bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HacDivSel: A program implementing a new haplotype-F_{ST} composite method 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}, 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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

In this work two complementary methods for detection of divergent selection between pairs of populations connected by migration are introduced. The new statistics do not require knowledge on the ancestral or derived allelic state and are robust to false positives. Additionally, it is not necessary to perform neutral simulations to obtain critical cut-off values for the identification of candidates. The utility of the methods is demonstrated in simulations. 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 highest F_{ST}s are not a priori selected to identify loci. On the contrary, candidate loci are chosen based upon haplotype 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 indicate that outcome. It is shown that powers ranging from 70-90% 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 involving population bottleneck and expansion. The second method is developed for cases with independently segregating markers, where the information contained in the haplotypes vanishes. In this case, the power to detect selection is attained by developing a new F_{ST} extreme-outlier test based on a *k*-means clustering algorithm. Both kinds of strategies have been implemented in the program HacDivSel.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 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 there is not outgroup so that the allelic state, ancestral or derived, is unknown. There are several methods aiming to detect selection in genome regions undergoing local adaptation. Some of them are based on finding outlier loci when measuring genetic differentiation between populations. From its original formulation (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).

Different problems have been encountered regarding to the use of F_{ST} measures as a guide in the search for selection. 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 (Bonhomme *et al.* 2010). Moreover, several non-local processes, 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 to

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

work-well when there are many loci under selection (Bierne *et al.* 2013; Li *et al.* 2012). All of the above provokes that, still under optimal conditions, some of the methods tend to produce many false positives (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) 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 an outgroup to account for the population structure.

Furthermore, if two populations connected by migration are under divergent selection, the selected alleles will be at more or less intermediate frequencies. Depending on the interplay between selection, migration and drift, a critical migration threshold will exist to maintain polymorphism (Yeaman& Otto 2011; Yeaman& Whitlock 2011). In this case, the detection of selection would be difficult using frequency spectrum methods but see (Achaz 2009), it would be also difficult under most of the F_{ST}-based existing approaches, but see hapFLK (Fariello *et al.* 2013). However, there is no program specialized to detect divergent selection at genomic or sub-genomic level, while working without information on the structure of the population tree, the state, ancestral or derived, of the alleles and on the genotypes.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The goal of the present work is to develop such a program, for detecting ongoing divergent selection in pairs of populations with gene flow. This new method should be adequate for working with non-model species. It would also desirable that selection could be detected without simulating any neutral demographic scenario. Moreover, it should work at the genomic level with automatic decision-making for the window size and must be protected from false positives. The latter is one of the main problems associated to the genome-wide selection detection methods. The algorithm we will describe achieves these goals, at least on the several scenarios assayed; being powerful for detecting single or polygenic selection when two populations connected by migration undergo divergent selection.

The design of the work is as follows:

In the first part, the model is developed, which includes the computation of a normalized variance difference, nvd, for detecting haplotype patterns under selection. The nvd value is then used to perform an F_{ST} index measure. Additionally, the algorithm for a conservative, extreme positive outlier test, to deal with the case of fully unlinked SNPs, is also developed. These methods are implemented in the computer program HacDivSel.

Second, the simulation setting for testing the methods is explained.

Finally, the results concerning power and false positive rate (declared significant versus true nulls) are given and discussed.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The model

Generalized HAC variance difference

For a given sample, let define the major-allele-reference haplotype (MARH) as the haplotype 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. An haplotype allelic class (HAC) groups the haplotypes having the same mutational distance. Therefore, 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 or 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 '1' is the minor allele. 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 (close to the MARH) and consequently

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

provoking lower variance values. 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 (Hussin *et al.* 2010). 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 immediately generalized to

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

Where, *f* is the frequency of the derived allele, 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* and 1-*f* are interchangeable.

A drawback in the gSvd statistic is its dependence on the window size *L* 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 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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to the window size will be dependent on the recombination rate and on the effect of selection. Thus, it is desirable to develop a HAC-based statistic not dependent on the window size. In what follows, the between partition variance difference will be 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 any given 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 given candidate SNP. We have the following relationships for the mean m and sample variance S^2 values (the subscripts identify the partition, see Appendix 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 $\bar{S} = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n-1}$; n_1 and n_2 are the sample sizes ($n_1 > n_2$ 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

$$vd_i = \frac{(n-1)S^2-(n-2)S_1^2}{n_2-1} \times 4f_i(1-f_i) \qquad \text{with } n_2 > 1 \qquad (2)$$

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 (Hussin *et al.* 2010; Rivas *et al.* 2015), the sample variance S^2 will be a function

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 has 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

$$nvd_i = \frac{2(\mathrm{n}-2)[(n-1)S^2-(n-2)S_1^2]}{(n_2-1)\mathrm{nL}^2} \times 4f_i(1-f_i) \qquad (4)$$

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* can be 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 selective sites. Therefore, we compute the normalized variance difference in (4) for the data obtained by merging the shared SNPs from

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the two population samples. Note however that the reference haplotype (MARH) is computed just from one of the populations (by default population 1).

Recall that (4) is already normalized by the square of the window size *L*. However, the problem remains, as with other genome-wide estimators based on sliding-window approaches, on how to choose an optimal window size when computing the quantities of interest. A solution to this problem is to automate the choice of the window size 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 for 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 selection is not at work, the maximum nvd value would be a false positive. Unfortunately, we ignore the distribution of the statistic and cannot decide if a given nvd 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. In 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}.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sign test

From a lower bound of (4) we derive the quantity (see Appendix for details)

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

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. Performing neutral simulations to assess significance in a given data set could be challenging especially with non-model organisms because the demographic history and related evolutionary parameters are not well known. As an alternative, it is possible to combine 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 asses. The joint use of these methods produces the combined measure nvdF_{ST}. The rationale of the approach is that if divergent selection acts on an specific site then the F_{ST} at that site will be higher compared to the overall F_{ST}. To obtain a *p*-value an obvious strategy would be to perform an LK test (Lewontin& Krakauer 1973). However, LK and its derivatives have several problems and a tendency to produce high rates of false positives (De Mita *et al.*

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2013; De Villemereuil *et al.* 2014; Lotterhos& Whitlock 2014). Instead, we proceed as follows, let *i* to be a 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. In doing so, the mean frequency for every SNP in the pooled populations is considered as this is 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. Thus, a low *p*-value indicates that the observed I_i was larger than the resampled ones most of the times. Note that, for each site, the resampling procedure has variance *pqn* which will be larger at intermediate frequencies. For a candidate with a high mean frequency (i.e. both populations has high frequency) the value of I_i is low and the *p*-value will be high. If the mean frequency is low the situation is similar, I_i is low and the *p*-value high. When the mean 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 seeking 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. Note that if $Nm < 1$ it is likely that the shared SNPs are scarce. Therefore, by considering only shared SNPs, we avoid to test data with $Nm < 1$ that could incur in false positives.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Multiple test correction by the effective number of independent SNPs, sign test and significance

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 $\gamma = 0.05$.

Thus, the algorithm nvdF_{ST} ends with a candidate being considered significant only when its *p*-value (computed as explained in the previous section) is lower than *c* and the sign in (5) is positive.

The k-means extreme positive outlier (EPO) test

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 be accessible. 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 above, 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 test takes advantage on the fact that the involved regions under divergent selection may produce a kind of 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. This approach is heuristic, i.e. a problem solving method that finds a satisfactory (but not necessarily optimal) solution to a difficult problem.

The rationale of the algorithm is as follows:

The first step consists in computing the extreme positive outliers (EPO) in the sense of Tukey i.e. those sites having a F_{ST} value higher than 3IQR where IQR is the interquartile range (Tukey 1977). The second step identifies different classes inside the extreme outlier set. This is done by a *k*-means algorithm which is a well-known method widely used in clustering and digital image processing (Schubert *et al.* 2012; Vattani 2011). Here, a *k*-modal distribution is assumed and all the elements of the EPO 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 the adjacent inter-mode middle point are maintained in the set. By

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

default k = 2 and two modes {0, 1} were used. Finally, for each of the candidates remaining in the EPO set a 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 EPO 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 must be in *MS*-format (Hudson 2002) and 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 -name sel.txt -sample 50 -candidates 10 -SL 0.05 -output anyname

Where -*sample* is the sample size for each population and the label -*candidates* 10, indicates that the ten highest nvd values must 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 EPO 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.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 real 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 (Sadedin *et al.* 2009; Thibert-Plante & Gavrilets 2013). Therefore, the simulation design includes both a single selective locus model plus some simulations 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 1). 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 = ± 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.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. 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 $\rho = 4Nr$ = {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.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 implemented 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* section in the Appendix).

Results

Combined method (nvdF_{ST})

Under the single locus architecture, the power of nvdF_{ST} vary between 80-90% for both medium (60 SNPs/Mb) and high density (250 SNPs/Mb) maps. (Table 2). These results can be compared with previously published analysis (Rivas *et al.* 2015) of the same files corresponding to rows 1-6 in Table 2. 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 (Table 2 in Rivas *et al.* 2015).

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

When these methods were applied considering the two populations merged then the best powers were attained by SvdM ranging from 42 to 94% (Table 3 in 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 2, nvdF_{ST} performs quite well with powers from 81 to 93% 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 2) the power is also good (80-99%). 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.

An exception in the power attained by nvdF_{ST} occurs when all SNPs segregate independently (last row in Table 2). 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 test and a candidate with low F_{ST} index measure.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Performance of the combined method (nvdF_{ST}) with *n* true selective sites. 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%)	Localization (kb)
65	12	0	1	87	1.7	±460
63	12	4	1	93	2.7	±200
60	12	12	1	91	1.2	±30
251	60	0	1	81	1.4	±60
232	60	4	1	84	6	±20
249	60	60	1	86	2.6	±1
282	60	60	5	80-99	2.6	<±1
318	60	∞	1	0	0	-

∑: Mean number of shared SNPs per Mb. θ: Mutation rate. ρ: Recombination rate. FPR: false positive rate. ∞: independently segregating sites.

Extreme Positive Outlier (EPO) test

The EPO test is very conservative as can be observed in Table 3 where the false positive rate is 0 in every case. Its power increases with the density and the independence of the markers reaching 60% 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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

polygenic setting (row with n = 5 in Table 3). Therefore, the EPO 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 (EPO) with n true selective sites. Selection was α = 600 and Nm = 10. Mean localization is given in distance kb from the real selective position.

∑	θ	ρ	n	%Power EPO	%FPR (γ = 5%)	Localization (kb)
65	12	0	1	0	0	-
63	12	4	1	0.2	0	±3
60	12	12	1	1.1	0	±77
251	60	0	1	0.9	0	0
232	60	4	1	1.5	0	±137
249	60	60	1	56	0	±0.07
282	60	60	5	0	0	-
318	60	∞	1	60	0	0

∑: Mean number of shared SNPs per Mb. θ: Mutation rate. ρ: Recombination rate. FPR: false positive rate. ∞: independently segregating sites.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Position effect

The ability of nvdF_{ST} to locate the position of the selective site, increases with the marker density and the recombination rate (Table 2). 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). However, the localization is also dependent on where the selective site is placed within the chromosome. Best outcomes, 0-1 kb from the real position, are obtained when the site is at the center of the chromosome. The farther from the center the worse the ability to correctly localize the selective positions (Table 4). This issue have already been shown for other HAC-based methods (Rivas *et al.* 2015). The problem is partially solved using the EPO test, because at least for the cases with recombination enough for this method to work well, the selective sites are localized at distances ranging from 0 to 36 kb from its real position about 64-79% of the times (cases with ρ = 60 in Table 4).

Table 4. Performance of nvdF_{ST} and EPO 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 2 and are omitted.

∑	θ	ρ	%Power nvdF_{ST}, EPO	Position (kb)	Localization (kb) nvdF_{ST}, EPO
259	60	0	83, 1	0	+484, +453
255	60	0	83, 1.5	10	+436, +479

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

256	60	0	83, 0.8	100	+349, +353
255	60	0	80, 0.7	250	±198, ±197
230	60	4	76, 2.4	0	+327, +125
226	60	4	77, 3.2	10	+329, +154
233	60	4	80, 1.7	100	+229, +138
229	60	4	83, 1.7	250	±124, ±22
262	60	60	63, 79	0	+119, +36
261	60	60	67, 78	10	+111, +30
257	60	60	82, 78	100	±45, ±6
252	60	60	86, 64	250	±10, ±0.07

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

Bottleneck-expansion scenarios

The robustness of the methods was additionally tested under a bottleneck-expansion scenario that is known to leave signatures that mimic the effect of positive selection. Thus, no selection was applied to this scenario in order to test its effect on the false positive rate of the methods. As a result, for nvdF_{ST} the false positive rate is maintained below the nominal level (4.8%) and for the EPO test it continues to be 0%. Therefore, both methods seem to be robust to these kinds of confounding scenarios.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Isolation and Nm = 50 scenarios

For the isolation scenario (*Nm* = 0), the tests simply produce no output. This happens because the program works with the shared variation between populations. If the isolation occurred from time ago there are not shared SNPs and the program is not able to produce an output.

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 33-61% with a false positive rate of 0% (Table 5). Therefore, the test is being 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 EPO test it has no power to detect selection in the given scenario when *Nm* equals 50.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 5. Performance of the combined method (nvdF_{ST}) with a single selective site in the short term (500 generations). 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%)	Localization (kb)
116	60	0	57	0	±150
180	60	4	61	0	±132
178	60	60	33	0	±4

∑: Mean number of shared SNPs per Mb. θ: Mutation rate. ρ: Recombination rate. FPR: false positive rate.

Short-term strong and long-term weak selection scenarios

The performance under the strong selection scenario is presented in Table 6. 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 46-66% with 0 false positive rate. These results can be compared with gSvd methods, Svd and SvdM, in Table 6 (*t* = 500 generations) from Rivas *et al*. (Rivas *et al.* 2015). The results in Rivas *et al*. 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 asses significance with these methods the exact neutral demography was simulated by Rivas and coworkers.

Concerning very weak selection in long-term scenarios (Table 6, α = 140) the power varied between 47-52% and localization between 2-32kb from the real selective position.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 6. 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%)	Localization (kb)
112	60	0	6000	500	46	0	±44
32*	60	4	6000	500	62	0	<±1
62	60	60	6000	500	66	0	±77
165	60	0	140	5,000	47	3.4	±32
156	60	4	140	5,000	52	5.6	±11.8
135	60	60	140	5,000	49	2.6	±1.7

∑: Mean number of shared SNPs per Mb. θ: Mutation rate. ρ: Recombination rate. *t*: number of generations. FPR: false positive rate. *: only 40 runs having a minimum of 25 SNPs.

Discussion

The goal in this work was to develop a method 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 model should be

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

useful for non-model species and it should not be necessary to perform neutral simulations to obtain critical cut-off values for the candidates.

It has been shown that combining an haplotype-based method with the F_{ST} differentiation measure, the so-called nvdF_{ST}, is quite powerful strategy for detecting divergent selection. However, when the whole set of markers is segregating independently, the information in the haplotype vanishes. Therefore, nvdF_{ST} is complemented with a new F_{ST} outlier test called EPO that was developed as a very conservative test because the mentioned tendency of such methods to produce false positives. Its conservativeness is attained by assuming that among the detected outliers, the ones which are due to divergent selection would be clustered apart from neutral ones. The implemented extreme positive outlier test behaves acceptably well when markers are independent, reaching powers between 60-80% while maintaining a false positive rate of virtually zero.

All the developed methods work without any a priori knowledge about candidate positions neither any information about the state, ancestral or derived, of the alleles. Consequently, they seem an interesting option for exploratory studies in non-model species.

In general, the F_{ST}-based methods are affected by the presence of 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 overall mean is high. The nvdF_{ST} performs even better in this scenario because the distributed selective signal facilitates the discovery of the corresponding patterns. Since only the F_{ST} of the specific site is compared with the overall F_{ST} and the null distribution is obtained using inter-

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

population mean frequencies, the nvdF_{ST} maintains high power under the polygenic setting. Nevertheless, the localization of various selective sites is not easy at least under the sample size assayed.

Besides the detection of the signal of selection in the sequences, the accuracy in inferring the location of the selected allele was also studied. When the target site is located at the centre of the studied region and the overall recombination rate is at least 0.3 cM/Mb ($\rho \geq 12$), the nvdF_{ST} method performs quite well under weak selection, with the inferred location within 32 kb of distance from the true location in the worst case. Under strong selection, the localization is worst, 77 kb, but this could be due to the lower number of segregating sites (only 62 in Table 6). The performance is poorer when the target is not at the center. In this case, with recombination of 1.5 cM/Mb, the inferred location changes from an almost perfect localization (<1 kb) to distances of 10-120 kb as the target is shifted away. Fortunately, the complementary EPO test works well, at least when the overall recombination rate is above 1.5 cM/Mb, giving detection powers near 80% and seems not to be so affected for such positional effect, resulting in distances from the true selective position between <1 - 36 kb.

Clearly, inferring the selective location 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 EPO 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.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As a conclusion, nvdF_{ST} combines haplotype and population differentiation information for the detection of divergent selection and seems to work well when knowledge of the haplotype phase is at hand. The complementary EPO method is a conservative alternative useful when the full set of SNPs is unlinked. Both strategies can be applied without the need of performing neutral simulations and have false positive rates below the desired nominal level.

Bibliography

Achaz G (2009) Frequency spectrum neutrality tests: one for all and all for one. *Genetics* **183**, 249-258.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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

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.

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.

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.

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Sadedin S, Hollander J, Panova M, Johannesson K, Gavrilets S (2009) Case studies and mathematical models of ecological speciation. 3: Ecotype formation in a Swedish snail. *Mol Ecol* **18**, 4006-4023.

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.

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.

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.

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). Simulated data: DRYAD entry doi: doi:10.5061/dryad.18f8h.

Acknowledgements

I thank E. Rolán-Alvarez and three anonymous reviewers 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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Crecimiento, GPC2013-011) and fondos FEDER. The author declares to have no conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Appendix

Effect of window size

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).

General variance difference

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

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

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Note that

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

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

Now consider the variance

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

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

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

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

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

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

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

From (A-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

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

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

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

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

Realize that the first term in the sum is contributing to increase the variance difference whenever $(n-1)S^2 \geq (n-2)S_1^2$. 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 $kn = n_2$ decrease so both are contributing to increase the negative term under

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

selection and diminish the value of the statistic. Consequently, it is adequate to discard the second term just to obtain

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

Lowe 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

$$\frac{4(n-1)S^2-nL^2}{nL^2} \quad \text{(A-8)}$$

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-8) 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-8) 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-8) is equal or lower than

$$\frac{4(n-1)S^2-2\ \sum_i \mathrm{hac}_{1i}^2}{nL^2} \quad \text{(A-9)}$$

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-9) could be

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted September 8, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

higher or lower than (A-8) depending on the HAC values of the first partition. In any case, a negative value in (A-9) 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 require (A-9) to be positive to count a given candidate as significant.

Critical migration threshold

Our simulation model is a particular case (with symmetric migration and intermediate dominance) of the model in Yeaman and Otto (2011) that these authors develop to study the interplay between drift, divergent selection and migration on the maintenance of polymorphism between the interconnected populations. Thus, we can compute 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) after substituting the fitness relationships from our system, we obtain the critical migration threshold:

$$m_{crit} = \frac{1}{2}\frac{\left(\frac{\alpha}{2}\right)^2 - 1}{\left(\frac{\alpha}{2}\right)^2 + 4N} \qquad \text{(A-10)}$$

where $\alpha = 4Ns$. 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 $\alpha = 140$ 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.