminishes with the highest

recombination rate. This may occur because the sign test is rejecting several cases due to the combined effect of gene flow and recombination that generates intermediate values of m_1 and m_2. Indeed, for a given selection intensity, the higher the *Nm* requires tighter linkage for the establishment of divergent alleles (Yeaman & Whitlock 2011). Therefore, the decrease in power for the higher *Nm* is not surprising. Concerning the EOS test it has no power to detect selection in the given scenario when *Nm* equals 50.

Table 5. Performance of $nvdF_{ST}$ in the short term (500 generations) with a single selective site. Selection was α = 600 and *Nm* = 50. Mean localization is given in distance kb from the real selective position.

∑	θ	ρ	%Power	%FPR (γ = 5%)	*q*-value	Localization (kb)
116	60	0	56	0	0.0098	±152
180	60	4	59	0	0.0042	±123
178	60	60	34	0.1	0.0096	±4

∑: Mean number of shared SNPs per Mb. θ: Mutation rate. ρ: Recombination rate. FPR: false positive rate. *q*-value: mean estimated *q*-value for the significant tests.

Comparison of EOS with BayeScan method

We have used BayeScan 2.1 (Foll & Gaggiotti 2008) to analyze the data corresponding to independent markers (last row in Table 1) and linked markers (ρ = 60, n=1, antepenultimate row in Table 1). Only SNPs shared between populations and with a minimum allele frequency (maf) of 2 per population (4%) were considered. The parameters for BayeScan

were the default ones. We have assessed as significant those runs having a Bayes factor higher than 3 (BF3) or those having a factor higher than 100 (BF100). In Figure 1 we may appreciate that with linked markers the power attained using BF3 is about 47% but at the cost of a prohibitive 26% of false positives. The situation is better with independent markers, still under the very stringent BF100, Bayescan maintains a power of 82% although the cost in false positives is still high (8%). On the contrary EOS maintains power of 60% both with linked and independent markers virtually without false positives.

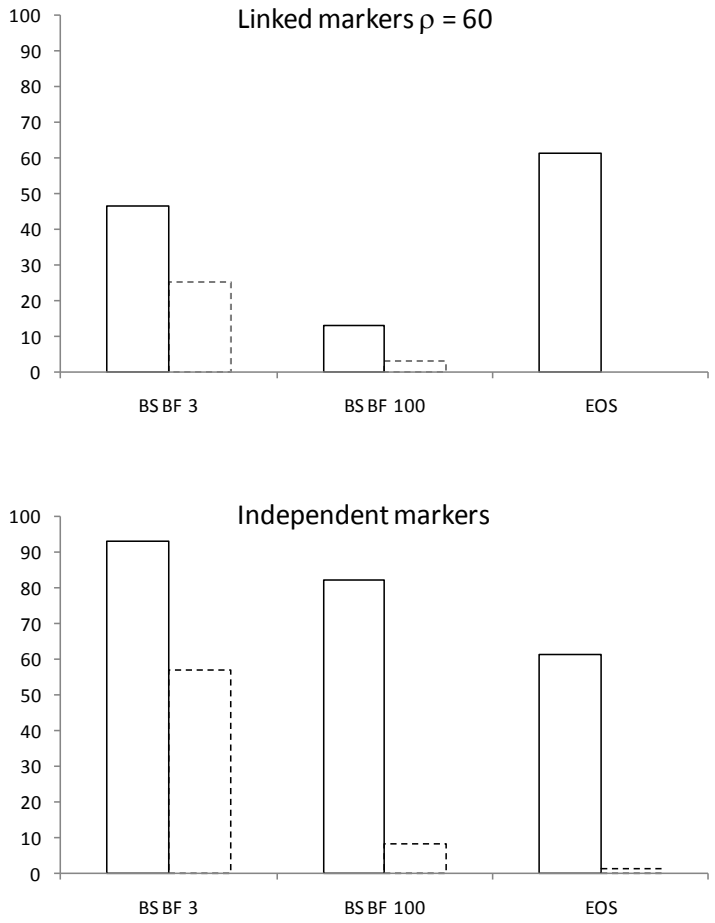

Figure 1. Comparison between BayeScan (BS) and EOS methods. BF: Bayes Factor. Continuous line: Power. Dashed line: False positive rate.

Empirical Data

We applied EOS to analyze a published data set from *Littorina saxatilis* species, concretely the separate-island filtered loci from Ravinet *et al.* (2016). We have discarded the loci with null allele frequency equal or higher than 0.5 jointly with those polymorphisms non-shared between ecotypes. We additionally require a maf of minimum 4 alleles per metapopulation sample size. Thus, we have excluded about 10-20% of the original individual-island filtered loci. The results of the outlier analysis between ecotypes using EOS are shown in Table 6. We may appreciate that the number of outliers detected as significant after EOS test is much less than in the original study since we find a total of 69 outliers in the three islands while they found 406 (RAD loci in Table 2 of Ravinet *et al.* 2016). This is not surprising given the conservative nature and low false positive rates of EOS. Note that, the shared by all outliers apart, Jutholmen and Ramsö share 2 outliers while Saltö has no outlier in common with Jutholmen and just 1 with Ramsö.

Table 6. Outliers detected after EOS analysis of individual-island filtered loci from *Littorina saxatilis* data (Ravinet *et al.* 2016).

Island	Unique	Only with Jutholmen	Only with Ramsö	Only with Saltö	Shared all	Total
Jutholmen	27	__	2	0	2	31
Ramsö	24	2	__	1	2	29
Saltö	6	0	1	__	2	9

For the outliers in EOS, the F_{ST} between ecotypes ranges between 0.4-0.6 (Table 7). The *q*-values are high (0.52 – 0.76) although we already saw in the simulations that this may be

indicating high linkage between the markers more than an inflated false positive rate (see also De Villemereuil *et al.* 2014).

Table 7. Summary of EOS analysis for the between ecotypes *Littorina saxatilis* data (Ravinet *et al.* 2016).

Island	Nonoutliers	Outliers not in EOS	EOS	F_{ST}	F_{ST_EOS}	$pval_{EOS}$	$qval_{EOS}$
Jutholmen	4564	91	31	0.045	0.40	0.004	0.52
Ramsö	4602	82	29	0.064	0.53	0.005	0.63
Saltö	4632	51	9	0.060	0.60	0.002	0.76

F_{ST}: Mean F_{ST} for the analyzed loci. F_{ST_EOS}: Mean F_{ST} for the loci included in the extreme outlier set. $pval_{EOS}$: Mean *p*-values across the loci included in the extreme outlier set. $qval_{EOS}$: Mean *q*-values across the loci included in the extreme outlier set.

Discussion

The goal in this work was to develop two methods, haplotype-based and outlier-based, for detection of divergent selection in pairs of populations connected by migration with the requisite of being protected from false positives which is a known concern for differentiation-based methods (De Mita *et al.* 2013; De Villemereuil *et al.* 2014; Lotterhos & Whitlock 2014). Additionally, the methods should be useful for non-model species and it should not be necessary to perform neutral simulations to obtain critical cut-off values for the candidates.

For the first method, it has been shown that combining haplotype-based and F_{ST} differentiation information, the so-called $nvdF_{ST}$, is a quite powerful strategy for detecting divergent selection. However, when the whole set of markers is segregating independently

there is no haplotypic information. Therefore, a second method was developed based on the idea that outliers due to the effect of divergent selection would cluster apart from those caused by different demography issues. This extreme outlier set test, EOS, was intended to be conservative because the mentioned tendency of outlier-based methods to produce false positives. Under the simulated scenarios, the EOS test behaves acceptably well when markers are independent or under weak linkage, reaching powers between 60-90% while maintaining false positive rate below the nominal level. On the contrary, BayeScan, one of the state-of-the-art genome-scan methods, commits 8% of false positives in the best case.

Polygenic Architecture

In general, the F_{ST}-based methods cannot detect selection in polygenic scenarios (Bierne *et al.* 2013; De Villemereuil *et al.* 2014) because those tests are specifically designed for finding larger than average F_{ST} values which are difficult to discover if the frequency differences are slight for the polygenic loci. On the contrary, the $nvdF_{ST}$ performs even better in this scenario because the distributed selective signal facilitates the discovery of the corresponding patterns by *nvd*. Since only the F_{ST} of the specific site indicated by *nvd* is compared with the overall F_{ST} and the null distribution is obtained using inter-population mean frequencies, the $nvdF_{ST}$ maintains high power under the polygenic setting for detecting at least one selective site.

Position Effect

Besides the detection of the signal of selection, we have also inferred the location of the selective site. It has been shown that under $nvdF_{ST}$ the localization is better when the selective site is at the center of the chromosome. The EOS test is not so affected by the position of the selective site. The ability of localizing the selective position is still a pending issue for many of the selection detection methods. There is also plenty of room for improvement under the $nvdF_{ST}$ and EOS methods in this regard, for example, trying to further explore the relationship between recombination and the window sizes producing the highest scores. Indeed, the interplay among divergent selection, recombination, drift and migration should be considered for further improving the efficiency of the methods.

Empirical Data

Local adaptation can occur most probably due to alleles with large effect but also under a polygenic architecture (Whitlock 2015; Yeaman 2015). In the latter, it is possible that the genes responsible for the adaptation be transient so that they vary over time. Hence, the geographic structure and the migration selection balance can generate complex patterns on the distribution of genetic variation (Debarre *et al.* 2015). Thus, there is great complexity in the natural systems where local adaptation occurs (Whitlock 2015). That been said, the *L. saxatilis* ecotypes are an especially interesting system to study local adaptation in presence of gene flow (Johannesson 2015). This system has an exceptional level of replication at different extent as country, distinct localities within country and finally the micro-geographical level of the ecotypes. In the case of the Swedish populations, the pattern of differentiation can be separated in factors such as localities and habitats variation among

islands, that may be caused by genetic drift, and variation between habitats, that may be caused by divergent selection (Johannesson 2015). There are also different mechanisms by which parallel adaptation may occur, resulting in different predictions about the proportion of shared adaptive variation among localities. Moreover, the combined effect of population structure with the local adaptation pressures may help to explain the proportion of shared outliers that can be encountered when studying such parallel adaptation processes. Regarding the *L. saxatilis* system in Spanish and Swedish populations it seems to involve a small proportion of shared genomic divergence (Hollander *et al.* 2015; Johannesson 2015; Ravinet *et al.* 2016). The EOS analysis of Ravinet *et al.* data supports their finding that the majority of genomic variation linked to the evolution of ecotypes is not shared between the studied islands. At the same time, we identify far fewer outliers, with Saltö having the lowest number. This may explain our results showing reduced shared divergence between Saltö and the two other islands. On the contrary, Ravinet *et al.* find more outliers in the Saltö data and correspondingly, increased numbers of shared outliers between Saltö and the other islands. This could be due to an excess of false positive outliers hiding the pattern or alternatively our finding can be an artefact due to the low number of outliers detected by EOS.

As a conclusion, $nvdF_{ST}$ combines haplotype and population differentiation information and may be a helpful tool to explore patterns of divergent selection when knowledge of the haplotype phase is at hand. Alternatively, the EOS method is a conservative outlier test useful when the full set of SNPs is unlinked or under weak linkage. Both strategies have low false positive rate and can be applied without the need of performing neutral simulations.

Bibliography

Akey JM (2009) Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research* **19**, 711-722.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a High-Density SNP Map for Signatures of Natural Selection. *Genome Research* **12**, 1805-1814.

Alachiotis N, Stamatakis A, Pavlidis P (2012) OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics* **28**, 2274-2275.

Bierne N, Roze D, Welch JJ (2013) Pervasive selection or is it…? why are FST outliers sometimes so frequent? *Molecular Ecology* **22**, 2061-2064.

Bonhomme M, Chevalet C, Servin B*, et al.* (2010) Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. *Genetics* **186**, 241-262.

Bourret V, Kent MP, Primmer CR*, et al.* (2013) SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (Salmo salar). *Molecular Ecology* **22**, 532.

Carvajal-Rodriguez A (2008) GENOMEPOP: A program to simulate genomes in populations. *BMC Bioinformatics* **9**, 223.

Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory* Harper & Row, New York.

Charlesworth B, Charlesworth D (2010) *Elements of evolutionary genetics* Roberts and Company Publishers, Greenwood Village, Colo.

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. *Genome Research* **20**, 393-402.

Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* **87**, 52-58.

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

De Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology* **23**, 2006-2019.

Debarre F, Yeaman S, Guillaume F (2015) Evolution of Quantitative Traits under a Migration-Selection Balance: When Does Skew Matter?*. *The American Naturalist* **0**, S000.

Devlin B, Risch N (1995) A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics* **29**, 311-322.

Dharmadhikari SW, Joag-Dev K (1989) Upper bounds for the variances of certain random variables. *Communications in statistics-theory and methods* **18**, 3235-3247.

Duforet-Frebourg N, Bazin E, Blum MGB (2014) Genome Scans for Detecting Footprints of Local Adaptation Using a Bayesian Factor Model. *Molecular Biology and Evolution* **31**, 2483-2495.

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

Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013) Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. *Genetics* **193**, 929-941.

Ferretti L, Ramos-Onsins SE, Pérez-Enciso M (2013) Population genomics from pool sequencing. *Molecular Ecology* **22**, 5561-5576.

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

Frichot E, Schoville SD, Bouchard G, FranÃ§ois O (2013) Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution* **30**, 1687-1699.

Friguet C (2012) A general approach to account for dependence in large-scale multiple testing. *Journal de la Societé Francaise de Statistique* **153**, 100-122.

Hollander J, Galindo J, Butlin RK (2015) Selection on outlier loci and their association with adaptive phenotypes in Littorina saxatilis contact zones. *Journal of Evolutionary Biology* **28**, 328-337.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338.

Hussin J, Nadeau P, Lefebvre J-F, Labuda D (2010) Haplotype allelic classes for detecting ongoing positive selection. *BMC Bioinformatics* **11**, 65.

Johannesson K (2015) What can be learnt from a snail? *Evolutionary Applications*, n/a.

Lewontin RC (1988) On measures of gametic disequilibrium. *Genetics* **120**, 849-852.

Lewontin RC, Krakauer J (1973) Distribution of Gene Frequency as a Test of Theory of Selective Neutrality of Polymorphisms. *Genetics* **74**, 175-195.

Li J, Li H, Jacobsson M*, et al.* (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology* **21**, 28-44.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology* **23**, 2178.

Maruki T, Kumar S, Kim Y (2012) Purifying Selection Modulates the Estimates of Population Differentiation and Confounds Genome-Wide Comparisons across Single-Nucleotide Polymorphisms. *Molecular Biology and Evolution* **29**, 3617-3623.

Meinshausen N, Rice J (2006) Estimating the Proportion of False Null Hypotheses among a Large Number of Independently Tested Hypotheses. *The Annals of Statistics* **34**, 373.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* **70**, 3321-3323.

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

Ravinet M, Westram A, Johannesson K*, et al.* (2016) Shared and nonshared genomic divergence in parallel ecotypes of Littorina saxatilis at a local scale. *Molecular Ecology*, 287-305.

Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L (2011) SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along

gradients of ecological speciation in lake whitefish species pairs (Coregonus spp.). *Molecular Ecology* **20**, 545.

Rivas MJ, Dominguez-Garcia S, Carvajal-Rodriguez A (2015) Detecting the Genomic Signature of Divergent Selection in Presence of Gene Flow. *Current Genomics* **16**, 203-212.

Rolan-Alvarez E (2007) Sympatric speciation as a by-product of ecological adaptation in the Galician Littorina saxatilis hybrid zone. *Journal of Molluscan Studies* **73**, 1-10.

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

Schubert E, Zimek A, Kriegel H-P (2012) Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery* **28**, 190-237.

Sidak Z (1967) Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association* **62**, 626-633.

Storey J (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* **31**, 2013-2035.

Storey JD (2001) Estimating false discovery rates under dependence, with applications to DNA microarrays.

Storey JD (2002) A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**, 479.

Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **66**, 187-205.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445.

Thibert-Plante X, Gavrilets S (2013) Evolution of mate choice and the so-called magic traits in ecological speciation. *Ecol Lett*.

Tukey JW (1977) *Exploratory data analysis* Addison-Wesley, Reading, Mass.

Vattani A (2011) k-means Requires Exponentially Many Iterations Even in the Plane. *Discrete & Computational Geometry* **45**, 596-616.

Whitlock M, C., Lotterhos K, E. (2015) Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of FST. *The American Naturalist* **186**, S24-S36.

Whitlock MC (2015) Modern Approaches to Local Adaptation. *The American Naturalist* **186**, S1.

Yeaman S (2015) Local Adaptation by Alleles of Small Effect*. *The American Naturalist* **186**, S74-S89.

Yeaman S, Otto SP (2011) Establishment and Maintenance of Adaptive Genetic Divergence under Migration, Selection, and Drift. *Evolution* **65**, 2123-2129.

Yeaman S, Whitlock MC (2011) The Genetic Architecture of Adaptation under Migration-Selection Balance. *Evolution* **65**, 1897-1911.

Software and Data Accessibility

The computer program HacDivSel implementing the methods explained in this article jointly with the user manual, are available from the author web site http://acraaj.webs.uvigo.es/software/HacDivSel.zip.

Data: DRYAD entry doi: .

Acknowledgements

I thank E. Rolán-Alvarez for useful comments on the manuscript. This work was supported by Ministerio de Economía y competitividad (CGL2012-39861-C02-01 and BFU2013-44635-P), Xunta de Galicia (Grupo con Potencial de Crecimiento, GPC2013-011) and fondos FEDER. The author declares to have no conflict of interest.

Appendix

A-1) Effect of window size on variance difference original statistics

We can appreciate the effect of a window size *L* on the computation of the original *gSvd* measure as follows. Recall that the HAC distance *d* between and haplotype *h* and a reference *R* both of length *L* is

$$d = \sum_{i=1}^{L} I(h_i \neq R_i)$$

where *I*(A) is the indicator function of the event *A*. Thus, $d \in [0, L]$ so that, given an increase of the window size by *Q* (*Q* > 1), then $d \in [0, QL]$. Therefore, the change in window size is a change in the scale of the HAC distances. Depending on the distribution under the new window size the magnitude of the change in the scale can be *Q* or more generally $Q' \in (1, Q]$. Thus, a window size increase of *Q* has a quadratic impact onto s^2 and Δ as defined in (1). Then, if we define *gSvd* under L_A, we have

$$gSvd_i = \frac{V_{2i} - V_{1i}}{L_A} \times f_i(1 - f_i)^a \times b$$

and if we change to window size $L_B = QL_A$ we might have

$$gSvd_{LB} = QgSvd_{LA}$$

For the equation to be exact it is also necessary that the change of window size do not alter the frequency distribution so that the relationship $v_B = Q^2 v_A$ and $\Delta_B = Q^2\Delta_A$ holds on, if not, the change will be better defined by $Q' \in (1, Q]$. In any case this explains why the normalization of gSvd by *L* is not very effective on avoiding a systematic increase of the statistic under higher window sizes (Hussin *et al.* 2010) (Rivas *et al.* 2015).

A-2) General variance difference

Consider the frequencies of a given haplotype *i* in the partition 1 and 2

$$f_{i1} = \frac{n_{i1}}{n_1} \qquad f_{i2} = \frac{n_{i2}}{n_2} \qquad f_i = \frac{n_{i1}+n_{i2}}{n} = \frac{f_{i1}n_1 + f_{i2}n_2}{n} \qquad \text{(A-2-1)}$$

Let d_i be the HAC distances for each haplotype *i* and with some abuse of notation *F*, F_1, F_2 the frequency distribution in the whole sample and in the partitions P_1 and P_2 respectively.

$$m = \sum_{i}^{n} \frac{d_i}{n} = \sum^{F} d_i f_i = \sum^{F} d_i \frac{f_{i1}n_1 + f_{i2}n_2}{n} = \sum^{F1} d_i \frac{f_{i1}n_1}{n} + \sum^{F2} d_i \frac{f_{i2}n_2}{n}$$

Note that

$m_1 = \sum^{F1} d_i f_{i1}$ and $m_2 = \sum^{F2} d_i f_{i2}$ and then

$m = \frac{n_1 m_1 + n_2 m_2}{n}$ (A-2-2)

Now consider the variance

$v = \sum^{F} (d_i - m)^2 f_i = \sum^{F} d_i^2 f_i - m^2$ (A-2-3)

$v_1 = \sum^{F1} d_i^2 f_{i1} - m_1^2 \qquad v_2 = \sum^{F2} d_i^2 f_{i2} - m_2^2.$

Substituting (A-1) in (A-3) and after some rearrangement we finally get

$v - \bar{V} = \Delta$ (A-2-4),

where n_1 and n_2 are the sample sizes and v_1 and v_2 the variances at each partition, $\bar{V} = \frac{n_1 v_1 + n_2 v_2}{n}$ and $\Delta = \frac{n_1 n_2}{n^2}(m_1 - m_2)^2 = \frac{n_1 n_2}{n^2}\Delta_m^2$.

Note that max(Δ) = $L^2/4$ = max(v) and min(Δ) = 0.

If we consider the sampling variance S^2 instead of the variance we have similarly

$$S^2 - \bar{S} = \frac{n}{n-1}\Delta \qquad \text{(A-2-5)}$$

being $\bar{S} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n-1}$.

From (A-2-5) and defining k as the fraction of sequences in the minor partition then n_1 = (1-k)n and n_2 = kn with $k \in$(MAF, 0.5) then we can express s^2 as

$S^2 = \frac{[(1-k)n-1]S_1^2 + (kn-1)S_2^2}{n-1} + \frac{n(1-k)k}{n-1}\Delta_m^2$ and

$$S_2^2 = \frac{(n-1)S^2 - [(1-k)n-1]S_1^2 - n(1-k)k\Delta_m^2}{kn-1}$$

So that the variance difference is broken down in the sum of two terms

$S_2^2 - S_1^2 = \frac{(n-1)S^2 - (n-2)S_1^2}{kn-1} - \frac{n(1-k)k\Delta_m^2}{kn-1}$

Reordering terms we have

$\frac{S_2^2 - S_1^2}{(1-k)k} = \frac{(n-1)S^2 - (n-2)S_1^2}{(kn-1)(1-k)k} - \frac{n\Delta_m^2}{kn-1}$ (A-2-6)

Realize that the first term in the sum is contributing to increase the variance difference whenever (n-1)S^2 ≥ (n-2)S_1^2. Note also that (1-k)k in the denominator has it maximum value when k = 0.5. In the second term, Δ_m^2 increases with directional selection (m_1 => 0 because the haplotypes in P_1 are expected, by definition, to be closer to the reference configuration) while kn (= n_2) decreases, so both are contributing to increase the negative term under

selection and diminish the value of the statistic. Thus, it is adequate to discard the second term in the variance difference (A-2-6). Now, recall the generalized Svd (gSvd, see Model section) defined for any SNP *i* as

$$gSvd_i = \frac{V_{2i} - V_{1i}}{L} \times f_i(1 - f_i)^a \times b$$

and, after discarding the second term in (A-2-6) substitute the $(V_{2i} - V_{1i})/L$ term in $gSvd_i$ to obtain

$$vd_i = \frac{(n-1)S^2 - (n-2)S_1^2}{kn-1} \times f_i(1 - f_i)^a \times b \quad \text{(A-2-7)}$$

that corresponds to formula (2) in the main text.

A-3) Lower bound and sign test

If the selective gene is at intermediate frequencies then $4f(1-f)$ would be close to 1, $n_1 = n_2 = n/2$ and the maximum variance in the first partition is $(n-2)\ S_1^2{}_{max} = nL^2/4$. By substituting in (4) we get

$$\frac{4(n-1)S^2 - nL^2}{nL^2} \quad \text{(A-3-1)}$$

which is a lower bound for *nvd* under a given S^2. Note that the variance in the first partition should not be at its maximum if selection is acting. Therefore a value as low as in (A-3-1) is not expected under selection. The lower bound still depends on the variance in the second partition and on the absolute value of the difference between the partition means $|m_1 - m_2|$. If the variance in the second partition is maximum it will be equal to the variance in the first and (A-3-1) becomes zero. With small variance in the second partition, the lower bound

will be negative only if $|m_1 - m_2|$ is low, just like can be expected under neutrality. Note that, if $n_1 = n/2$, (A-3-1) is equal or lower than

$$\frac{4(n-1)S^2 - 2\sum_i \mathrm{hac}_{1i}^2}{nL^2} \qquad \text{(A-3-2)}$$

where hac_{1i} are the HAC values measured at each haplotype *i* in the partition 1 and the sum is over the n_1 sequences in that partition. However if $n_1 > n/2$, the quantity in (A-3-2) could be higher or lower than (A-3-1) depending on the HAC values of the first partition. In any case, a negative value in (A-3-2) may be caused by m_1 being equal or higher than m_2 and suggests, whether it be $n_1 = n/2$ or not, that the value of *nvd* is not the result of divergent selection. Indeed, we call (A-3-2) the divergent selection sign (*dss*, formula 5 in the main text) and require it to be positive to count a given candidate as significant.

A-4) Bounds on FDR and q-value estimation

For a given test *i* with *p*-value p_i the FDR and the *q*-value after performing *S* tests is just (Storey 2002; Storey & Tibshirani 2003)

$$\mathrm{FDR}(p_i) = p_i * \pi_0 * S / max(\#\{p \leq p_i\}, 1)$$

where π_0 is the proportion of true nulls and $\#\{p < p_i\}$ corresponds to the position of p_i in the sorted (in ascending order) list of *p*-values. Then the *q*-value for p_i is obtained as the minimum of the FDR set for the *p*-values equal or higher to p_i.

$$q(p_i) = min\{\mathrm{FDR}(t)\} \text{ with } t \geq p_i.$$

In this work we estimated π_1, i.e. the proportion of the false null hypothesis using a method that is specially aided for cases when this proportion is very small (Meinshausen & Rice 2006) and then we obtained $\pi_0 = 1 - \pi_1$. This fits our expectation of detecting some few positions in the genome belonging to the alternative non-neutral distribution.

Lower bound

In the EOS test and because we have a sample-dependent upper bound G^*_{STmax} for the G_{ST} estimator we can correct for the minimum *q*-value achievable in that sample. Then for a given sample with G_{ST} mean *m,* the lower-bound *p*-value for the G_{ST} test *f* will be a function $p_{LB} = f(G^*_{STmax}, m) \geq 0$ that can be computed for each sample and consequently we can guess a minimum *q*-value. Therefore for the lower bound *p*-value, p_{LB} we have

$$FDR(p_{LB}) = p_{LB} * \pi_0 * S/max(\#\{p \leq p_{LB}\}, 1) \leq p_{LB} * \pi_0 * S$$ so that a lower-bound FDR is

$$FDR_{LB} = p_{LB} * \pi_0 * S/max(\#\{p \leq \alpha\}, 1) < FDR(p_{LB})$$

since $\#\{p \leq p_{LB}\} < \#\{p \leq \alpha\}$ and provided that $p_{LB} < \alpha$.

Note that with $\pi_0 < 1$ i.e. some non-null may exist and because p_{LB} is above 0 then we cannot reach an FDR of 0 still in the case when the lowest *p*-value corresponds to the true effect.

Now, let

$q(p_{LB}) = min\{FDR(t)\}$ with $t \geq p_{LB}$ then a lower bound for the *q*-value is

$$q_{LB} = min\{ q(p_{LB}), FDR_{LB} \}.$$

Upper bound

Let $p_i = \alpha$ then we get FDR(α) = $\alpha * \pi_0 * S/max(\#\{p \leq \alpha\}, 1)$ and $q(\alpha) = min\{FDR(t)\}$ with $t \geq \alpha$ is the minimum FDR that can be committed when calling significant a given test at this threshold. If the *p*-values are uniformly distributed the expected $q(\alpha)$ is π_0 however if the distribution is weighted towards 0, as expected when we have a mixture of null and alternative distributions, then $q(\alpha) < \pi_0$ is expected. Now, if we set a threshold of 1 i.e. the whole distribution of values, now we get necessarily a FDR of π_0 as

$q(1) = FDR(1) = 1 * \pi_0 * S/S = \pi_0$ and $\pi_0 \leq 1$ so that the upper bound is $q_{UB} = 1$.

Corrected q value

Under some circumstances may be of interest to correct for the bias effect of not being able of reaching minimum *p*-values due to the sample-dependent G_{ST} upper-bound. Thus we define

$$q'(p_i) = (q(p_i) - q_{LB}) / (1 - q_{LB})$$

Dependence

When considering many SNPs through the genome, the condition of independence is rarely maintained. In general, FDR-based estimates become more conservative as the dependence is stronger (Friguet 2012; Storey 2001; Storey *et al.* 2004). An important aspect when computing the FDR and associated *q*-values is the estimation of the proportion of true null

hypotheses, π_0. As indicated above we estimate π_0 through π_1. The impact of dependence structures on the estimation of π_1 has proved to be negligible compared to the conservative impact of the FDR estimation. This is not surprising because most SNPs belongs to the true null distribution so we do not expect the density of correlated values to be weighted towards 0, thus for p_i sufficiently low it should be true that $\#\{p \leq p_i\} < Sp_i$ i.e. $FDR(p_i) > \pi_0$ and we tend to have conservative estimates of FDR. We have confirmed this when comparing *q*-value estimates from dependent versus independent data. For example, comparing *q*-values for the EOS test in files with linked SNPs (ρ = 60) versus files with non-linked SNPs we obtained q=q'= 2.4 x 10^{-6} on average when markers are independent versus q= 0.63, q' = 0.5 when each pair of markers are linked with recombination of 0.015x10^{-6} (ρ = 60).

A-5) Lower and upper bounds for G_{ST} and F_{ST} estimators

Let Np be the number of populations, na is the number of alleles, *maf* is the minimum relative allelic frequency and n_i is the sample size for population i.

a) G_{STmax}

For G_{ST} (Nei 1973) we develop the formulas just for the one locus case without loss of generality. We will obtain the maximum G_{ST} noted as G_{STmax} and an upper bound noted as

$G^*_{STmax} = 1 - H^*_{smin}/H^*_{Tmax}$ (A-5-1)

Where

$$\mathrm{H}^*_{\mathrm{smin}} = 1 - (1 - maf)^2 - (maf)^2$$

and

$$H^*_{Tmax} = \mathrm{H}^*_{\mathrm{smin}} + (1 - maf)^2 \left(\frac{Np - 1}{Np}\right) + (maf)^2 \left(\frac{Np - 1}{Np}\right)$$

Let $G_{ST} = 1 - H_s/H_t$ with $H_s = 1 - \sum_{i=1}^{na} p_i^2$ averaged for the different populations and H_t is the same computation performed with the allelic metapopulation frequencies (Charlesworth & Charlesworth 2010). Thus, we are interested in computing the maximum G_{ST} when the minimum allele frequency (*maf*) is not 0 and additionally want to show that G^*_{STmax} is an upper bound of such value independently of the number of alleles considered. In doing so, we first compute the minimum for the subpopulation heterozigosity then we compute the maximum for the pooled heterozygosity and then we compute the maximum G_{ST} using these two values. This maximum will depend on the number of alleles segregating at each population. Finally we demonstrate that (A-5-1) is an upper-bound for G_{ST} whatever the number of alleles.

Let first look for the minimum H_s at each population. Usually this occurs when one allele is at maximum frequency i.e. 1 giving H_s=0. However in our case the maximum frequency is 1 – *maf* and the sum of frequencies can be expressed as

$$\sum_{i=1}^{na} p_i = (1 - maf) + \sum_{2}^{na} p_j \text{ with } \sum_{2}^{na} p_j = maf$$

Therefore

$$H_{\mathrm{smin}} = 1 - \sum_{i=1}^{na} {p_i}^2 = 1 - (1 - maf)^2 - \sum_{2}^{na} {p_j}^2$$

because $\left(\sum_{j=2}^{na} p_j\right)^2 = \sum_{j=2}^{na} p_j^2 + 2\sum_{\substack{j=2 \\ k>j}}^{na} p_j p_k$ then we can rewrite

$$H_{\text{smin}} = 1 - \sum_{i=1}^{na} p_i{}^2 = 1 - (1-maf)^2 - \left(\sum_{j=2}^{na} p_j\right)^2 + 2\sum_{\substack{j=2 \\ k>j}}^{na} p_j p_k$$

$$H_{\text{smin}} = 1 - (1-maf)^2 - (maf)^2 + 2\sum_{\substack{j=2 \\ k>j}}^{na} p_j p_k$$

The average for *Np* populations (for convenience of notation from herein we use in the summatory $j \neq k$ instead j=2 and k>j)

$$\overline{\mathrm{H}}_{\text{smin}} = 1 - (1-maf)^2 - (maf)^2 + \frac{2\sum_{j\neq k}^{na} p_j p_k + 2\sum_{j\neq k}^{na} p_j^{'} p_k^{'} + \ldots + 2\sum_{j\neq k}^{na} p_j^{Np-1} p_k^{Np-1}}{Np}$$

If *na* = 2 then

$$\overline{\mathrm{H}}_{\text{smin}} = 1 - (1-maf)^2 - (maf)^2 = \mathrm{H}^*_{\text{smin}}$$

if *na* > 2 then is obvious that $\overline{\mathrm{H}}_{\text{smin}} > \overline{\mathrm{H}}^*_{\text{smin}}$.

Now we are interested in H_{Tmax} i.e. the maximum pooled heterozygosity. It is easy to see that the maximum H_t occurs when the highest frequency allele at each population is at its maximum i.e. 1-*maf* and there are no shared alleles between populations. Therefore, for the case of two populations we have

$\sum_{i=1}^{na} p_i = (1-maf) + \sum_{2}^{na} p_j$ for the first population and

$\sum_{i=1}^{na} p_i^{'} = (1-maf) + \sum_{2}^{na} p_j^{'}$ for the second population and so on if there are more populations.

After pooling we have the sum of frequencies in the whole metapopulation

$$\sum_{i=1}^{na} \frac{p_i + p_i^{'} + .. + p_i^{Np-1}}{Np} =$$

$$= \frac{(1-maf)}{Np} + \frac{(1-maf)}{Np} + \cdots + \frac{(1-maf)}{Np} + \frac{\sum_2^{na} p_j}{Np} + \frac{\sum_2^{na} p_j^{'}}{Np} + \cdots + \frac{\sum_2^{na} p_j^{Np-1}}{Np}$$

Thus noting now p_i as the pooled frequency of allele *i*

$$H_t = 1 - \sum_{i=1}^{Np*na} p_i^2 =$$

$$= 1 - \frac{(1-maf)^2}{Np^2} - \cdots - \frac{(1-maf)^2}{Np^2} - \frac{\left(\sum_2^{na} p_j\right)^2}{Np^2} - \cdots - \frac{\left(\sum_2^{na} p_j^{Np-1}\right)^2}{Np^2} =$$

$$= 1 - \frac{(1-maf)^2}{Np} - \frac{\left(\sum_2^{na} p_j\right)^2}{Np^2} - \cdots - \frac{\left(\sum_2^{na} p_j^{Np-1}\right)^2}{Np^2} =$$

$$= 1 - (1-maf)^2 + (1-maf)^2 - \frac{(1-maf)^2}{Np} - \frac{\left(\sum_2^{na} p_j\right)^2}{Np^2} - \cdots - \frac{\left(\sum_2^{na} p_j^{Np-1}\right)^2}{Np^2} =$$

$$= 1 - (1-maf)^2 + (1-maf)^2\left(\frac{Np-1}{Np}\right) - \frac{\left(\sum_2^{na} p_j\right)^2}{Np^2} - \cdots - \frac{\left(\sum_2^{na} p_j^{Np-1}\right)^2}{Np^2} =$$

and rearranging terms for *maf* in a similar way as we did for H_s we finally get

$$H_{Tmax} = \mathrm{H}^*_{\mathrm{Tmax}} + \frac{2\sum_{j\neq k}^{na} p_j p_k + 2\sum_{j\neq k}^{na} p_j^{'} p_k^{'} + \ldots + 2\sum_{j\neq k}^{na} p_j^{Np-1} p_k^{Np-1}}{Np} \qquad \text{(A-5-2)}$$

with

$$H^*_{Tmax} = \mathrm{H}^*_{\mathrm{smin}} + (1-maf)^2\left(\frac{Np-1}{Np}\right) + (maf)^2\left(\frac{Np-1}{Np}\right)$$

We already has obtained the maximum G_{ST} which is $G_{STmax} = 1 - H_{smin}/H_{Tmax}$. Now we need only to show that $H_{smin}/H_{Tmax} > H^*_{smin}/H^*_{Tmax}$ (A-5-3). First note that $H_{smin} = H^*_{smin} + C$

with $C = \frac{2\sum_{j\neq k}^{na} p_j p_k + 2\sum_{j\neq k}^{na} p_j^{'} p_k^{'} + \ldots + 2\sum_{j\neq k}^{na} p_j^{Np-1} p_k^{Np-1}}{Np}$.

Similarly $H_{Tmax} = H^*_{Tmax} + C$ and from (A-5-2) we appreciate that $H^*_{smin} < H^*_{Tmax}$ so we can express $H^*_{Tmax} = kH^*_{smin}$ with $k>1$. We will proof (A-5-3) by contradiction so let assume that $H_{smin}/H_{Tmax} \leq H^*_{smin}/H^*_{Tmax}$ this implies that

$(H^*_{smin} + C)/ (kH^*_{smin} + C) \leq H^*_{smin}/ kH^*_{smin}$ rearranging terms we get $k \leq 1$which is false. Thus, $H_{smin}/H_{Tmax} > H^*_{smin}/H^*_{Tmax}$ and therefore $G_{STmax} = 1 - H_{smin}/H_{Tmax} < G^*_{STmax} = 1 - H^*_{smin}/H^*_{Tmax}$ so G^*_{STmax} is an upper bound of G_{STmax}.

b) G_{STmin}

It is quite immediate to show that $G_{STmin} = 0$. Consider a scenario in which every population has the same heterozygosis with the same alleles then $H_s = H_T$ and $G_{STmin} = 0$ and this is in fact the minimum and the lower bound.

3) F_{STmax}

For a sequence of biallelic SNPs we will use the F_{ST} estimation as defined in (Ferretti *et al.* 2013) to obtain as an upper bound

$$F_{STmax} = \frac{(Np-1)\left[Np - 2(1-maf)maf\sum_{k=1}^{Np}\frac{n_k}{n_k-1}\right]}{Np(Np-1) + 2(1-maf)maf\sum_{k=1}^{Np}\frac{n_k}{n_k-1}}$$

Thus we proceed as follows; first we will show that the maximum pooled heterozygosity depends on the mean population heterozygosity. Then we compute the minimum for the

subpopulation heterozigosity and show that the corresponding pooled heterozygosity is a maximum so finally the maximum F_{ST} is again $F_{STmax} = 1 - H_{smin}/H_{Tmax}$.

In (Ferretti *et al.* 2013) H_T is defined as

$$H_T = \frac{\bar{H}_S}{Np} + \frac{2}{Np^2}\sum_{k=2}^{Np}\sum_{k'=1}^{k-1}\theta_{\pi a}\,(k,k')$$

So, for computing H_{Tmax} we first seek for the maximum $\theta_{\pi a}$. This maximum will occur when sequences between populations are completely different. Because there are only two alleles and the minimum allele frequency is not 0 but *maf* the value $\theta_{\pi a}$ computed in this way will be an upper bound and the real maximum would be more or less close to that depending on the relationship between the sample size *n* and the sequence length *L*. In any case this upper bound is valid to ensure an upper bound for F_{STmax}.

$max\theta_{\pi a} = \frac{n_i n_j L}{n_i n_j L} = 1$ for any given pair of populations *i*, *j*. Therefore

$H_{Tmax} = \frac{\bar{H}_S}{Np} + \frac{2}{Np^2}\frac{Np(Np-1)}{2} = \frac{\bar{H}_S + Np - 1}{Np}$ and

$$F_{ST} = 1 - \frac{\bar{H}_S}{H_{Tmax}} = 1 - \frac{\bar{H}_S}{\frac{\bar{H}_S + Np - 1}{Np}} = 1 - \frac{Np\bar{H}_S}{\bar{H}_S + Np - 1}$$

by taking the derivative of F_{ST} with respect to H_s it is clear that F_{ST} decreases with H_s (the derivative is negative) so the lower the H_s the higher the F_{ST} thus we should compute the minimum H_s. We know that

$H_S = \frac{\sum_{Np} \theta_\pi}{Np}$ where θ_π is the mean number of differences between pair of sequences of length *L*. The minimum number of differences at one site will occur if an allele frequency at this site is maximum e.g. if the allele is at frequency 1 the differences at this site are 0. In our case the maximum frequency allele is (1-*maf*) that in a sample of size *n* implies *n*(1-*maf*) copies of this allele and *n*(*maf*) copies of the alternative so the number of differences at this site are n^2(1-*maf*)(*maf*) and for *L* sites is Ln^2(1-*maf*)(*maf*). The mean is through *Ln*(*n*-1)/2 pairs so for a given population

$$\theta_{\pi min} = \frac{Ln^2(1-maf)maf}{\frac{Ln(n-1)}{2}} = \frac{2n(1-maf)maf}{(n-1)}$$

then for *Np* populations with different sample sizes

$$H_{Smin} = \frac{\sum_{Np} \theta_{\pi min}}{Np} = \frac{2(1-maf)maf}{Np} \sum_{k=1}^{Np} \frac{n_k}{n_k - 1}$$

and finally the upper bound F_{ST} is

$$F_{STmax} = 1 - \frac{\overline{H}_{Smin}}{H_{Tmax}} = 1 - \frac{Np\overline{H}_{Smin}}{\overline{H}_{Smin} + Np - 1}$$

by substituting H_{smin} and some rearrangement we get

$$F_{STmax} = \frac{(Np-1)\left[Np - 2(1-maf)maf \sum_{k=1}^{Np} \frac{n_k}{n_k - 1}\right]}{Np(Np-1) + 2(1-maf)maf \sum_{k=1}^{Np} \frac{n_k}{n_k - 1}}$$

as an F_{ST} upper bound in a biallelic setting with *maf* frequencies in a number of *Np* populations with different sample sizes.

4) F_{STmin}

Finally, we will show that the lower bound for F_{ST} is $F_{STmin} = 0$. Let $F_{STmin} = 1 - H_{smax}/H_T$.

For simplicity and without loss of generality let assume that the sample size is even in every population. At any site the maximum number of differences occurs when allelic frequencies are at intermediate frequency and this number is $n^2/4$ (or $(n^2-1)/4$ if odd). So, for L sites we have $Ln^2/4$ differences. The mean through $Ln(n-1)/2$ pairs for a given population is

$$\theta_{\pi max} = \frac{\frac{Ln^2}{4}}{\frac{Ln(n-1)}{2}} = \frac{n}{2(n-1)}$$

$$H_{Smax} = \frac{\sum_{Np} \theta_{\pi max}}{Np} = \frac{1}{2Np} \sum_{k=1}^{Np} \frac{n_k}{n_k - 1}$$

which corresponds to H_{smin} as computed above when *maf* is substituted by 0.5.

As we already shown that F_{ST} decreases with H_s we compute the corresponding pooled heterozygosis the H_s is maximum

$$H_T = \frac{\overline{H}_{Smax}}{Np} + \frac{2}{Np^2} \sum_{k=2}^{Np} \sum_{k'=1}^{k-1} \theta_{\pi a}\ (k, k')$$

Because there are only two alleles and the alleles in any population are at intermediate frequencies, the number of differences in a given site between any pair of populations *i,j* is $n_i n_j/2$ and the average value for L sites and pairs of sequences, $\theta_{\pi a}$, is $(Ln_i n_j/2)/\ Ln_i n_j = 1/2$ for the pair of populations *i,j*. So

$$H_T = \frac{\overline{H}_{Smax}}{Np} + \frac{(Np-1)}{Np}\frac{1}{2}$$

$$F_{STmin} = 1 - \frac{\overline{H}_{Smax}}{H_T} = 1 - \frac{\overline{H}_{Smax}}{\frac{\overline{H}_{Smax}}{Np} + \frac{(Np-1)}{Np}\frac{1}{2}}$$

$$F_{STmin} = 1 - \frac{2Np\overline{H}_{Smax}}{2\overline{H}_{Smax} + Np - 1} = \frac{(Np-1)(1-2\overline{H}_{Smax})}{2\overline{H}_{Smax} + Np - 1}$$

Thus $F_{STmin} > 0$ implies that $1 > 2H_{smax}$ which in turn implies

$$Np > \sum_{k=1}^{Np} \frac{n_k}{n_k - 1}$$

because $n_k/(n_k - 1) > 1$ then to be true the above inequality it is a necessary condition that $Np > Np$ so it follows that $F_{STmin} \leq 0$. Because we force F_{ST} to be 0 the lower-bound will be $F_{STmin} = 0$.

A-6) Simulations and analysis

There are several examples of adaptation to divergent environments connected by migration such as the intertidal marine snail *L. saxattilis* (Rolan-Alvarez 2007), wild populations of *S. salar* (Bourret *et al.* 2013), lake whitefish species (Renaut *et al.* 2011) and so on. To perform simulations as realistic as possible, a model resembling the most favorable conditions for the formation of ecotypes under local adaptation with gene flow was implemented. Some

relevant demographic information from *L. saxatilis*, such as migration rates and population size as estimated from field data (Rolan-Alvarez 2007), was used. Concerning selection intensities, we considered moderate selection pressures and few loci with large effects (Thibert-Plante & Gavrilets 2013). Therefore, the simulation design includes a single selective locus model plus one case under a polygenic architecture with 5 selective loci. Two populations of facultative hermaphrodites were simulated under divergent selection and migration. Each individual consisted of a diploid chromosome of length 1Mb. The contribution of each selective locus to the fitness was 1-*hs* with $h = 0.5$ in the heterozygote or $h = 1$ otherwise (Table S1). In the polygenic case the fitness was obtained by multiplying the contribution at each locus. In both populations the most frequent initial allele was the ancestral. The selection coefficient for the ancestral allele was always $s = 0$ while $s = \pm 0.15$ for the derived. That is, in population 1 the favored allele was the derived (negative *s*, i.e. $1 + h|s|$ in the derived) which was at initial frequency of 10^{-3} while in the other population the favored was the ancestral (positive *s*, i.e. $1 - h|s|$ in the derived) and was initially fixed.

Table S1. Fitness Model. The ancestral allele is noted with uppercase *A* and the derived as *a*.

Population	**Genotypes**		
	AA	***Aa***	***aa***
1	1	$1 + \|s\|/2$	$1 + \|s\|$
2	1	$1- \|s\|/2$	$1 - \|s\|$

$|s|$: absolute value of the selection coefficient.

In the single locus model the selective site was located at different relative positions 0, 0.01, 0.1, 0.25 and 0.5. In the polygenic model the positions of the five sites were 4×10^{-6}, 0.2, 0.5, 0.7 and 0.9. Under both architectures, the overall selection pressure corresponded to α = 4*Ns* = 600 with *N* = 1000. Simulations were run in long term scenarios during 5,000 and 10,000 generations and in short-term scenarios during 500 generations. Some extra cases with weaker selection α = 140 (*s* = ± 0.07, *N* = 500) in the long-term (5,000 generations) and stronger selection, α = 6000 (*s* = ± 0.15, *N* = 10,000) in the short-term were also run.

The mating was random within each population. The between population migration was *Nm* = 10 plus some cases with *Nm* = 0 or *Nm* = 50 in a short-term scenario. Recombination ranged from complete linkage between pairs of adjacent SNPs (no recombination, ρ = 0), intermediate values ρ = 4*Nr* = {4, 12, 60} and fully independent SNPs.

A bottleneck-expansion scenario was also studied consisting in a neutral case with equal mutation and recombination rates, $\theta = \rho$ = 60, and a reduction to *N* = 10 in one of the populations in the generation 5,000 with the subsequent expansion following a logistic growth with rate 2 and K_{max} = 1000.

For every selective case, 1000 runs of the corresponding neutral model were simulated. To study the false positive rate (FPR) produced by the selection detection tests, the significant results obtained in the neutral cases were counted. The simulations were performed using the last version of the program GenomePop2 (Carvajal-Rodriguez 2008).

In most scenarios, the number of SNPs in the data ranged between 100 and 500 per Mb. However, only the SNPs shared between populations were considered thus giving numbers between 60-300 SNPs per Mb i.e. medium to high density SNP maps.

The interplay between divergent selection, drift and migration (Yeaman & Otto 2011) under the given simulation setting should permit that the adaptive divergence among demes persists despite the homogeneity effects of migration (see *Critical migration threshold* below).

A-7) Critical migration threshold

Our simulation model is a particular case (with symmetric migration and intermediate dominance) of the model in Yeaman and Otto (2011). These authors develop the model to study the interplay of drift, divergent selection and migration on the maintenance of polymorphism between interconnected populations. They provide a measure, the critical migration threshold, below which adaptive divergence among demes is likely to persist. By rearranging terms in equation (11) from Yeaman and Otto (2011) and after substituting the fitness relationships from our system, we obtain the critical migration threshold for our model:

$$m_{crit} = \frac{1}{2}\frac{\left(\frac{\alpha}{2}\right)^2 - 1}{\left(\frac{\alpha}{2}\right)^2 + 4N} \qquad \text{(A-7-1)}$$

where α = 4*Ns* . For each selective pressure, we can therefore compute the critical number of migrants (Nm_{crit}) below which the selective polymorphism should be present in the data. The weaker the selection the lower the threshold so, for α = 140 the minimum critical number of migrants is 355 individuals. Thus our highest migration *Nm* = 50 is far below the threshold. This means that both scenarios *Nm* = 10 and 50, would tend to maintain the locally adaptive allele for every selective scenario assayed (weak, intermediate and strong) despite the homogeneity effects of migration.