bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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: Two new methods (haplotype-based and outlier-based) for the detection of divergent selection in pairs of populations.

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, 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 July 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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

The detection of genomic regions involved in local adaptation is an important topic in current population genetics. There are several detection strategies available depending on the kind of genetic and demographic information at hand. A common drawback is the high risk of false positives. In this study, we introduce two complementary methods for the detection of divergent selection from populations connected by migration. Both methods have been developed with the aim of being robust to false positives. The first method combines haplotype information with inter-population differentiation (F_{ST}). Evidence of divergent selection is concluded only when both the haplotype pattern and the F_{ST} value support it. The second method is developed for independently segregating markers i.e. there is no haplotype information at hand. In this case, the power to detect selection is attained by developing a new outlier test based on detecting a bimodal distribution. The test computes the F_{ST} outliers and then assumes that those of interest would have a different mode which is detected by a clustering algorithm. The utility of the two methods is demonstrated through simulations and the analysis of real data. The simulation results show power ranging from 60-94% in several of the scenarios whilst the false positive rate is controlled below the nominal level in every scenario. The analysis of real samples consisted of phased data from the HapMap project and unphased data from intertidal marine snail ecotypes. The software HacDivSel implements the methods explained in this manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 an important focus on the detection of the signature of natural selection at the molecular level. The detection of the selection effect in a given DNA region matters because it may connect that region with a key functionality, with past or ongoing selective events and in general, create a deeper understanding of the evolutionary processes. More specifically, it may help to understand the evolutionary mechanisms allowing the species adaptation to local conditions. The study of local adaptation processes implies that some genetic variant is favored by the environmental conditions. The positively selected locus increases in frequency and the pattern of variation around that locus will change through the process known as selective sweep (Smith and Haigh 1974; Nielsen et al. 2005).

There are many tests designed to detect different kinds of effects produced by selective sweeps. Such effects can involve skewed site frequency spectra, high linkage disequilibrium or high rates of genetic divergence (Crisci et al. 2013; Jensen, Foll, and Bernatchez 2016). The information required by those tests is variable: they could need knowledge about candidate adaptive loci, the haplotypic phase, the recombination rates, or the ancestral/derived status at each segregating site. This kind of information is often available for model organisms and so, in previous years, most of the effort was focused in humans and other model species.

In the case of non-model organisms, the most useful methods for studying local adaptation have been those based on measuring genetic differentiation between populations. The idea behind these methods is that the loci involved in local adaptation would be outliers, i.e. would have unusually large values of F_{ST}. From its original formulation (LK test Lewontin and

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Krakauer 1973), this technique has been improved in several ways depending on the summary statistic's -the F_{ST} or any other differentiation index - expected neutral distribution. That is, in order to account for more realistic situations, the different methods change the assumptions of the null demographic model (reviewed in Whitlock and Lotterhos 2015).

However, one of the main drawbacks of the outlier-based methods is the difficulty of defining an accurate null model since there are several historical events and demographic scenarios, other than local selection, that can produce similar F_{ST} patterns. Still in the presence of local adaptation we can expect different F_{ST} patterns, since the populations involved may be more or less connected by migration, and this will influence the structure of the genetic variation both at intra and inter-population levels. Consequently, when using outlier methods based on the deviation over an expected demographic null model, we always face the risk of having an excess of false positives (Perez-Figueroa et al. 2010; Bierne, Roze, and Welch 2013; De Mita et al. 2013; Lotterhos and Whitlock 2014).

Another problem concerning outlier-based methods is that they have low power when the overall F_{ST} is high as it can happen when the genetic basis of the adaptation is polygenic or when the populations under study are subspecies (De Villemereuil et al. 2014; Whitlock and Lotterhos 2015). Fortunately, over recent years, the amount of information available on the genomes of several species has increased (Ellegren 2014) and consequently new and more sophisticated methods can now be applied to detect local adaptation in non-model organisms. As mentioned above, the linkage disequilibrium (LD) is the basis for several computational methods used in the detection of selective sweeps (reviewed in Crisci et al. 2012). Some LD-based methods try to identify the region with maximized LD (Kim and

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Nielsen 2004; Pavlidis, Jensen, and Stephan 2010) while others explore the pattern of LD decay from candidate SNPs (Voight et al. 2006; Sabeti et al. 2007). However, only a few LD-based methods have considered structured populations as the evolutionary scenario of interest. Still, there are rather different scenarios that can be evaluated under structured populations (Vatsiou, Bazin, and Gaggiotti 2016). Therefore, although LD-based tests can be powerful, and robust for detecting selective sweeps in isolated or simple structured population scenarios with low migration rate, they are not, under several other realistic scenarios (Crisci et al. 2012; Crisci et al. 2013; Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015; Vatsiou, Bazin, and Gaggiotti 2016).

In addition, the possibility of observing local adaptation with gene flow depends on the demography and on the genetic basis of the traits involved (Yeaman and Otto 2011). This complicates the performance of the methods under moderate-to-high migration scenarios (Vatsiou, Bazin, and Gaggiotti 2016). Thus, even if haplotype phase information is at hand, specific methods should be developed to detect local adaptation under structured population scenarios (Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015).

The aim of this paper is to present two complementary methods specialized in detecting divergent selection in pairs of populations with gene flow. Both methods are adequate for working with non-model species although the first requires an approximate knowledge of the phase of the SNPs under study. If the SNPs under evaluation are not linked then the second method should be applied. Our working definition for non-model species involves those for which we could have partial information about the haplotypic phase but we do not have estimates of recombination rates, neither do we have information on potentially

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

adaptive loci; nor do we know the ancestral/derived status at each segregating site. This definition also implies that we barely have good knowledge about the demography of the populations under study and so, we cannot reliably use a simulated neutral distribution to assess significance.

The first method combines haplotype-based information with a diversity-based F_{ST} measure. It is a sliding-window approach that uses automatic decision-making for applying different window sizes. The second method is not haplotype-based and performs a two step F_{ST} outlier test. The first step of the algorithm consists of a heuristic search for different outlier clusters, the second step is just a conditional LK test that will be performed only if more than one cluster is found and in this case the test is applied through the cluster with the higher F_{ST} values.

The design of the work is as follows: We begin developing the model for the haplotype-based method ($nvdF_{ST}$) and then, to deal with the case of fully unlinked SNPs, we build up the algorithm for a conservative, extreme outlier set test (EOS). In the results section we first compare the ROC (receiving operating characteristic) space, the two new methods with an LD-based method (Omegaplus, Alachiotis, Stamatakis, and Pavlidis 2012) and with an outlier-based method (BayeScan, Foll and Gaggiotti 2008). Finally, the $nvdF_{ST}$ is applied to genome-wide phased data from the HapMap project (Consortium 2010) and the EOS test to a recently published data set of *Littorina saxatilis* species. We will show that our methods are able to detect divergent selection under gene flow while being robust to false positives.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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} model

In this section we improve a previous haplotype-based method for detecting divergent selection (Hussin et al. 2010; Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015). The new statistic is called *nvd*F_{ST} because it combines a normalized variance difference, *nvd*, with the F_{ST} index. The *nvd* part performs a sliding-window approach to identify sites with specific selective patterns. When combined with the F_{ST}, it allows the significance of the candidate sites to be assessed without the need of simulating neutral demography scenarios. Before developing the *nvd* formula, we reviewed some concepts related to haplotype allelic classes.

Generalized HAC variance difference

A major-allele-reference haplotype (MARH) is a haplotype that carries only major frequency alleles (Hussin et al. 2010). Therefore, we can define the mutational distance between any haplotype and MARH as the number of sites (SNPs) carrying a non-major (i.e. minor) 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 value of the mutational distance corresponding to each haplotype allelic class. That is, 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 haplotypes (and their HAC classes) into P_1, the subset carrying the most frequent (major) allele at the SNP i and P_2 the subset with the remaining haplotypes carrying the minor allele at i. That is, let '0' to be the

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

major allele for the SNP *i* and '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. Within each subset we can compute the variance of the HACs. Thus, in P_1 we have the variance S^2_{1i} and correspondingly variance S^2_{2i} in P_2 where *i* refer to the SNP for which we have performed the partition.

The rationale of the HAC-based methods relies on the sweeping effect of the selectively favored alleles. Therefore, if the SNP *i* is under ongoing selection then the variance in the partition 1 (S^2_{1i}) will tend to be zero because the allele at higher frequency (i.e. the allele of the SNP *i* in the partition 1) is being favored 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 (S^2_{2i}) should not be affected by the sweeping effect because it does not carry the favored allele. So, the difference S^2_{2i} - S^2_{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

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

Where f_i is the frequency of the derived allele of the SNP *i*, and the parameters *b* and *a* 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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

SvdM (Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015). Note that when taking $a = 1$ it is not necessary to distinguish between ancestral and derived alleles because f_i and $1\text{-}f_i$ 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, Dominguez-Garcia, and Carvajal-Rodriguez 2015). Although gSvd is normalized by L, the effect of the window size on the computation of variances is quadratic (see Supplementary 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 bias due to the change in the window size is important because the partitions P_1 and P_2 may experience different scaling effects, which would increase the noise in the estimation. The change in the scale due to the window size will also be dependent on the recombination and selection rates. Thus, it is desirable to develop a HAC-based statistic that does not increase with 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.

Normalized variance difference (nvd)

We have seen that for any haplotype sample of size n, we can compute the statistic gSvd which basically is a difference between the HAC variances of the partitions P_1 and P_2. Remarkably, the corresponding HAC means and variances at each partition are related via the general mean and variance in that sample. Consider, for any candidate SNP, the mean

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

HAC distance, *m*, of the sample, and m_1 and m_2, the means corresponding to the partitions P_1 and P_2, respectively. We have the following relationships for the mean *m* and sample variance S^2 (the subscripts 1 or 2 identify each partition; see the 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 at each partition ($n_1 \geq n_2$ by definition) and $\Delta = \frac{n_1 n_2}{n^2}(m_1 - m_2)^2$.

Using the relationships in (1), it is possible to compute the variance difference as appears in gSvd. However, we can substitute the parameters *b* and *a* by *a* = 1 and *b* = 4 as these are the values that permit to ignore the allelic state while maximizing the frequency product. In addition, it is also possible to consider the difference between means term (delta) in order to engage it in the detection of selection (see details in the Appendix). Thus, we finally obtain a new statistic for the variance difference of the candidate SNP *i*

$$vd_i = vd_{i0} + 4f_i(1-f_i)(S_{2i}^2 - S_{1i}^2) \qquad (2)$$

where $vd_{i0} = \frac{n\Delta}{n_2 - 1} \times 4f_i(1-f_i)$

Therefore, the effect of selection upon vd_i is two-fold. The first term of the sum in (2) corresponds to the effect of the difference between means and the second between variances. Clearly, increasing S_{2i} or decreasing S_{1i}, as expected under selection, will increase the value of the statistic. If S_{1i} and S_{2i} are equal then the value of vd_i is independent of the variances and just relies on the term vd_{i0} corresponding to the partitions' mean (m_1 and m_2)

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

and the candidate SNP frequencies. Please note that vd_{i0} is also the value of vd_i when both variances are 0.

It is worth mentioning that, because the HAC values are bounded by 0 and L, the two parts of vd_i are not independent. So that, if we have an extreme value for the HAC mean in the selective partition, for example $m_1 = 0$, this implies that $S_1^2 = 0$ since every haplotype has to have a HAC of 0 to get that mean value. Note however, that the opposite is not true: a value of $S_1^2 = 0$ does not imply necessarily that $m_1 = 0$.

Having said that, it can be shown (see Appendix A-2) that when the variance of the sample (without the partitions) is maximized, we get an upper bound (d_{max}) for vd_i

$$d_{max} = \frac{nL^2}{2(n-2)} \qquad \text{so that,}$$

$$vd_i \leq d_{max} \qquad (3)$$

If we divide (2) by d_{max} we have a normalized variance difference

$$nvd_i = \frac{vd_{i0} + 4f_i(1-f_i)(S_2^2 - S_1^2)}{d_{max}} \qquad (4)$$

The quantity from (4) can be computed for each SNP in a sample of sequences of any given length L and the SNP yielding the maximum *nvd* may be considered as a candidate for selection. Furthermore, in a two population setting, 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 merge the shared SNPs from the two population samples and then we

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

compute the normalized variance difference using (4). Note however, that the reference haplotype (MARH) is defined just from one of the populations (by default the population with the highest sample size).

Because (4) is normalized by d_{max} the problem of scaling-up on the window size has been solved (see Figure S1 in Appendix A-2). However, the problem of choosing an optimal window size remains. A possible solution is to automate the choice by selecting the size which gives the maximum value for the statistic (Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015) or alternatively, trying different window sizes and giving the corresponding results for each window at the output. Therefore, in a given window size, the program evaluates all the SNPs and selects the ones that have the maximum *nvd* (one or more depending on the proportion of candidates we are interested in evaluating) and then, repeat the process for a different window size.

At this point we already have a HAC-based statistic, *nvd*, that does not increase on the window size and that should produce higher positive values for pairs of populations undergoing divergent selection. However, even if there is no selection, the maximum *nvd* value could also be positive. Unfortunately, we ignore the distribution of the statistic and cannot decide if a given maximum 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 whether the value obtained for a given sample is due to the effect of selection, particularly because we desire to put great emphasis on avoiding false positives.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Consequently, we will perform two more measures before giving a diagnosis about the presence of divergent selection. The first is a sign test based on the lower bound of *nvd*, the second is the comparison between the F_{ST} of the SNP having the maximum *nvd* and the overall F_{ST}.

Sign test

In our definition of *nvd* the term vd_{i0} cannot be negative. For that reason, it could happen that with a negative difference in variances and a high mean HAC value in the first partition, as expected under neutrality, the value of *nvd* could still be positive. Therefore, we use a lower bound of *nvd* to derive the quantity called test selection sign (*tss*, see Appendix A-3 for details) that would have negative value when the HAC values in the first partition are high

$$tss = \frac{4(n-1)S^2 - 2\sum_h \mathrm{hac}_{1h}^2}{nL^2} \qquad (5)$$

Where hac_{1h} are the HAC values measured at each haplotype *h* 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.

Now, even if we have a candidate position identified by its high *nvd* value and by the positive sign of *tss*, we still lack a method for obtaining *p*-values associated to the sites chosen by the *nvd* algorithm. We can solve this problem if we combine the information on the selective candidate SNP, as given by *nvd*, with the F_{ST} interpopulation differentiation index at that site. The joint use of these methods produces the combined measure $nvdF_{ST}$.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Combined method: *nvdF*$_{ST}$

First, it is important to note that, when computing *nvd*, we have considered only the SNPs shared between both populations in order to avoid low informative loci with high sampling variance (reviewed in Whitlock and Lotterhos 2015). Thus, we have an *nvd* value that may indicate the presence of divergent selection in a pair of populations connected by migration. The rationale of the *nvdF*$_{ST}$ approach is that if divergent selection acts on a specific site then the F_{ST} at that site would be higher when compared to the overall F_{ST}. Then, we proceed as follows, let *i* be the candidate site chosen because it has the maximum *nvd* value, then we calculate the index $I_i = F_{STi} - F_{ST}$ comparing the F_{ST} measured at the candidate site with the overall. The F_{ST} values were computed following the algorithm in Ferretti *et al* (Ferretti, Ramos-Onsins, and Pérez-Enciso 2013). To obtain the *p*-value we do not perform an LK test (Lewontin and Krakauer 1973) because first, the candidate was not chosen for being an outlier and second, we are considering linked rather than independent sites.

To get the *p*-value for a given index I_i, the data is resampled several times (500 by default; 100 if we evaluate more than 10^4 SNPs) to generate an empirical distribution. The expected frequency of each SNP is obtained as its mean frequency between populations as this is the expectation under the homogenizing effect of migration (Crow and Kimura 1970). If the sample sizes are different the mean is weighted by the sample size. Then, for each resampling 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 ($I'_i = F_{STi\_resample\ iteration} - F_{ST\_resample\ iteration}$) were larger than I_i.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

If we inspect the *p* values for all the SNPs and ignore the sign test, the above procedure is just testing, for each SNP, the hypothesis of panmixia. However, we are not doing so, i.e. we are not choosing the significant *p*-values and so we are not targeting SNPs that reject panmixia. On the contrary, we are selecting some SNPs based on their high *nvd* value and positive sign test. Only then, can we check if such SNPs reject panmixia, since it is well-known that divergent selection rejects that hypothesis for some genes (Hey 2006; Sousa et al. 2013). Regarding the number of candidates ranked by high *nvd* value, we can decide to consider just the best one, a few or a given percentage (say 0.1%) of candidates.

For candidates with similar frequencies at both populations we expect low index I_i and correspondingly high *p*-values. When the pooled frequency is intermediate; two situations are possible: first, each population has a similar intermediate frequency which again implies 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. Note that, for each site, the resampling procedure has variance $p(1-p)n$ which is larger at intermediate pooled frequency values. Thus, the method has the desired property to be more conservative at intermediate pooled frequencies, which minimizes the possibility of false positives.

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

The computation of $nvdF_{ST}$ has required as many tests as the candidates considered under the different window sizes assayed. Thus, it is desirable to apply a multiple test correction for the number of independent candidates. We use as a proxy for that number, the independent SNPs included between the left and right-most candidate positions.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 roughly estimate the number of independent SNPs, we calculate the linkage disequilibrium measure *D*' (Lewontin 1988; Devlin and Risch 1995) 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 $[w_{ini}, w_{end})$. The Šidák correction (Sidak 1967; Cheverud 2001) can now be applied to get the adjusted significance level $c = 1 - (1 - \gamma)^{1/Meffs}$ with nominal γ (= 0.05 by default). Thus, the algorithm $nvdF_{ST}$ would finally suggest a candidate as significant only when the sign as computed in (5) was positive and the *p*-value (as obtained in the previous section) is lower than *c*.

The *q*-values (Storey 2003) can be seen as a multiple testing analogs of *p*-values. They have been proposed as a useful approach for evaluating method performance in terms of false discoveries (De Villemereuil et al. 2014). Accordingly, we estimate the *q*-values (see Appendix A-4 for details on the calculation) and provide those corresponding to each inspected *p*-value. In order to obtain the *q*-values we must compute the *p*-values for all the SNPs (shared SNPs at frequency higher than *maf*). Certainly, we do not inspect all the computed *p*-values just those corresponding to a desired number of candidates, ranked by their highest *nvd* values. For example, consider that the highest *p*-value is 0.99 and the lowest is 10^{-6}. If we select just one candidate we do so because it has the highest *nvd* value and only a posteriori we concern about its *p*-value. If it happens that the associated *p*-value is 0.2 we give both the *p*-value and its corresponding *q*-value and output this as a non-significant result.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 extreme outlier set test (*EOS*)

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

As already 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. 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. The subsequent LK test is performed only when this kind of outliers is detected. As F_{ST} estimator we use G_{ST} (Nei 1973).

The rationale of the algorithm is as follows: the first step consists of 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. This is done by a *k*-means algorithm (Vattani 2011; Schubert, Zimek, and Kriegel 2012). The algorithm permits to classify all the elements in the outlier set in one of the *k* classes. Once all the elements are classified, 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. For the sake of computational efficiency we use $k = 2$ and consider two modes $\{0, F_{STu}\}$ for computing the cutoff. The modes corresponding to lower (0) and upper (F_{STu}) bounds for the F_{ST} estimator (see Appendix A-5). The cutoff point is defined as

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 overall F_{ST} + F_{STu} / 3, i.e. the mean plus the square root of the upper-bound for the F_{ST} variance under an asymmetric unimodal distribution (Dharmadhikari and Joag-Dev 1989). Finally, for each of the candidates remaining in the EOS after the cutoff, the LK test (Lewontin and Krakauer 1973) is performed to compute its *p*-value. The Šidák correction (Sidak 1967; Cheverud 2001) for the number of remaining outliers in the set is applied to get the significance level. Each *p*-value is accompanied by its corresponding *q*-value.

Software description

Both *nvd*F_{ST} and the EOS test have been implemented in the program HacDivSel. Complete details of the software and extensive explanations for its use can be found in the accompanying manual. The input program files for the haplotype-based test can be in the SNPs x haplotypes HapMap3, MS (Hudson 2002) or Fasta formats. If the data does not include haplotype information then the Plink (Purcell et al. 2007) flat file (map/ped), Genepop (Rousset 2008) or BayeScan (Foll and Gaggiotti 2008) formats can be used. When using Fasta the sample size should be the same for both populations. A typical command line for calling the program in order to analyze a MS format file named *sel.txt* containing 100 sequences, 50 from each population, would be

HacDivSel -input sel.txt -sample 50 -sample2 50 -candidates 10 -SL 0.05 -output anyname -format ms –maf 4

Where 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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

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. The command line options for the different analysis performed, both for simulated and real data, are available at the supplementary material.

Simulations

There are several examples of adaptation to divergent environments connected by migration such as the intertidal marine snail *L. saxattilis* (Rolan-Alvarez 2007), some 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, we use some relevant demographic information from *L. saxatilis*, such as migration rates and population size as estimated from field data (Rolan-Alvarez 2007). Concerning selection intensities, we considered moderate selection pressures and few loci with large effects (Thibert-Plante and Gavrilets 2013). Therefore, a model resembling the most favorable conditions for the formation of ecotypes under local adaptation with migration was implemented.

Two populations of 1000 facultative hermaphrodites were simulated. The selective scenario (α = 4*Ns*) is divergent so that the allele favored in one population is the deleterious in the other. 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. The selection coefficient for the ancestral allele was always *s* = 0 while *s* = ± 0.15

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 the derived. That is, the ancestral was the favored allele in one population (positive *s* in the derived) while the derived was the favored in the other population (negative *s*, see Table S1 in Appendix A-6). In the polygenic case the fitness was obtained by multiplying the contribution at each locus. This simulation model involves low and high mutation rate ($\theta=4N\mu$) and different recombination rates ($\rho = 4Nr \in \{0, 4, 12, 60, \text{unlinked}\}$) and extends previous work (Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015) by adding new parameter values and demographic scenarios. At the end of each run, 50 haplotypes were sampled from each population. The whole setting is fully explained in the Supplementary Appendix (Appendix A-6). The simulations were performed using the last version of the program GenomePop2 (Carvajal-Rodriguez 2008).

The simulated data is also utilized for the comparison of $nvdF_{ST}$ with OmegaPlus (Alachiotis, Stamatakis, and Pavlidis 2012) and of the EOS test with BayeScan 2.1 (Foll and Gaggiotti 2008). We chose OmegaPlus because of its ease of use, its efficiency and good performance, compared with other haplotype-based methods, under non-equilibrium demographic conditions (Crisci et al. 2013). Certainly, OmegaPlus was not developed to detect local selection but selective sweeps in single populations. We will apply it to each single population separately and just to show that, even using haplotype information, it is necessary to apply methods developed specifically to detect divergent selection in presence of gene flow. Several combinations of the parameters for OmegaPlus were tested in (Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015), here we selected the ones with the best performance. BayeScan was chosen because it is one of the main state-of-the-art outlier-based programs. The parameters for BayeScan were the default ones as this is a conservative setting and we were interested in comparing the false positive rates. Only SNPs

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

shared between populations and with a minimum allele frequency (maf) of 2 per population (4%) were considered.

Results

In what follows we present the results for both simulated and real data. The power of a test (true positive rate) is measured as the % of runs in which selection was detected from simulated selective scenarios and the false positive rate (FPR) is measured as the % of runs in which selection (at any position) was detected from simulated neutral scenarios. The *q*-value (Storey 2003) is an estimate from the results (see Appendix A-4).

Within each run, we have also measured the per genome falsely detected number of sites, as well as the FDR (the proportion of falsely detected sites from the total detected). As these quantities were low and do not have any significant impact on power and FPR as defined above we will postpone the mention of these measures until the discussion section.

ROC curves

The ROC curve is a useful way to explore the performance of any detection statistic because it plots, under different conditions, the true positive rate (TPR or power, y-axis) against the false positive rate (FPR, x-axis). Obviously, the best detection method would have high power and low FPR, which corresponds to the points in the upper-left corner of the ROC space. On the contrary, values close to the diagonal or no-discrimination line (NDL) would be considered as poor performing methods. Recall that we have developed two methods

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

specially intended to be robust to false positives so we expect them to occupy at least the left side of the ROC.

In the ROC space of the Figure 1 we can appreciate the performance of *nvdF*$_{ST}$, EOS, OmegaPlus and BayeScan methods through the α = 600 and *Nm*=10, simulated scenarios. These scenarios include two mutation and three different recombination rates (see also Table 1 and Appendix A-6) plus the case with independent markers. EOS and BayeScan were compared only under the independent and the highest recombination scenarios (only two points; see below) while the haplotype-based methods were compared for the 6 mutation-recombination scenarios. Detailed results for each method and simulation setting are given in the next sections but the main general picture can already be obtained from this ROC curve.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

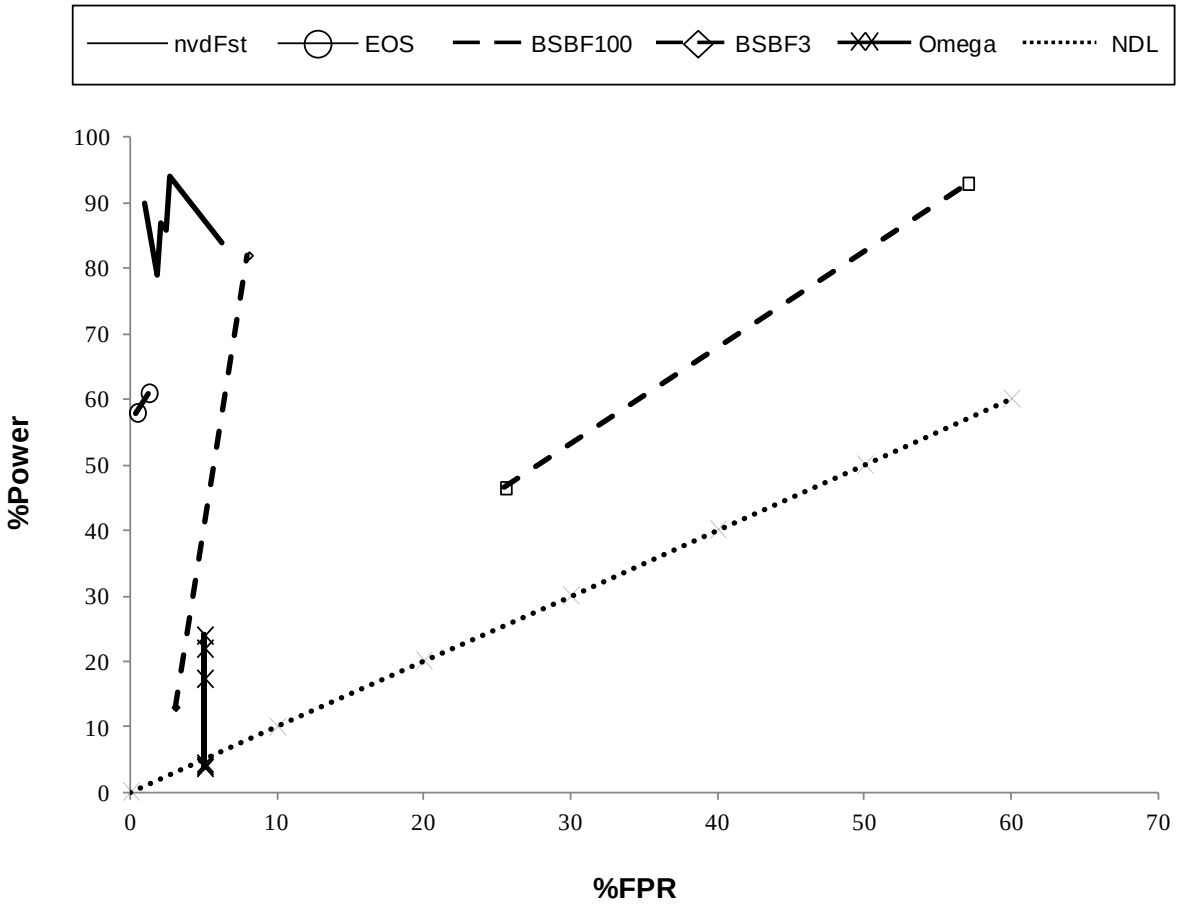

Figure 1. ROC curves for *nvdF*$_{ST}$, EOS, OmegaPlus (Omega) and BayeScan (Bayes factor 3: BSBF3 and Bayes factor 100: BSBF100). NDL: No-discrimination line. FPR: False Positive Rate. There are 6 points (2 mutation x 3 recombination) in the haplotype-based method's curves (*nvdF*$_{ST}$ and OmegaPlus) and 2 points (unlinked and high recombination) in the outlier-based method's curves (EOS and BayeScan).

The method *nvdF*$_{ST}$ is the best performing one since it occupies the upper-left corner. However, the scenario with independent (unlinked) markers is not visible in the *nvdF*$_{ST}$ plot since it has a (0, 0) coordinate (see Table 1 and next section below). For the *nvdF*$_{ST}$ plotted scenarios, the power ranges between 80-90% (y-axis) and FPR (x-axis) is below 5%. On the contrary, OmegaPlus is too close to the NDL diagonal under the same divergent selection scenarios.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 mentioned, the BayeScan and EOS methods are plotted only for the unlinked and the weak-linkage (ρ = 60; 1.5 cM/Mb) marker scenarios. The method BayeScan with a Bayes factor above 3 (BSBF3) is positioned too far to the right side which indicates false positive rates that are too high for both the unlinked (upper point in BSBF3 line) and the weak-linkage (lower point in BSBF3 line) scenarios. However, with a Bayes factor of 100, BayeScan works very well for the unlinked scenario (upper point in BSBF100 line) but fails when markers are even slightly linked (the lower point in BSBF100 line). The EOS test performance for the slightly linked and unlinked scenarios is not bad in terms of power, and very good in terms of FPR since it occupies the left side of the plot.

In the following sections we detail these and other results for $nvdF_{ST}$ and EOS under the different simulated scenarios.

Combined Method ($nvdF_{ST}$)

Under a single locus architecture with selection α = 4*Ns* = 600 and migration *Nm*=10, the power of $nvdF_{ST}$ vary between 79-94% for both medium (60 SNPs/Mb) and high density (250 SNPs/Mb) maps (Table 1 and Figure 1). These results can be compared with published analysis (Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015) with the methods Svd, SvdM and OmegaPlus (Alachiotis, Stamatakis, and Pavlidis 2012) for which similar best results were obtained by Svd and SvdM for the same cases with high mutation and recombination (Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015). However, 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, and consequently, the results in the Rivas

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

and coworkers study, were obtained having the exact neutral demography at hand. As it can be appreciated from rows 1 to 6 in Table 1 —that matches the scenarios in (Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015)—the $nvdF_{ST}$ performs well without the need of performing additional neutral simulations. Also, the false positive rate and the *q*-value are low in all the scenarios. The given results are for 10,000 generations; the cases with 5,000 generations were similar.

Under the polygenic architecture (*n* = 5 in Table 1) at least one candidate is found 99% of the times and more than one, 80% of the time. However, the number of correctly identified sites is 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 in the sign test and a candidate with low F_{ST} index measure.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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. 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 $\alpha = 4Ns = 600$ and migration $Nm = 10$. 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 Selection and Long-term Weak Selection Scenarios

The performance of $nvdF_{ST}$ under the strong selection scenario (α = 6000) in the short-term (500 generations) varies between 44% for fully linked to 67% for weak linked markers (Table 2). Not surprisingly, the number of segregating sites is considerably reduced. In fact the

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

minimum window size allowed by the program had to be shortened from 51 to 25 to perform the analyses. Notably the false positive rate (FPR) was 0.

Concerning weak selection in long-term scenarios (Table 2, α = 140) the power varies 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.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Extreme Outlier Set Test (EOS)

We applied the EOS test under the single locus architecture with selection α = 600 and migration *Nm*=10. As desired, the test is very conservative with false positive rates below the nominal 0.05 in every case (Table 3). Not surprisingly for an outlier-based method, the test has no power if the markers are strongly linked (ρ from 0 to 12) or under a polygenic setting (row with *n* = 5 in Table 3). However, in the cases of independent SNPs and also with recombination of 1.5 cM/Mb in maps with 250-300 SNPs/Mb the power rises up to 60%. Therefore, the EOS test is complementary to $nvdF_{ST}$ having its maximum power when the latter has its minimum.

Concerning false positives, note —in the last three rows of Table 3— the low false positive rate (FPR) values that are indicating the low percentage of outliers detected as selective in the corresponding neutral scenario (1 false outlier in 1.2% of the runs for independent markers and 1-4 false outliers in 0.4% of the runs for linked markers). Thus, the test worked correctly by avoiding false selective sites under the neutral setting. However, the *q*-value estimates varied a lot depending on the linkage between markers. Recall that by *q*-value, we mean 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 the *q*-value is very low (3×10^{-6}) for independently segregating sites but rise up to 0.5 for the same scenario when markers are linked. This happens because when computing the *q*-values, the algorithm assumes that a number of independent *S* tests were performed and corrects under this assumption (see Appendix A-4). However, in the case of linked markers the number of independent tests is smaller and this generates the inflated *q*-values.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 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 in the significant tests. ∞: independently segregating sites.

Position Effect

The ability to locate the physical 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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

since they were low (in the order of hundreds of bases or below 5 kb in the worst case of fully linked markers). The *nvdF*$_{ST}$ method performs acceptably well when the target site is located at the centre of the studied region, the selection is not too strong ($\alpha \leq 600$) and the overall recombination rate is at least 0.3 cM/Mb ($\rho \geq 12$). In these cases, the selective location is estimated, at worst, within 33 kb of distance from the true location (Table 1). In the case with strong selection, the localization is still bad under high recombination (Table 2, $\alpha = 6000$). However, this could be due to the lower number of segregating sites (only 62 in Table 2).

In addition, it should be noted that the proportion of detected sites that lies at, or close to, the true selective position depends on the linkage relationships among the markers, which is obviously influenced by the recombination rate. Thus, if we consider the distance in centimorgans (cM), the mentioned differences in our results, in distances measured in Kb, between the distinct recombination rates are not so important since all detected sites lay within 1 map unit from the true selective position. For example, if we require that any detected SNP should be within 0.15 map units from the true position then under a recombination rate of 0.3 cM/Mb we accept any detected site no more than 500 Kb away from the right position. However, in the cases under recombination rate of 1.5 cM/Mb we accept sites that are no more than 100 Kb away.

Importantly, the localization is also dependent on 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-

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

122 kb as the target is shifted away from the center. This issue has already been shown for other HAC-based methods (Rivas, Dominguez-Garcia, and Carvajal-Rodriguez 2015).

The problem is partially solved, when the recombination is high (1.5 cM/Mb), by using the EOS test. In this case, the test has high power (67-93%) and good localization of the selective position. In fact, the position of the selective site is almost perfectly estimated (few bases or kb) when the true position is not at the extremes. Even if the target sites are at the extremes the localization is within 40 kb (see cases with ρ = 60 in Table 4).

In the case of independent markers with the selective site located at the center (Table 3) the localization by EOS was perfect in 98% of the replicates. However, in the table the average appears as ±7 kb because of two runs in which the localization failed by almost 400 kb. These two runs correspond to lower F_{ST} index values that were marginally considered given the cutoff for selecting the SNPs. Thus, we could count these two as false positives or alternatively, by making a bit more astringent the EOS classification in the upper class, we would have already discarded 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 only from 61 to 59% while getting perfect localization of the selective SNPs in every run.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 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.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Bottleneck-expansion Scenarios

Bottleneck-expansion scenarios are known to leave signatures that mimic the effect of positive selection. We tested the robustness of the tests by looking for false positives when applying the methods to a bottleneck and expansion scenario under a neutral setting. The bottleneck was simulated by a reduction of one of the populations to 1% of the original size (*N* from 1000 to 10). Afterwards, the population expansion was implemented by increasing the population size following a logistic growth model (see details in Appendix A-6). The methods performed well, the false positive rate is maintained below the nominal level with 4.6% and 1% for $nvdF_{ST}$ and EOS tests, respectively.

High Migration Scenario

For the short-term (500 generations) scenario with *Nm* = 50 (5%), $nvdF_{ST}$ 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). Concerning the EOS test, it works only under weak linkage and with a power of 16%. EOS committed no false positives under this setting.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 *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.

HapMap Data

In order to check how the *nvdF*$_{ST}$ method works under real data, we have analyzed phased haplotypes from human chromosome 2. Concretely, we analyzed the chromosome 2 from northern and western European (CEU); East Asian (ASN: CHB + JPT) and Yoruba in Nigeria (YRI) from Phase III of the International HapMap Project (Consortium 2010). The data consisted in 116,430 SNPs in the unrelated samples downloaded from the project page (http://hapmap.ncbi.nlm.nih.gov/, last accessed May 29, 2016). The original data files as well the command line options for the analysis with HacDivSel, jointly with the obtained output files are available as supplementary material.

The HacDiveSel program automatically filters the data for using only SNPs shared between each pair of populations, therefore we analyzed 86,483 SNPs under the ASN-CEU

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

comparison; 79,272 under the ASC-YRI and 74,582 for the CEU-YRI. We set the program to select the 0.1% highest *nvd* values within each assayed window size. For these candidates we only considered those having a positive value under the sign test (5) jointly with a significant *p*-value under the F_{ST} test.

In Table 6 (see also supplementary files) we provide the genomic regions in which we found significant SNPs and the corresponding genes within these regions. The estimated *q*-values were below 0.01 in every case. We found some of the highest *nvd* values, involving more than 40 SNPs, within the region 135-136 Mb in CEU-YRI comparison which includes SNPs in a quite precise localization (135.78-135.83) of the lactase (LCT) gene jointly with some well-known linked candidates for recent selective sweeps as RAB3GAP1, ZRANB3, MCM6 and R3HDM1 genes (Bersaglieri et al. 2004; Voight et al. 2006; Sabeti et al. 2007). The SNPs in this region showed high F_{ST} values (Table 6). This result was expected as it is known that Yoruba does not show signature of selection at LCT while the signature is strong in CEU (Voight et al. 2006).

Another strong signal also occurred for the 218.9 - 219.4 region which includes genes previously reported under divergent selection as TTLL4 (Sabeti et al. 2007) as well as genes related to growth and immune maturation as NHEJ1 (Buck et al. 2006), this region also presents some significant signal in the ASN-YRI comparison. Also within this region, we have detected in the CEU-YRI comparison, the IL34 gene that has been previously reported as giving strong signature of selection in the YRI population (Ferrer-Admetlla et al. 2014). Another interesting signal in this region is at the SNP rs613539 which corresponds to SLC23A3 gene that has been claimed to be a candidate for schizophrenia susceptibility in the

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Japanese population (Shibata et al. 2013). This SNP has been significant in CEU-YRI and in ASN-YRI comparisons but not in the ASN-CEU one (Table 6 and Supplementary Files).

For the CEU-YRI comparison two more regions were detected including NBAS and CLASP1. NBAS has also been detected in the ASN-CEU pair (see below). The CLASP1 gene is close to GLI2 already reported as a top region by the XP-CLR method (Chen, Patterson, and Reich 2010). In the ASN-YRI comparison, the region 203.4-203.9 were detected which is close to the NIF3L1 gene that includes an SNP previously reported under non-synonymous differentiation in these populations (Sabeti et al. 2007).

Finally, the significant *nvd* candidates having lowest F_{ST} values correspond to the regions detected in the CEU-ASN comparison, which also involves a lower number of regions. In this comparison, the highest *nvd* occurred in the NBAS region that had also the strongest signal in the CEU-YRI pair. Another interesting region is the 27.9-28.1 which includes the BRE gen (Table 6).

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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. Top significant divergent selection regions of the human chromosome 2 based on the *nvd*F_{ST} test for the 0.1% highest *nvd* values at different window sizes.

Populations	# significant SNPs	Positions (Mb)	F_{ST}	Genes
ASN-CEU	67	15.49-15.54	0.12-0.17	NBAS
		27.95-28.1	0.21	RBKS, MRPL33, BRE
		63.6-63.98	0.15-0.21	WDPCP, UGP2, ACA59
		83.23-83.24	0.16	Intergenic
ASN-YRI	46	27.98-28.09	0.25	MRPL33, BRE
		56.89	0.16	intergenic
		135.69-135.72	0.23-0.35	ZRANB3
		203.4-203.9	0.18-0.29	ICA1L, WDR12, CARF, CYP20A1
		117.28-117.30	0.22-0.36	intergenic
		210.26	0.19	MAP2
		218.99-219.2	0.29-0.53	CTDSP, VIL1, AK302678, NHEJ1-XLF, SLC23A3, piR-39082, ABCB6, AK091345, ATG9A
CEU-YRI	116	15.52-15.53	0.29	NBAS
		27.9	0.25	RBKS-MRPL33
		122.02-122.04	0.22-0.23	CLASP1
		135.4-136.9	0.29-0.63	ACMSD, CCNT2-AS1, MAP3K19, RAB3GAP1, ZRANB3, LCT, MCM6, R3HDM1
		178.08	0.41	AGP
		218.8-219.4	0.28-0.43	ARPC2, LINC00608, VIL1, USP37, NHEJ1-XLF, RQCD1, SLC23A3, IL34, ZNF142, BCS1L, GLB1L, STK36, TTLL4, PTPRN, CYP27A1

The positions correspond to the minimum and maximum positions for each given region. The F_{ST} range corresponds to the minimum and maximum F_{ST} for each region.

Littorina saxatilis **Data**

The rough periwinkle (*Littorina saxatilis*) is a marine gastropod mollusk that represents an interesting system for studying adaptive divergence and parallel speciation at different

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

spatial scales. In Europe, *L. saxatilis* has adapted to different shore habitats resulting in an exposed-to-wave ecotype and other non-exposed but crab-accessible ecotype. The crab ecotype has large thick shells while the wave ecotype consists in smaller snails with thin shells and a larger shell aperture. Several experimental studies have shown that these ecotypes have been able of evolving local adaptation in the face of gene flow even at small spatial scales (Rolan-Alvarez, Austin, and Boulding 2015).

Ravinet *et al.* (2016) have recently published a study in which they use RAD loci as dominant markers to quantify shared genomic divergence amongst *L. saxatilis* ecotype pairs (wave vs crab) on three close islands on the Skagerrak coast of Sweden. These islands are connected by weak gene flow. In their outlier analysis they filtered the data for sex-linked loci and null alleles. The filtering was applied for each island separately.

We applied the EOS test to analyze the separate-island filtered loci from Ravinet *et al.* We have discarded the loci with null allele frequency equal or higher than 0.5. We also discarded those polymorphisms not shared between ecotypes from the same location. Additionally, we required a minimum frequency allele of 4 per metapopulation sample size. Thus, we have excluded about 10-20% of the original individual-island filtered loci. The results of the between ecotypes outlier analysis using EOS are shown in Table 7. We can appreciate that the number of outliers detected as significant is much less than in the original study since we find a total of 69 outliers in the three islands while the number originally found was 406 (RAD loci in Table 2 of Ravinet et al. 2016). This is not surprising given the conservative nature and low false positive rate of EOS. However, note that we find a 2.9% (2/69) of SNPs shared by all islands which is quite similar to the 2.2% (9/406) found in the original study.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Considering the islands by pairs, Jutholmen and Ramsö share 2 outliers while Saltö has no outlier in common with Jutholmen and just 1 with Ramsö.

Table 7. Outliers detected after EOS analysis of the individual-island filtered loci from *Littorina saxatilis* data. Numbers in parenthesis refer to the results in the original analysis (Ravinet et al. 2016).

Island	Unique	Only with Jutholmen	Only with Ramsö	Only with Saltö	Shared all	Total
Jutholmen	27 (59)	__	2 (13)	0 (16)	2 (9)	31 (97)
Ramsö	24 (86)	2 (13)	__	1 (21)	2 (9)	29 (129)
Saltö	6 (134)	0 (16)	1 (21)	__	2 (9)	9 (180)

For the outliers in EOS, the F_{ST} between ecotypes ranges between 0.4-0.6 (Table 8). The *q*-values are high (0.52 - 0.76) although we already know by the simulations that this may be indicating linkage between the markers more than a real false positive rate (see also De Villemereuil et al. 2014).

Table 8. 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.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Discussion

The aim of this study was to develop two methods, haplotype-based and outlier-based, for the detection of divergent selection in pairs of populations connected by migration. We also intended that the methods be robust to false positives. High rate of false positives is a known concern of outlier-based methods (De Mita et al. 2013; De Villemereuil et al. 2014; Lotterhos and Whitlock 2014) thus, EOS was especially designed to minimize the false positive rate. Additionally, both methods should be useful for non-model species and so, it should not be necessary to have a priori functional information on candidate regions or perform neutral simulations to obtain critical cut-off values.

For the *nvdF*$_{ST}$ method, it has been shown that combining haplotype-based and F_{ST} differentiation information is a quite powerful strategy for detecting divergent selection. However, the *nvdF*$_{ST}$ algorithm does not perform well when the whole set of markers is segregating independently. To deal with this latter case, a second method was proposed based on the idea that the outliers caused by 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 aforementioned 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.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 Architecture

In general, the F_{ST}-based methods cannot detect selection in polygenic scenarios (Bierne, Roze, and Welch 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 or if the overall F_{ST} is high. On the contrary, the *nvd*F_{ST} performs even better in this scenario. The explanation for this good performance is that the distributed selective signal facilitates the discovery of the corresponding patterns of divergent selection. These patterns imply high frequency at the target site in one population and low in another. Therefore, the F_{ST} index under the neutral expectation would be very low compared with the observed one and so, still under high overall F_{ST}, *nvd*F_{ST} maintains high power under the polygenic setting.

Position Effect, Positive Predictive Value and FDR by site

Besides the detection of the signal of selection, we have also inferred the location of the selective site. It has been shown that under *nvd*F_{ST} the localization is better when the selective site is at the center of the chromosome. 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 *nvd*F_{ST} and EOS methods in this regard. Trying, for example, to further explore the relationship between recombination and the window sizes that yield 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 July 28, 2016. 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.

Concerning the measuring of positive and false rates by site, it should be noted first that under the neutral scenario we always considered a false positive when any site within the chromosome has been signaled as selective. Thus, when we are saying that FPR is below 5% this means that, in more than 95% of the neutral files no single site was detected as selective. That being said, we can consider the number of sites advocated as selective in those few runs in which neutrality was falsely rejected. In the case of the EOS test this number was just 1 single site per Mb when markers are independent or 1-4 when they are linked.

In the case of $nvdF_{ST}$ the situation is different just because various window sizes are assayed and a number k of candidates is considered, so that the maximum number of false positives by genome can be at most wk where w is the number of different window sizes. However, the number of false positives would also be dependent on the recombination rate. In most cases we have assayed 2-4 window sizes and just 1 candidate and the mean number of falsely detected sites in the neutral files was close to 1.

When a higher number of candidates (k>1) was studied we found in general less than k false positives in those neutral files in which selection was detected. In any case, the number of false positives within each run was always below the 1% of the total number of possible sites. For example, in the extreme case of inspecting the 99.9% of the sites (about 300 candidates) in a neutral scenario we found less than 1% of false positives runs (less than 10 out of 1000 runs) for every window size. Within those false positives, most of the files had just 1 site falsely assigned as selective and when not, the maximum number of false candidates was always below 1% (e.g. 3 in 340).

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 the selective scenarios we are also interested in the per site comparison, i.e. in the proportion of detected sites that lies at, or close to, the true selective position. The linkage relationship will depend on the recombination rate and so we would measure the distance in centimorgans (cM). Therefore, if we consider a given position as being true only when it lies at distance less than *t* cM from the real position, then the positive predictive value (PPV) can be measured as PPV = #(detected sites at distance ≤ *t*) / # detected sites. Similarly, the per-site false discovery rate is, FDR_I = 1 -PPV.

Thus, given the positions obtained in our simulations, when we set *t* = 0.15 cM the PPV is in most of the cases 100% i.e. all detected positions lie within the acceptance region except for some cases under the highest recombination rate (ρ = 60). In the latter cases under the $nvdF_{ST}$ test, the PPV decreases to 85-95% with a worst case of 70% when the selective site is located at the very extreme of the chromosome. Thus, in these latter cases ($nvdF_{ST}$; ρ = 60), when claiming that selection has been detected we are assuming that 5-15% of the times the detected position could be far away from 0.15 cM. However, If we assume as correct every position within 1 cM distance from the true one (*t* = 1) then the per-site FDR_I is virtually zero in every studied case. The latter means that any detected site lies closer than 1 map unit from the real position.

High Migration Scenario

It is important to note, that under high migration, $nvdF_{ST}$ maintains reasonable power (60%). However, the power diminishes with the highest recombination rate. This may occur due to the combined effect of gene flow and recombination that generates intermediate HAC mean

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

values m_1 and m_2 and similar variances. Indeed, for a given selection intensity, the higher the *Nm* requires tighter linkage for the establishment of divergent alleles (Yeaman and Whitlock 2011).

In the case of the EOS test, there is an obvious tradeoff between the astringency of the cutoff point for the outlier set and the migration rate. Our cutoff depends on the F_{ST} upper-bound which is a function of the number of populations, the sample sizes and the minimum allele frequency. However, with higher migration the upper bound of the F_{ST} should be lower because as migration increases the number of populations becomes virtually one. Therefore, a possible solution to improve the efficiency of EOS under high migration would be to update the F_{ST} upper bound as a function of the migration rate.

HapMap data

We evaluated the applicability of the *nvd*F_{ST} method to real data by analyzing the human chromosome 2. For each pair of populations, the full set of SNPs in the phased haplotypes was directly analyzed without any assumption about specific candidate regions or reference SNPs.

Our method successfully detected the lactase persistence region that is known to expand over more than 1 Mb in chromosome 2 of European populations (Bersaglieri et al. 2004; Tishkoff et al. 2007). We have also detected some other regions previously reported as under ongoing selective sweeps in human populations (see Table 6 and supplementary Table S2). Worth mentioning is the SLC23A3 gene that has been associated to schizophrenia

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

susceptibility in the Japanese population (Shibata et al. 2013) and that we found as significant both in ASN-YRI and CEU-YRI comparisons.

We found also some regions not previously reported, for example, in the ASN-CEU comparison, 20 SNPs with very high *nvd* value within the NBAS gene. Mutations in this gene has been associated to short stature and different multisystem disorders (Maksimova et al. 2010; Segarra et al. 2015). Again in the ASN-CEU comparison, the BRE gene was found with quite high *nvd* values. This gene encodes a component of the BRCA1-A complex. These results illustrate how the $nvdF_{ST}$ method can be used in exploratory studies for detecting locally adapted polymorphisms that can be interesting candidates for association studies.

Littorina saxatilis Data

Local adaptation may occur most likely due to alleles with large effect but also under a polygenic architecture (Whitlock 2015; Yeaman 2015). In addition, the geographic structure and the migration-selection balance can generate complex patterns of the distribution of genetic variation (Debarre, Yeaman, and Guillaume 2015). Thus, the natural systems where local adaptation occurs can be of great complexity (Whitlock 2015). 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, localities within country, and 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 habitat variation among islands —that may be caused by genetic drift— and variation between habitats within localities that may be caused by divergent

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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 (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.

Regarding the shared genomic divergence of the *L. saxatilis* system in Swedish populations, it seems to be a small proportion of the total genomic divergence (Hollander, Galindo, and Butlin 2015; Johannesson 2015; Ravinet et al. 2016). That is, the majority of the genomic variation linked to the evolution of ecotypes is not shared between the studied islands. The EOS analysis of the Ravinet *et al.* data seems to support this finding. At the same time, we identify far fewer outliers, with Saltö —which is closer to the mainland— having the lowest number. However, this is the opposite of the result in Ravinet's study where Saltö had the highest number of outliers.

In any case, the lower number of outliers that we found in Saltö, could also explain our results of reduced shared divergence between Saltö and the two other islands. We cannot rule out that our findings can be artefact due to the conservativeness of EOS, however; alternatively, the previous results could be due to an excess of false positives, hiding the pattern of reduced shared divergence in Saltö.

To conclude, it is worth mentioning that a combination of multiple signals from different tests have been proposed as a way of improving power/false positive rate relationships for the selection detection methods (Zeng, Shi, and Wu 2007; Lin et al. 2011; Vatsiou, Bazin, and Gaggiotti 2016). Accordingly, the *nvd*F_{ST} test does just that. It combines haplotype and population differentiation information and may be a helpful tool to explore patterns of divergent selection when approximate knowledge of the haplotype phase is at hand.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Alternatively, the EOS method is a conservative outlier test useful when the full set of SNPs is unlinked or under weak linkage. Both strategies can be applied without the need of performing neutral simulations and have a low false positive rate.

Acknowledgements

I thank E. Rolán-Alvarez and Auriel Sumner-Hempel for useful comments on the manuscript. I also thank to two anonymous reviewers from Peerage of Science for their time and effort in the evaluation of the manuscript (final manuscript evaluation: publishable; PAQ = 4.3/5). 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.

Software 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.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Bibliography

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

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet **74**:1111-1120.

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

Bourret, V., M. P. Kent, C. R. Primmer, A. Vasemägi, S. Karlsson, K. Hindar, P. McGinnity, E. Verspoor, L. Bernatchez, and S. Lien. 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.

Buck, D., L. Malivert, R. g. de Chasseval, A. Barraud, M.-C. FondanÃ¨che, O. Sanal, A. Plebani, J.-L. StÃ©phan, M. Hufnagel, and F. o. le Deist. 2006. Cernunnos, a novel nonhomologous end-joining factor, is mutated in human immunodeficiency with microcephaly. Cell **124**:287-299.

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

Consortium, T. I. H. 2010. Integrating common and rare genetic variation in diverse human populations. Nature **467**:52-58.

Crisci, J. L., Y.-P. Poh, A. Bean, A. Simkin, and J. D. Jensen. 2012. Recent Progress in Polymorphism-Based Population Genetic Inference. Journal of Heredity.

Crisci, J. L., Y.-P. Poh, S. Mahajan, and J. D. Jensen. 2013. The Impact of Equilibrium Assumptions on Tests of Selection. Frontiers in Genetics **4**.

Crow, J. F., and M. Kimura. 1970. An Introduction to Population Genetics Theory. Harper & Row, New York.

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

Cheverud, J. M. 2001. A simple correction for multiple comparisons in interval mapping genome scans. Heredity **87**:52-58.

De Mita, S., A.-C. Thuillet, L. Gay, N. Ahmadi, S. Manel, J. Ronfort, and Y. Vigouroux. 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., É. Frichot, É. Bazin, O. François, and O. E. Gaggiotti. 2014. Genome scan methods against more complex models: when and how much should we trust them? Molecular Ecology **23**:2006-2019.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

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

Ellegren, H. 2014. Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution **29**:51-63.

Ferrer-Admetlla, A., M. Liang, T. Korneliussen, and R. Nielsen. 2014. On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. Molecular Biology and Evolution **31**:1275-1291.

Ferretti, L., S. E. Ramos-Onsins, and M. Pérez-Enciso. 2013. Population genomics from pool sequencing. Molecular Ecology **22**:5561-5576.

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

Hey, J. 2006. Recent advances in assessing gene flow between diverging populations and species. Current Opinion in Genetics & Development **16**:592-596.

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

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

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

Jensen, J. D., M. Foll, and L. Bernatchez. 2016. The past, present and future of genomic scans for selection. Molecular Ecology **25**:1-4.

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

Kim, Y., and R. Nielsen. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics **167**:1513-1524.

Lewontin, R. C. 1988. On measures of gametic disequilibrium. Genetics **120**:849-852.

Lewontin, R. C., and J. Krakauer. 1973. Distribution of Gene Frequency as a Test of Theory of Selective Neutrality of Polymorphisms. Genetics **74**:175-195.

Lin, K., H. Li, C. Schlotterer, and A. Futschik. 2011. Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics **187**:229-244.

Lotterhos, K. E., and M. C. Whitlock. 2014. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Molecular Ecology **23**:2178.

Maksimova, N., K. Hara, I. Nikolaeva, T. Chun-Feng, T. Usui, M. Takagi, Y. Nishihira, A. Miyashita, H. Fujiwara, and T. Oyama. 2010. Neuroblastoma amplified sequence gene is associated with a novel short stature syndrome characterised by optic nerve atrophy and Pelgerâ€“HuÃ«t anomaly. Journal of medical genetics:jmg. 2009.074815.

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

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005. Genomic scans for selective sweeps using SNP data. Genome Res **15**:1566-1575.

Pavlidis, P., J. D. Jensen, and W. Stephan. 2010. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics **185**:907-922.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

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

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81**:559-575.

Ravinet, M., A. Westram, K. Johannesson, R. Butlin, C. André, and M. Panova. 2016. Shared and nonshared genomic divergence in parallel ecotypes of Littorina saxatilis at a local scale. Molecular Ecology:287-305.

Renaut, S., A. W. Nolte, S. M. Rogers, N. Derome, and L. Bernatchez. 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, M. J., S. Dominguez-Garcia, and A. Carvajal-Rodriguez. 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.

Rolan-Alvarez, E., C. Austin, and E. G. Boulding. 2015. The contribution of the genus Littorina to the field of evolutionary ecology. Oceanography and Marine Biology: an Annual Review **53**:157-214.

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

Sabeti, P. C.P. VarillyB. FryJ. LohmuellerE. HostetterC. CotsapasX. XieE. H. ByrneS. A. McCarrollR. GaudetS. F. SchaffnerE. S. LanderK. A. FrazerD. G. BallingerD. R. CoxD. A. HindsL. L. StuveR. A. GibbsJ. W. BelmontA. BoudreauP. HardenbolS. M. LealS. PasternakD. A. WheelerT. D. WillisF. YuH. YangC. ZengY. GaoH. HuW. HuC. LiW. LinS. LiuH. PanX. TangJ. WangW. WangJ. YuB. ZhangQ. ZhangH. ZhaoH. ZhaoJ. ZhouS. B. GabrielR. BarryB. BlumenstielA. CamargoM. DefeliceM. FaggartM. GoyetteS. GuptaJ. MooreH. NguyenR. C. OnofrioM. ParkinJ. RoyE. StahlE. WinchesterL. ZiaugraD. AltshulerY. ShenZ. YaoW. HuangX. ChuY. HeL. JinY. LiuY. ShenW. SunH. WangY. WangY. WangX. XiongL. XuM. M. WayeS. K. TsuiH. XueJ. T. WongL. M. GalverJ. B. FanK. GundersonS. S. MurrayA. R. OliphantM. S. CheeA. MontpetitF. ChagnonV. FerrettiM. LeboeufJ. F. OlivierM. S. PhillipsS. RoumyC. SalleeA. VernerT. J. HudsonP. Y. KwokD. CaiD. C. KoboldtR. D. MillerL. PawlikowskaP. Taillon-MillerM. XiaoL. C. TsuiW. MakY. Q. SongP. K. TamY. NakamuraT. KawaguchiT. KitamotoT. MorizonoA. NagashimaY. OhnishiA. SekineT. TanakaT. TsunodaP. DeloukasC. P. BirdM. DelgadoE. T. DermitzakisR. GwilliamS. HuntJ. MorrisonD. PowellB. E. StrangerP. WhittakerD. R. BentleyM. J. DalyP. I. de BakkerJ. BarrettY. R. ChretienJ. MallerS. McCarrollN. PattersonI. Pe'erA. PriceS. PurcellD. J. RichterP. SabetiR. SaxenaS. F. SchaffnerP. C. ShamP. VarillyD. AltshulerL. D. SteinL. KrishnanA. V. SmithM. K. Tello-RuizG. A. ThorissonA. ChakravartiP. E. ChenD. J. CutlerC. S. KashukS. LinG. R. AbecasisW. GuanY. LiH. M. MunroZ. S. QinD. J. ThomasG. McVeanA. AutonL. BottoloN. CardinS. EyheramendyC. FreemanJ. MarchiniS. MyersC. SpencerM. StephensP. DonnellyL. R. CardonG. ClarkeD. M. EvansA. P. MorrisB. S. WeirT. TsunodaT. A. JohnsonJ. C.

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

MullikinS. T. SherryM. FeoloA. SkolH. ZhangC. ZengH. ZhaoI. MatsudaY. FukushimaD. R. MacerE. SudaC. N. RotimiC. A. AdebamowoI. AjayiT. AniagwuP. A. MarshallC. NkwodimmahC. D. RoyalM. F. LeppertM. DixonA. PeifferR. QiuA. KentK. KatoN. NiikawaI. F. AdewoleB. M. KnoppersM. W. FosterE. W. ClaytonJ. WatkinR. A. GibbsJ. W. BelmontD. MuznyL. NazarethE. SodergrenG. M. WeinstockD. A. WheelerI. YakubS. B. GabrielR. C. OnofrioD. J. RichterL. ZiaugraB. W. BirrenM. J. DalyD. AltshulerR. K. WilsonL. L. FultonJ. RogersJ. BurtonN. P. CarterC. M. CleeM. GriffithsM. C. JonesK. McLayR. W. PlumbM. T. RossS. K. SimsD. L. WilleyZ. ChenH. HanL. KangM. GodboutJ. C. WallenburgP. L'ArchevequeG. BellemareK. SaekiH. WangD. AnH. FuQ. LiZ. WangR. WangA. L. HoldenL. D. BrooksJ. E. McEwenM. S. GuyerV. O. WangJ. L. PetersonM. ShiJ. SpiegelL. M. SungL. F. ZachariaF. S. CollinsK. KennedyR. Jamieson, and J. Stewart. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature **449**:913-918.

Schubert, E., A. Zimek, and H.-P. Kriegel. 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.

Segarra, N. G., D. Ballhausen, H. Crawford, M. Perreau, B. Campos-Xavier, K. van Spaendonck-Zwarts, C. Vermeer, M. Russo, P.-Y. Zambelli, B. Stevenson, B. Royer-Bertrand, C. Rivolta, F. Candotti, S. Unger, F. L. Munier, A. Superti-Furga, and L. Bonafé. 2015. NBAS mutations cause a multisystem disorder involving bone, connective tissue, liver, immune system, and retina. American Journal of Medical Genetics Part A **167**:2902-2912.

Shibata, H., K. Yamamoto, Z. Sun, A. Oka, H. Inoko, T. Arinami, T. Inada, H. Ujike, M. Itokawa, and M. Tochigi. 2013. Genome-wide association study of schizophrenia using microsatellite markers in the Japanese population. Psychiatric genetics **23**:117-123.

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

Smith, J. M., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genetical research **23**:23-35.

Sousa, V. C., M. Carneiro, N. Ferrand, and J. Hey. 2013. Identifying Loci Under Selection Against Gene Flow in Isolation-with-Migration Models. Genetics **194**:211-233.

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

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

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet **39**:31-40.

Tukey, J. W. 1977. Exploratory data analysis. Addison-Wesley, Reading, Mass.

Vatsiou, A. I., E. Bazin, and O. E. Gaggiotti. 2016. Detection of selective sweeps in structured populations: a comparison of recent methods. Molecular Ecology **25**:89-103.

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

bioRxiv preprint doi: https://doi.org/10.1101/026369; this version posted July 28, 2016. 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.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biol **4**:e72.

Whitlock, M., C., and K. Lotterhos, 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, M. C. 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., and S. P. Otto. 2011. Establishment and Maintenance of Adaptive Genetic Divergence under Migration, Selection, and Drift. Evolution **65**:2123-2129.

Yeaman, S., and M. C. Whitlock. 2011. The Genetic Architecture of Adaptation under Migration-Selection Balance. Evolution **65**:1897-1911.

Zeng, K., S. Shi, and C. I. Wu. 2007. Compound tests for the detection of hitchhiking under positive selection. Mol Biol Evol **24**:1898-1908.