UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole de biologie

A strategy to detect adaptive introgression in clinal populations

Travail de Maîtrise universitaire ès Sciences en Sciences moléculaires du vivant

Master Thesis of Science in Molecular Life Sciences

par

Marie ZUFFEREY

Directeur: Docteur Nils Arrigo
Superviseur: Docteur Nils Arrigo
Expert: Professeur Nadir Alvarez

Département d'écologie et d'évolution

Résumé

Les méthodes génomiques actuelles permettent d'étudier l'introgression adaptative, procédé par lequel le flux de gènes interspécifique accroît la fitness d'un individu, à un niveau de précision sans précédent. Néanmoins, identifier l'importance de tels transferts d'adaptations entre espèces reste un véritable défi, notamment en raison de notre compréhension limitée des bases génétiques de l'adaptation locale. Ainsi, la majorité des approches visant à détecter les loci adaptatifs entre populations et/ou le long de gradients environnementaux souffrent de taux élevés de faux positifs et s'accompagnent d'une valeur explicative limitée. De plus, la plupart de ces approches ne conviennent pas à l'étude de populations hybrides, se concentrant plutôt sur les différentiations intraspécifiques. Dans cette étude, nous proposons un cadre général pour identifier les loci ayant introgressé de façon adaptative impliqués dans l'adaptation le long de gradients environnementaux. Notre stratégie procède en trois étapes principales: 1) nous identifions les allèles d'origine introgressive; 2) nous relions ensuite leurs fréquences à un gradient environnemental et considérons comme outliers celles qui sont surreprésentées sous des conditions sélectives, par rapport à un arrière-plan génomique neutre (non soumis à sélection); 3) finalement, nous affinons notre liste d'outliers en nous concentrant sur ceux partagés entre plusieurs zones hybrides indépendantes. Nous validons notre approche au moyen de simulations implémentant de la sélection polygénique, sous quatre régimes différents de flux génétique et de sélection. La performance de notre approche est comparable à une autre méthode de détection d'outliers bien établie. Si un génome de référence est disponible, nous illustrons ensuite comment notre liste de loci candidats peut être améliorée en inspectant leur position chromosomique, tirant profit du fait qu'il est attendu que les faux positifs soient physiquement liés aux loci sous sélection, tout en ayant des fréquences légèrement inférieures dans les populations hybrides. En conclusion, notre approche pourrait servir à identifier avec plus d'assurance les loci sous-jacents à l'adaptation dans des populations hybrides. Et en combinaison avec une analyse fonctionnelle ultérieure, cela pourrait conduire à une amélioration de notre compréhension à la fois des processus d'introgression et de ceux d'adaptation.

Abstract

Adaptive introgression, the process by which interspecific gene flow increases individual fitness in a given environment, is being investigated with unprecedented accuracy owing to recent advances in genomics studies. However, unravelling the prevalence of such transfers among species remains a challenging task, notably because of our limited understanding of the genetic bases of local adaptation. Accordingly, the majority of approaches attempting to detect adaptive loci among populations and/or along environmental gradients suffer from high false positive rates and generally yield limited explicative value. In addition, most approaches are not suited for the study of admixed populations and rather focus on intraspecific differentiations. Here, we suggest a general framework to identify adaptively introgressed loci being involved in adaptation along environmental gradients. Our strategy proceeds in three main stages: 1) we identify alleles of introgressed origin; 2) we then relate their frequencies to an environmental gradient and consider as outliers those being over-represented under selective conditions, compared to the neutral (released from selection) genomic background; 3) finally, we refine our list of outliers by focusing only on those being shared among independent hybrid zones. We validate our approach using simulations implementing polygenic selection under four different gene flow and selection regimes. The performance of our approach is comparable to another well-established outlier detection method. If a reference genome is available, we further illustrate how our list of candidate loci can be improved by inspecting their chromosomal location on a reference genome, leveraging the fact that false positives are expected to be linked to selected loci, but show slightly lower frequencies in the admixed populations. In conclusion, our approach could serve to identify more confidently the loci underlying adaptation in admixed populations. And in combination with downstream functional analysis, it may lead to an improvement in our understanding of both introgression and adaptation processes.

A strategy to detect adaptive introgression in clinal populations

Marie Zufferey, Nils Arrigo

Department of Ecology and Evolution - University of Lausanne

Introduction

Unravelling the mechanisms promoting local adaptation is a key concern in ecology and evolutionary biology (Kawecki and Ebert 2004). For adaptation to take place, genetic variability upon which natural selection can act is essential. In this context, interspecific crosses (i.e. hybridization events) are a likely source of genetic variation and might act as a driver of evolution (Baack and Rieseberg 2007). As hybridization typically occurs more often than background mutations (Lynch 2007; Whitney et al. 2010), the role of genetic variants inherited through interspecific gene flow (i.e. introgression) in local adaptation processes might be prominent. Importantly, evolutionary changes requiring multiple allele substitutions or gene modifications might be more easily achieved through hybridization than by means of other triggers of genetic variation such as *de novo* mutations or standing variation (Barrett and Schluter 2007; Rieseberg 2009; Hedrick 2013).

Historically, the potential of hybridization as an evolutionary force has long been debated (e.g. Barton 2001). Its prevalence has however been increasingly appreciated during the last decade (Tigano and Friesen 2016), with the advent of modern molecular methods that outlined interspecific genetic exchanges in every life kingdoms (e.g. Cronn and Wendel 2003; Twyford and Ennos 2012; Mallet et al. 2015; Rosenzweig et al. 2016). Accordingly, recent investigations provided evidence that gene flow among species is more widespread than previously appreciated (Scascitelli et al. 2010 and references therein). In particular, several studies demonstrated the transfer of ecological and morphological adaptations between species (i.e. adaptive introgression; Hedrick 2013), for instance the flood tolerance in *Iris* species (Martin et al. 2006), the radiate morphology of *Senecio vulgaris* (Kim et al. 2008), or the wing patterns of *Heliconius* butterflies (Nadeau et al. 2012). Those examples highlight the unique opportunity that introgression represents to identify adaptive loci, via the fine-scale inspection of those genes being preferentially retained in the recipient genome under selective forces (e.g. Hamilton et al. 2013). For example, artificial phenotype-based selection and introgression has been successfully used with *Drosophila* spp. to dissect the genetic architecture of a complex behavioural trait (Earley and Jones 2011). Additionally, relying on ancestry information to detect adaptive loci, rather than using intraspecific sequence variants might be preferable as it is less impacted by allelic heterogeneity or multiple independent mutations, and more efficient for the detection of low polymorphic genes (McKeigue 2005; Seldin 2007). In addition, introgression-based approaches are particularly suitable when the parental populations are highly divergent (Darvasi and Shifman 2005; Smith and O'Brien 2005; Crawford and Nielsen 2013) and usually provide higher statistical power (McKeigue 2005), especially through the greater number of unique different loci identified in the parental lines (Earley and Jones 2011). From above it follows that hybrid zones or artificial introgression hold great potential for unravelling the genetic bases of adaptation (e.g. Lexer et al. 2004; Earley and Jones 2011).

Box 1 - More than five years of simulation studies Pérez-Figueroa and colleagues (2010) simulated two populations in an island model, with different proportions of true selective loci (ranging from 0 to 10%) to evaluate the efficiency of three population differentiation (PD) methods (DFDIST (Beaumont and Nichols 1996), DETSELD (Vitalis et al. 2003), BayeScan (Foll and O. Gaggiotti 2008)). They showed that BayeScan was the most powerful software in terms of true and false positives. Still centered on PD methods, Narum and Hess (2011) tested four models - either weak or strong selection occurring either in the same direction of gene flow or randomized - for ten populations along a clinal gradient (95 neutral markers, five additive quantitative traits - each with one locus). According to this study, BayeScan was again more preferable than other softwares (FDIST2 (Beaumont and Nichols 1996), Arlequin (Excoffier and Lischer 2010)) due to lower type I and II errors. Other researchers used six design combinations (two rates of self-fertilization, three different migration models) for 751 unlinked loci (among which ten markers, one under selection) to compare genetic-environment association (GEA) with PD methods (De Mita et al. 2013). This study highlighted PD methods as more specific (less false positives) than GEA methods, but these latter might have more power. Then, the issue of linkage between loci for PD methods was addressed by Vilas and colleagues (Vilas et al. 2012). In that work, the authors investigated the outlierness of neutral markers flanking loci controlling a quantitative trait under divergent selection in two subpopulations connected by migration. Another study (Jones et al. 2013) focused specifically on regression methods and simulated one clinal population and 100 loci (99 neutral, one under selection). Jones and colleagues noted a decrease of performance of these landscape genomic methods under the "weak" selection scenario, especially in terms of type II errors. De Villemereuil and coworkers (De Villemereuil et al. 2014) investigated performance of GEA and PD methods (linear regression, BayeScan, BayEnv (Coop et al. 2010) and LFMM (Frichot et al. 2013)) when polygenic selection is at work for different demographic scenarios and population structures (highly structured isolation, isolation with migration, stepping-stone). Overall, they stressed out that the power and error rate of the softwares depend on the scenario tested. Interestingly, the authors showed that BayeScan "was always less powerful than at least one of the other methods", the better compromise being offered by LFMM. Also for assessing performance of GEA and PD methods, Lotterhos and Whitlock (2015) conducted a comprehensive survey by testing for 20 different sampling designs, two distinct sampling approaches (pairs and transects) and 100 selected loci. In terms of power, their results suggested that GEA tests should be preferred under island model and PD under isolation by distance scenarios. As for the sampling, transect or random strategies generally came with lower power than paired ones.

The quest for identifying adaptive loci has spanned a large body of literature and promoted the development of statistical approaches concurrently to the accumulation of empirical genomic data. Those statistical approaches are classified in two broad categories that either rely on population differentiation (mostly based on F_{ST} or F_{ST}-like statistic) or proceed by regressing allele frequencies on environmental gradients (genetic-environment association) (e.g. Lotterhos and Whitlock 2015; François et al. 2016). Notwithstanding the wealth of analytical tools available, identifying loci under selection remains a challenging task. Accordingly, simulation-based studies (Box 1) have highlighted the prevalence of false positives, that calls for caution at the time of result interpretation. In addition, the appropriateness of each outlier detection method largely depends on the underlying demographic scenarios - which often remain elusive under field conditions - and

even on the sampling design (e.g. De Mita et al. 2013; Lotterhos and Whitlock 2015). Moreover, a recommendation that is often formulated is to combine different statistical tests (e.g. Narum and Hess 2011; De Villemereuil et al. 2014; François et al. 2016), as the consensus on true positives is expected to be higher than that on false positives. Finally, it is worth pointing out that these methods and simulations considered intraspecific populations and usually disregarded hybrid zone scenarios.

Here, we aim at bridging this gap, by presenting a strategy for the detection of loci that underly local adaptation in clinal admixed populations. We first assume that the species origin of alleles can be identified accurately. In a simulation context, this task is trivial but real-world datasets will require a preliminary classification of alleles as being either "resident" or "introgressed" (see Joly et al. 2009; Lawson et al. 2012; Ward and van Oosterhout 2016; or Twyford and Ennos 2012 for a review). Then, as the frequencies of introgressed alleles are expected to reflect the intensity of selection, we propose that the locus-specific area under the curve (AUC) obtained by regressing those frequencies to a clinal proxy can fittingly serve as a test statistic to detect adaptive loci. In fact, the AUC of the alleles under selection is expected to be markedly higher than that of the non adaptive loci. We benchmark our approach with simulated hybrid populations under positive selection for a complex trait, i.e. a trait controlled by a large number of genes. Four scenarios were tested, resulting from the combination of variable and/or uniform gene flow and selection strength. These two conditions potentially confound the detection of adaptive genes, as the former could obscure the signature of selection and the latter affect the performance of the outlier detection. Notably, we show that - even in the presence of strong confounding factors - our method compares favorably to another established software performing genome scan for selection. In addition, acknowledging the fact that detecting adaptive loci is subject to false positives, we further assess the usefulness of biological replicates in improving the specificity and sensitivity of our approach. Finally, if a reference genome is available, we further recommend the close examination of introgressed blocks that are putatively adaptive for an ultimate refinement of the list of candidate loci. We illustrate this latter point with a worked example. Overall, our three-step procedure yields considerable power to detect adaptive genes in populations sampled along an environmental gradient in a hybrid zone.

Material and methods

Outlier detection method

For the purpose of discriminating neutral markers from selected genes, we related allele frequencies to an environmental gradient and considered as outliers those being over-represented under selective conditions, compared to the neutral genomic background. Briefly, i) we computed the frequencies of introgressed alleles for every locus in each population, ii) regressed those frequencies on an explanatory variable (here, a geographical position along a selection gradient), iii) estimated the area under the obtained curve (AUC), and finally iv) inspected the distribution of the AUC values obtained across all loci to select outliers among those showing the largest AUC. We implemented two versions of this AUC calculation, either by taking the integral of a generalized linear model (i.e. a GLM fitted for each locus, with logit link function and binomial error distribution; using the *glm* function of the *stats* R package; R Core Team 2016) ("GLM AUC"), or by following the trapezoid rule (as implemented in the *auc* function of the *flux* R package; Jurasinski et al. 2014) ("Smooth AUC"). Notably, this latter does not require to meet any assumption of the linear models.

We then validated our AUC method by comparing its performance to *pcadapt* (Luu et al. 2016). This outlier detection method is a recently released algorithm relying on principal component analysis (PCA). Several reasons motivated the choice of this software. First and most importantly, the presence of admixed populations does not impact *pcadapt* (Luu et al. 2016). Then, a recent study showed that under isolation-by-distance model (the most appropriate to describe clinal populations) *pcadapt* detects more selected loci than other softwares (Lotterhos and Whitlock 2015; note that this was an earlier version - with a different implementation (test statistic) - of *pcadapt* than the one used for our analysis). Finally, it was also demonstrated to have a lower false discovery rate than the well-established BayeScan (Foll and Gaggiotti 2008) and to be one of the most powerful in the context of population divergence (Luu et al. 2016). From a computational standpoint, its fast running time was also a substantial advantage.

Implemented as an R package, *pcadapt* is individual-based (i.e. grouping individuals into populations is not required) and assumes that "markers excessively related to population structure are candidates for local adaptation" (Luu et al. 2016). It proceeds as follows. First, a PCA is performed to infer population structure on the basis of K first principal components (PC). After that, a vector containing K z-scores is calculated for each genetic marker (e.g. SNP), measuring to which extent a given locus is related to the K PCs. A Mahalanobis distance is then computed among markers, based on the z-scores, "to detect outliers for which the vector of z-scores does not follow the distribution of the main bulk of points" (Luu et al. 2016). This metric is "a multidimensional measure of the number of standard deviations that a point lies from the mean of a distribution" (Verity et al. 2016). In addition, it can be considered as robust as "the estimators of the mean and of the covariance matrix [...] required to compute the Mahalanobis distances, are not sensitive to the presence of outliers in the data set (Maronna & Zamar 2002)" (Luu et al. 2016). Correcting for covariance among samples, this distance is considered as a better statistic than the Euclidean distance for genome scan as it does not assume the independence of observations (Verity et al. 2016). However, it will tend to perform poorly if the distribution of the observations is complex or multimodal (Verity et al. 2016). In principle, the number of K eigenaxes must be chosen after visual inspection of a scree plot displaying the proportion of explained variance per principal component. After having checked some of these plots for the simulated data (Fig. S1/2), we chose $K = 10$ for all the analyses conducted in our study.

Our outlier AUC-based detection approach is a rank-based method: we considered as outliers the loci for which the AUC is above a given threshold, defined as a quantile - that has to be defined in an earlier stage - of the AUC estimate distribution. In the case of *pcadapt*, the cut-off was set as $\alpha = 1 - threshold$ and loci with a q-value smaller than α were classified as outliers (note that the q-values instead of the p-values were used to control the false discovery rate; computed with the *qvalue* R package; Storey 2015).

Simulations - overall model

We assumed two parental species, defined as "resident" and "external" that i) were interfertile and ii) differed for an adaptive complex trait (i.e. the external species being adapted). Due to the combined action of gene flow and selection, alleles could introgress from the external species into the resident and increase in frequency in populations where they were adaptive. We then considered a sampling scheme focused on the resident species, with specimens being collected in populations located along an environmental gradient and enduring varying selection pressures for the adaptive trait.

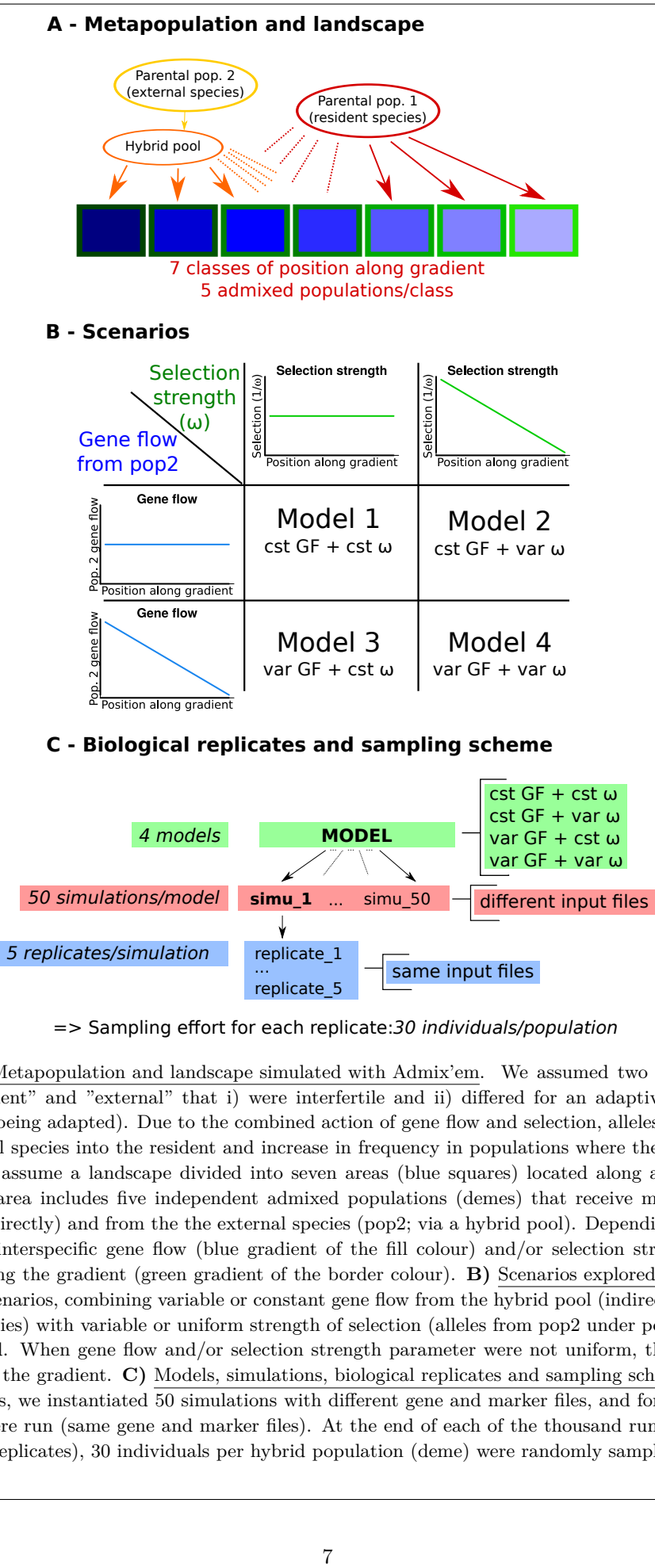

Figure 1: **A)** Metapopulation and landscape simulated with Admix'em. We assumed two parental species, defined as "resident" and "external" that i) were interfertile and ii) differed for an adaptive trait (i.e. the external species being adapted). Due to the combined action of gene flow and selection, alleles could introgress from the external species into the resident and increase in frequency in populations where they were adaptive. Our simulations assume a landscape divided into seven areas (blue squares) located along an environmental gradient. Each area includes five independent admixed populations (demes) that receive migrants from the resident (pop1; directly) and from the the external species (pop2; via a hybrid pool). Depending on the model, the amounts of interspecific gene flow (blue gradient of the fill colour) and/or selection strength parameter vary linearly along the gradient (green gradient of the border colour). **B)** Scenarios explored with Admix'em. Four different scenarios, combining variable or constant gene flow from the hybrid pool (indirect gene flow from the external species) with variable or uniform strength of selection (alleles from pop2 under positive selection), were investigated. When gene flow and/or selection strength parameter were not uniform, they were linearly decreasing along the gradient. **C)** Models, simulations, biological replicates and sampling scheme. For each of the four scenarios, we instantiated 50 simulations with different gene and marker files, and for each simulation five replicates were run (same gene and marker files). At the end of each of the thousand runs (4 models x 50 simulations x 5 replicates), 30 individuals per hybrid population (deme) were randomly sampled.

Simulations - genotypes, phenotypes and fitness function

We used Admix'em (Cui et al. 2016), a forward-in-time simulator, to generate genotypic data (configuration files for running Admix'em and allele tables were prepared with custom C++ scripts). This software models admixed populations where selection can be imposed on phenotypes defined using multiple loci. We considered diploid individuals, characterized by 951 neutral markers and 50 selected genes that are all randomly distributed along 10 chromosomes. The number of expected recombination events during meiosis was left to its default value (two per chromosome and per generation). The phenotype of individuals was set as the sum of allelic values at the 50 genes. These latter thus compose the architecture of the complex trait under selection. Allelic values were either 0 or 1 for the homozygous and 0.5 for the heterozygous alleles (mutations were not allowed). The external (adapted) species was declared as homozygous for the "1" allele across all genes (and thus showed a phenotypic value of 50) whereas the resident species was initially homozygous for the "0" allele and showed a phenotypic value of 0. The fitness of individuals was then computed according to Rhoné et al. (2011), using the following fitness function:

$$F(Z) = exp(-(Z - Z_{opt})^2/\omega^2)$$

where Z and Z_{opt} are the observed and optimal (corresponding here to that of the adapted species, i.e. Z_{opt} = 50) phenotypic values respectively, while $1/\omega$ represents the intensity of selection.

Simulations - metapopulation, landscape and scenarios

We implemented a metapopulation of 35 demes of the resident species evolving over 200 generations that experience varying levels of selection, and in which adaptive alleles from the external species were immigrating. In our models, three parental demes (pop1 - the resident species, pop2 - the external species, and hybFoo - an intermediate hybrid pool) served as an infinite source of parental and hybrid genotypes and sent migrants to those 35 demes where selection forces were at work (see Supplementary Materials for a comprehensive description; Tables S1/2/3).

Our 35 demes were distributed in a landscape organised along an environmental gradient, divided into seven spatial areas, that varied in levels of selection and/or interspecific gene flow (Fig. 1A). For convenience, we referred to those areas according to their position along the environmental gradient. The levels of selection and/or interspecific gene flow in each area were then determined according to their position along the gradient (see below). Each area contained five independent demes, and migration among demes/areas was not allowed (only interspecific gene flow could occur).

On this basis, we explored four different scenarios (Fig. 1B): i) model 1 (unif GF + unif ω): uniform interspecific gene flow and phenotype selection for all demes; ii) model 2 (unif GF + var ω): uniform gene flow, but the selection strength parameter ω varies linearly along the gradient; iii) model 3 (var GF + unif ω): uniform selection, but gene flow varies linearly along the gradient; iv) model 4: gene flow and ω vary linearly along the gradient (var GF + var ω). When selection intensity was not identical across demes (model 2 and model 4), the parameter ω was decreasing linearly along the gradient, defined by: $\omega = 0.04$ ("very strong selection"; Rhoné et al. 2011) at the origin and $\omega = 100$ ("almost neutral case"; Rhoné et al. 2011) at the distal extremity of the gradient. In the scenarios where selection was defined as uniform across

populations (model 1 and model 3), we opted for an intermediate level ($\omega = 50.02$). In models assuming variable gene flow (model 3 and model 4), this latter was also set to decrease linearly along the gradient, with 90% of the hybFoo deme being sent to populations from the first area and 0% at the most distant one (these percentages are then rescaled to sum up to one as hybFoo must be emptied at each generation; see details in Supplementary Materials). In scenarios considering uniform gene flow, all demes received the same number of migrants from hybFoo ($1/35 = 2.86\%$ of the hybFoo deme). Finally, all demes received migrants from the resident species (pop1) at the same rate ($0.8/35 = 2.29\%$ of the pop1 deme in model 1 and model 2; $0.7/35 = 2\%$ in model 3 and model 4).

Simulations - biological replicates and sampling scheme

Our study also aims at examining whether including additional biological replicates improves the detection of adaptive loci. For each simulation (50 per model; same model settings, but different gene and marker input files), we ran five biological replicates (same model settings and identical gene and marker input files) (Fig. 1C). In the case where the outlier detection was based on more than one replicate, we considered as candidate loci the outliers that were shared across the replicates (i.e. the intersect of all the respective lists of outliers).

At the end of each of the thousand runs (4 models x 50 simulations x 5 replicates), 30 individuals per hybrid population (deme) were randomly sampled. An allele table was then constructed indicating the ancestry of each locus for every individual (0 or 1 for the homozygotes - resident and external species respectively - and 0.5 for the heterozygotes). The frequencies of the introgressed alleles were computed thereupon.

Performance assessment

We compared the performance of our AUC-based outlier detection methods to that of *pcadapt* at a fixed threshold value (0.90) for an increasing number of biological replicates. To this end, we investigated four different metrics: the true positive rate or sensitivity (selected genes correctly identified over all selected genes; *TPR = true positives/(true positives + false negatives)*), true negative rate or specificity (neutral loci correctly identified as non adaptive over all neutral loci; *TNR = true negatives/(true negatives + false positives)*), false positive rate (neutral loci mistakenly identified as adaptive over all neutral loci; *FPR = false positives/(false positives + true negatives)*) and false negative rate (selected genes erroneously identified as neutral over all selected genes; *FNR = false negatives/(false negatives + true positives)*). Next, we assessed the best performance of the different methods for an increasing number of replicates with two different, albeit similar, statistics: the F1 score (van Rijsbergen 1979) and Youden's J (Youden 1950). For this purpose, we iterated over quantile thresholds from 0 to 1 by increments of 0.01 (same threshold for all replicates), computing the performance statistic at each point and retaining the best (i.e. closest to one) performance value. The Youden's J is defined as *J = Sensitivity + Specificity - 1* (where *Sensitivity = true positives/(true positives + false negatives)* and *Specificity = true negatives/(true negatives + false positives)*). It can range between -1 (worst) and 1 (best). The F1 score is the harmonic mean of sensitivity and precision (i.e. *F1 = 2 * true positives/(2 * true positives + false positives + false negatives)*), taking values between zero (worst) and one (best). All statistical analyses were conducted on R version 3.2.0 (R Core Team 2016).

Results

Signature of selection on allele frequencies

Inspecting allele frequencies along the environmental gradient clearly reveals that the introgressed alleles occur at higher frequencies in areas where they are adaptive (Fig. 2). Moreover, the arguably most striking feature that emerges from this analysis is the upper position of the gene curves with respect to the marker curves (i.e. lower introgressed allele frequencies of the neutral loci). Finally, we observed that the signature of selection is well apparent under each simulated scenario.

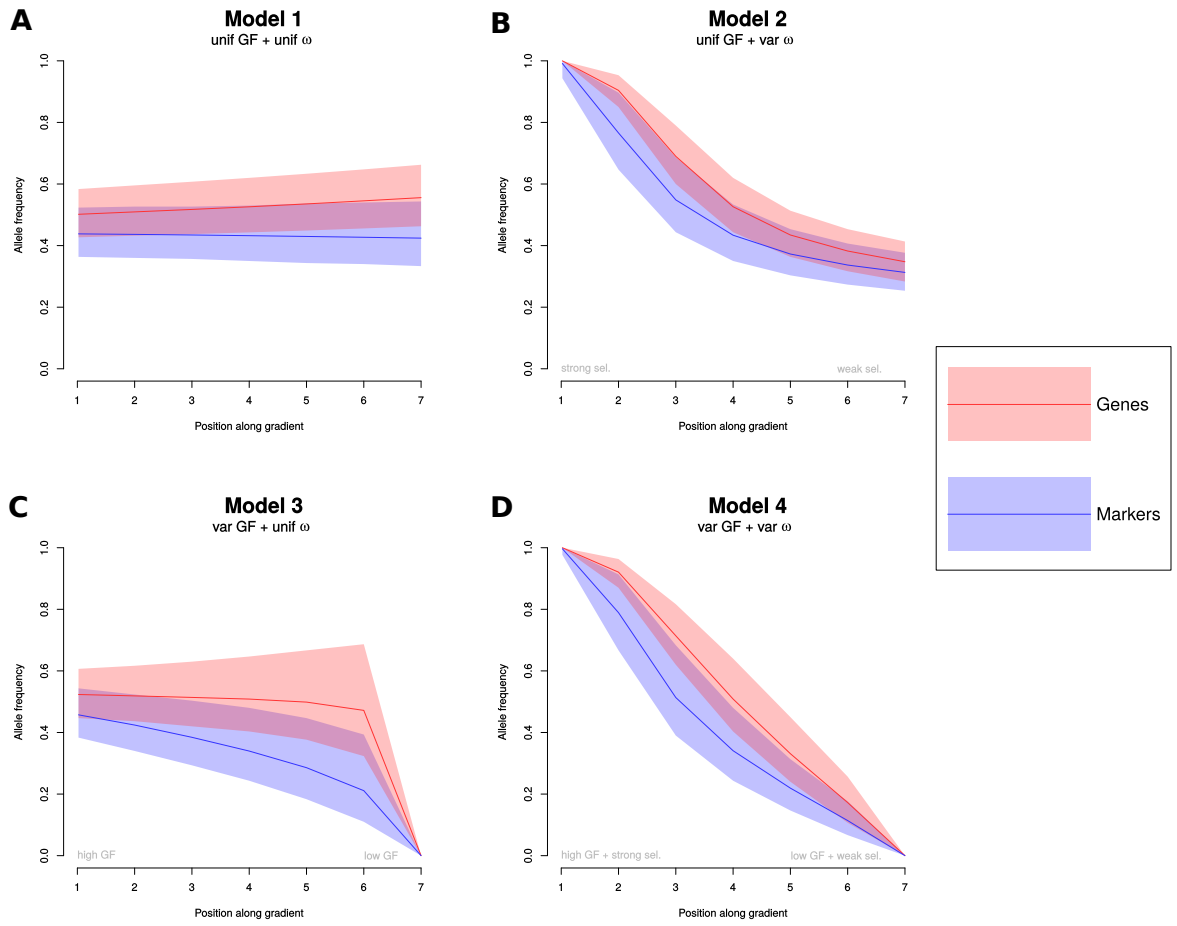

Figure 2: Signature of selection. We simulated 35 demes of hybrid populations arranged in seven geographical areas along an environmental gradient (see Figure 1). We then related the frequencies of introgressed alleles (y-axis) either for neutral markers (blue) or adaptive genes (red) as a function of the spatial location along the environmental gradient (x-axis). Four different scenarios were considered: **A)** model 1: uniform gene flow and uniform selection pressure (unif GF + unif ω); **B)** model 2: uniform gene flow and variable selection pressure (unif GF + var ω); **C)** model 3: variable gene flow and uniform selection pressure (var GF + unif ω); **D)** model 4: variable gene flow and variable selection pressure (var GF + var ω). For each of the four models are plotted the 0.95-quantiles of the frequencies of introgressed alleles for the neutral markers (blue area) and the positively selected genes (red area) along the environmental gradient. Solid lines indicate the mean values (blue for the neutral loci, red for the targeted genes).

Comparison and performance of outlier detection methods and the use of replicates

Next, we calculated for all simulations and for each locus the estimates of the AUC and the q-values derived from the *pcadapt* analysis. As computing the AUC using the integral of a GLM ("GLM AUC") or the trapezoid rule ("Smooth AUC") leads to highly similar results, we report only the "Smooth AUC" approach, which we believe is preferable as it does not stand on further modelling assumptions (results for the "GLM AUC" are given in Supplementary Materials; Fig. S4). Using four different metrics (FPR, FNR, TPR, TNR), we investigated the power and accuracy of the AUC-based methods and *pcadapt* at a given threshold (0.90), for an increasing number of biological replicates.

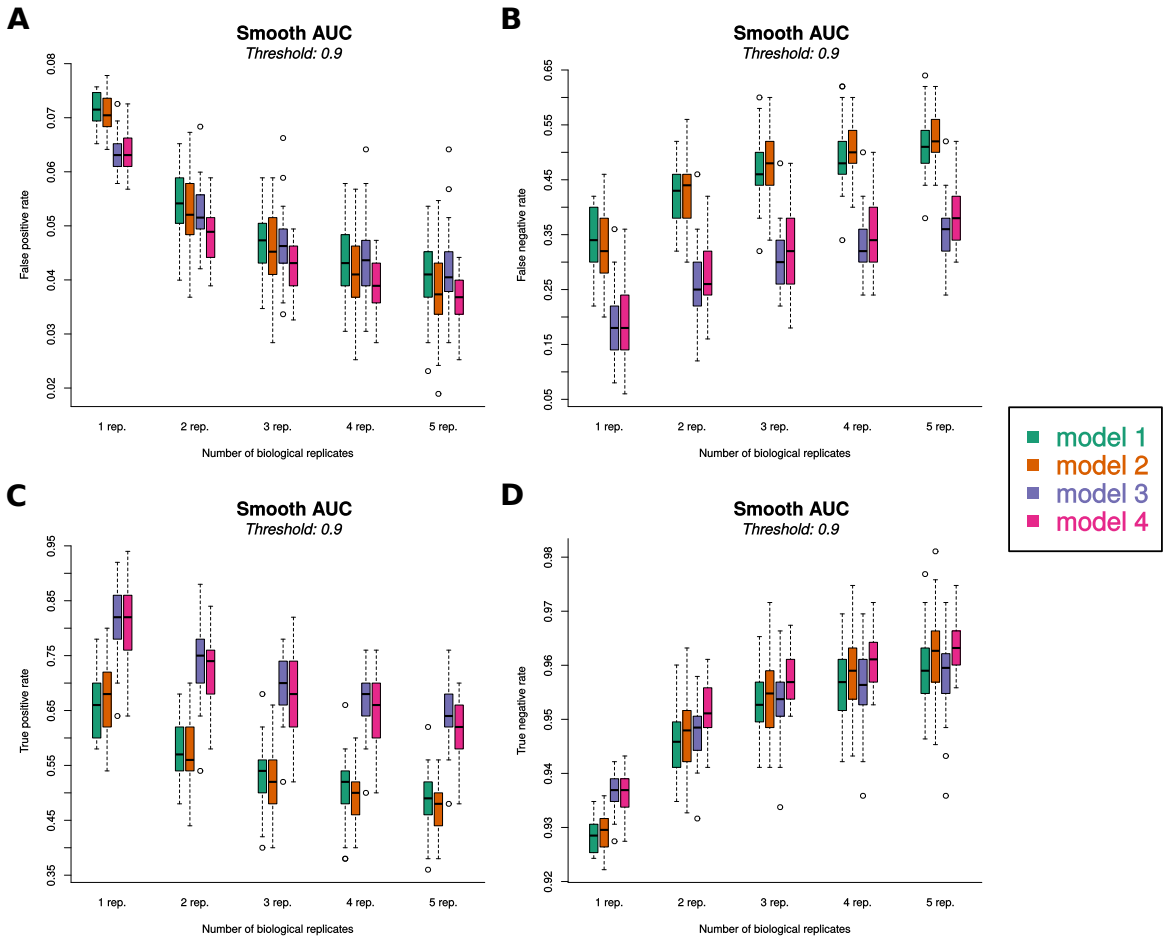

Figure 3: Power and accuracy of our AUC-based outlier detection method ("Smooth AUC"). 35 demes of hybrid populations arranged in seven classes of positions along an environmental gradient were simulated, under four different scenarios (refer to Figure 1 and 2 captions for model descriptions). Loci under positive selection were identified using the "Smooth AUC" method (as described in Material and methods). The performance of the outlier detection was assessed with four different metrics: **A)** false positive rate, **B)** false negative rate, **C)** true positive rate, **D)** true negative rate, computed by considering as candidate loci the outliers at the intersection of an increasing number of biological replicates (x-axis), for each of the four models simulated (50 simulations per model). The outlier detection threshold was set at the 0.90-quantile of the AUC estimate distribution.

By assessing the performance of the AUC-based method and that of *pcadapt* for an increasing number of biological replicates, we observed that the former outperforms the latter, and that adding biological replicates decreases the false positive rate. Indeed, from Figures 3 and 4 we see that the FPR decreases when more biological replicates are considered, but this diminution is less pronounced after adding a second replicate. Notably, the outlier detection with *pcadapt* is less accurate (higher FPR; Fig. 4A) than our AUC-based method (Fig. 3A). The TPR follows a similar decreasing trend, and the "Smooth AUC" (Fig. 3C) is again preferable since more sensitive (higher TPR) than *pcadapt* (Fig. 4C). Mirroring the true and false positive rates, we found an opposite increasing trend for the FNR and TNR. For the former, we noted that more genes are missed with *pcadapt* (higher FNR; Fig. 4B) than with "Smooth AUC" (Fig. 3B). For the latter, the difference between the two outlier detection methods is substantial when the analysis is based on a single biological replicate. In this case, "Smooth AUC" leads to a more specific identification of adaptive loci (higher TNR; Fig. 3D) than *pcadapt* (Fig. 4D). However, this difference tends to decrease with more replicates.

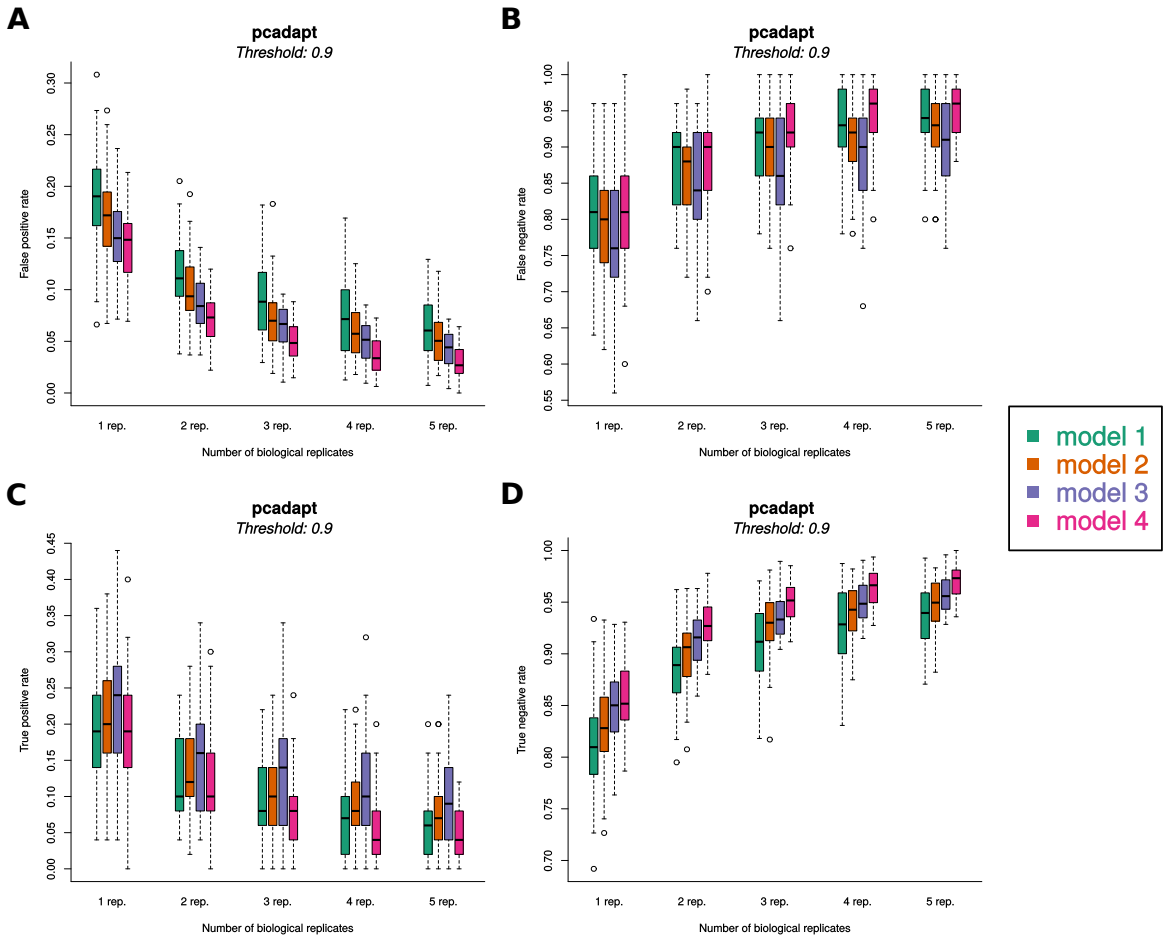

Figure 4: Power and accuracy of the *pcadapt* outlier detection method. The outlier detection threshold was set at a q-value of 0.10. Refer to Figure 3 caption for comprehensive figure description.

Interestingly, we also noted that *pcadapt* does not seem to be impacted by variation neither in gene flow nor in selection intensity, as the results of the four models are highly similar across the investigated scenarios (Fig. 4). By contrast, the models with variable gene flow (model 3 and model 4) facilitate the detection of outliers with the "Smooth AUC" approach, at least in terms of FNR and TPR (Fig. 3B/C). As for the FPR and TNR, adding biological replicates mitigates the effect of the model, as the departure between models 1/2 and 3/4 diminishes already after the second replicate (Fig. 3A/D).

Acknowledging that fixing the cut-off value to 0.90 was an arbitrary decision, we evaluated the performance of the "Smooth AUC" and *pcadapt* methods in their respective best case scenario. This latter was retrieved by iterating over quantile/q-value thresholds ranging from 0 to 1, and using the cut-off value that leads to the best performance (i.e. performance statistic closest to 1). As it is apparent from Figure 5, taking the AUC as a test statistic performs better than *pcadapt* (Fig. 5 for Youden's J; Fig. S3 for F1 score). Remarkably, better performance of outlier detection is achieved in model 3 and model 4, i.e. in simulations with variable gene flow (mean values for each model and replicate number are given in Tables S4/5). This observation is consistant with the greater overlap of gene and marker areas in the signature plots pictured in Figure 2. Notably, including additional replicates in the analysis improves the outlier detection with "Smooth AUC" (higher Youden's J; Fig. 5A), yet a plateau is readily reached after the second replicate. As for *pcadapt*, the inclusion of more replicates has no substantial impact on its performance, as the Youden's J oscillates roughly around zero in all cases.

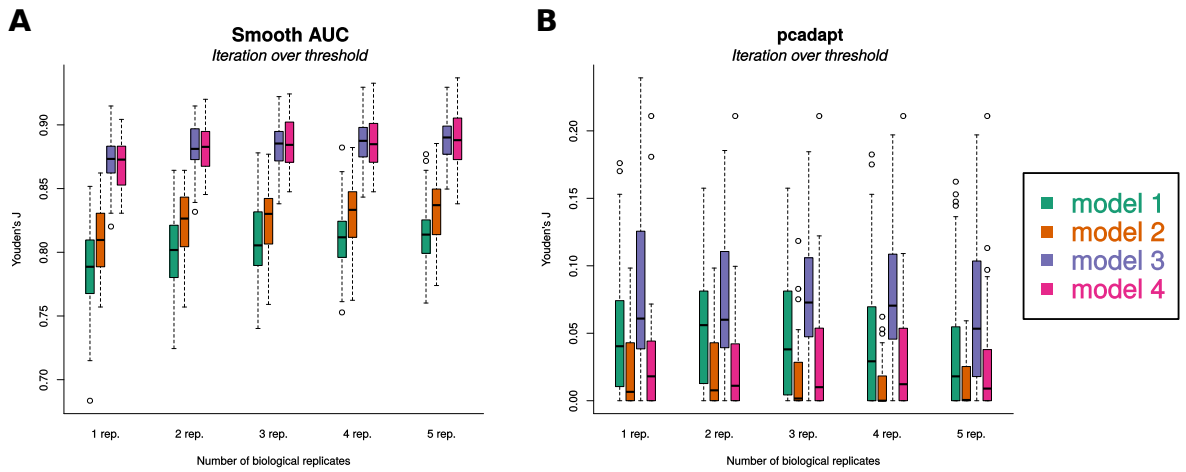

Figure 5: Performance of the outlier detection with increasing number of biological replicates. 35 demes of hybrid populations arranged in seven classes of positions along an environmental gradient were simulated, under four different scenarios (refer to Figure 1 and 2 captions for model descriptions). The performance of the outlier detection is assessed using the Youden's J statistic. Detection of adaptive loci was performed using **A)** our "Smooth AUC" method and **B)** *pcadapt*, as described in Material and methods. For the "Smooth AUC" method, the candidate loci are those for which the AUC is above a given threshold, defined as a quantile of the AUC estimates distribution. In the case of *pcadapt*, were considered as outliers the loci with a q-value smaller than α, where $\alpha = 1 - threshold$. When several biological replicates were included in the analysis (x-axis), candidate loci correspond to the outliers that are shared across the replicates (i.e. the intersect of all the respective lists of outliers). To overcome the arbitrary choice of the cut-off, "Smooth AUC" and *pcadapt* methods were evaluated in their respective best case scenario, by iterating over threshold values ranging from 0 to 1 and retaining the best (i.e. closest to one) Youden's J score.

Close-up on introgressed blocks to discard linked markers

If available, a reference genome enables the visual inspection of introgressed blocks and further improves the identification of adaptive genes. Indeed, we noted that after adding a third biological replicate to the analysis, the false positive rate stalls and the true positive rate does not drop below 0.45 for the "Smooth AUC" (Fig. 3A/C). This is an indication that neutral loci are erroneously - but consistently - identified as adaptive in all replicates. Furthermore, the pattern produced by plotting AUC estimates against distance to nearest gene (Fig. S5) suggests that those false positives are located in close proximity to the selected genes along the chromosome, presumably in linkage disequilibrium. Thus, we propose as a next step of our approach to investigate more closely blocks of adjacent outliers. We illustrate this stage with a worked example - loci of one chromosome of a simulation under the model 4 (var GF + var ω), taking the intersect of three biological replicates for the outlier detection. From Figures 6 and 7, it is apparent that outliers tend to cluster along the chromosome. Close-up of such introgressed blocks hence reveals that selected genes are often surrounded by the neutral markers identified as false positives. Especially, we noted that the frequencies of introgressed alleles of the focal loci, and, to a lesser extent, those of their closest neighbours are higher than that of the chromosomal background.

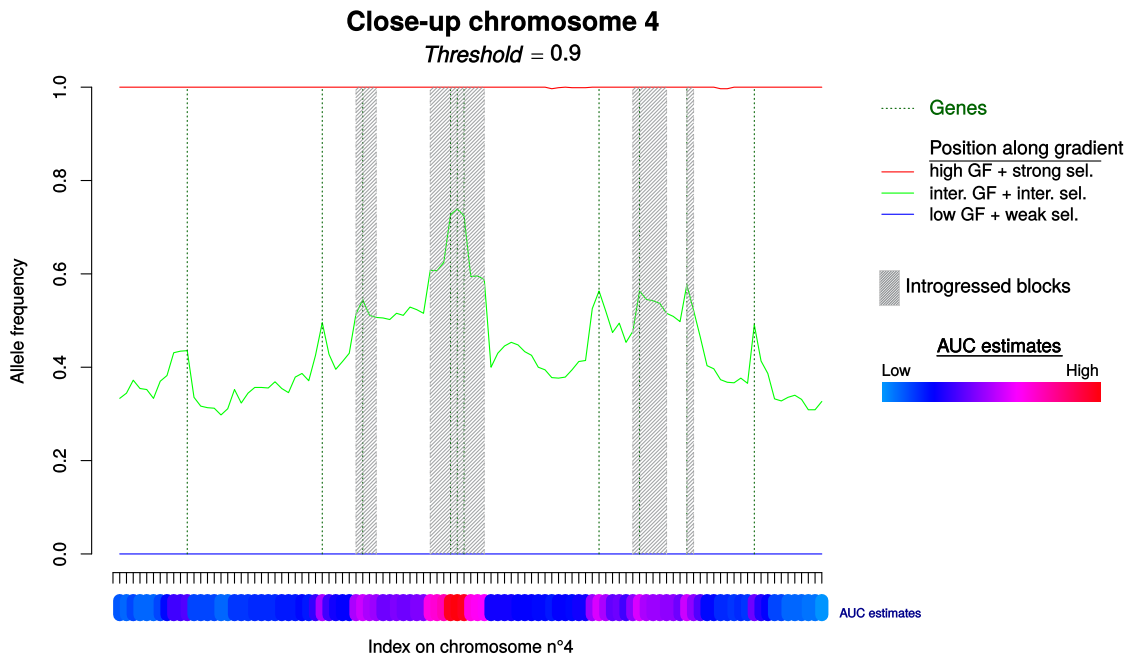

Figure 6: Worked example: close-up on a chromosome. Plot showing the variation in the frequencies of introgressed alleles (y-axis) along the fourth chromosome (discrete locus positions on the x-axis) for the model 4 (var ω + var GF), under strong (red curve), intermediate ("inter."; green curve) and low selection pressure ("sel.") and gene flow ("GF") (both inversely proportional to the position along the gradient). Green dashed lines indicate the presence of genes. Grey areas depict introgressed blocks (defined here as adjacent outliers, as detected at the 0.90-quantile threshold with the "Smooth AUC" method, for the intersection of three replicates of a single simulation). Under the x-axis, an indication of the AUC estimates as computed with the trapezoid rule ("Smooth AUC") for this simulation is given (colour gradient ranging from blue - low AUC values - to red - high AUC values).

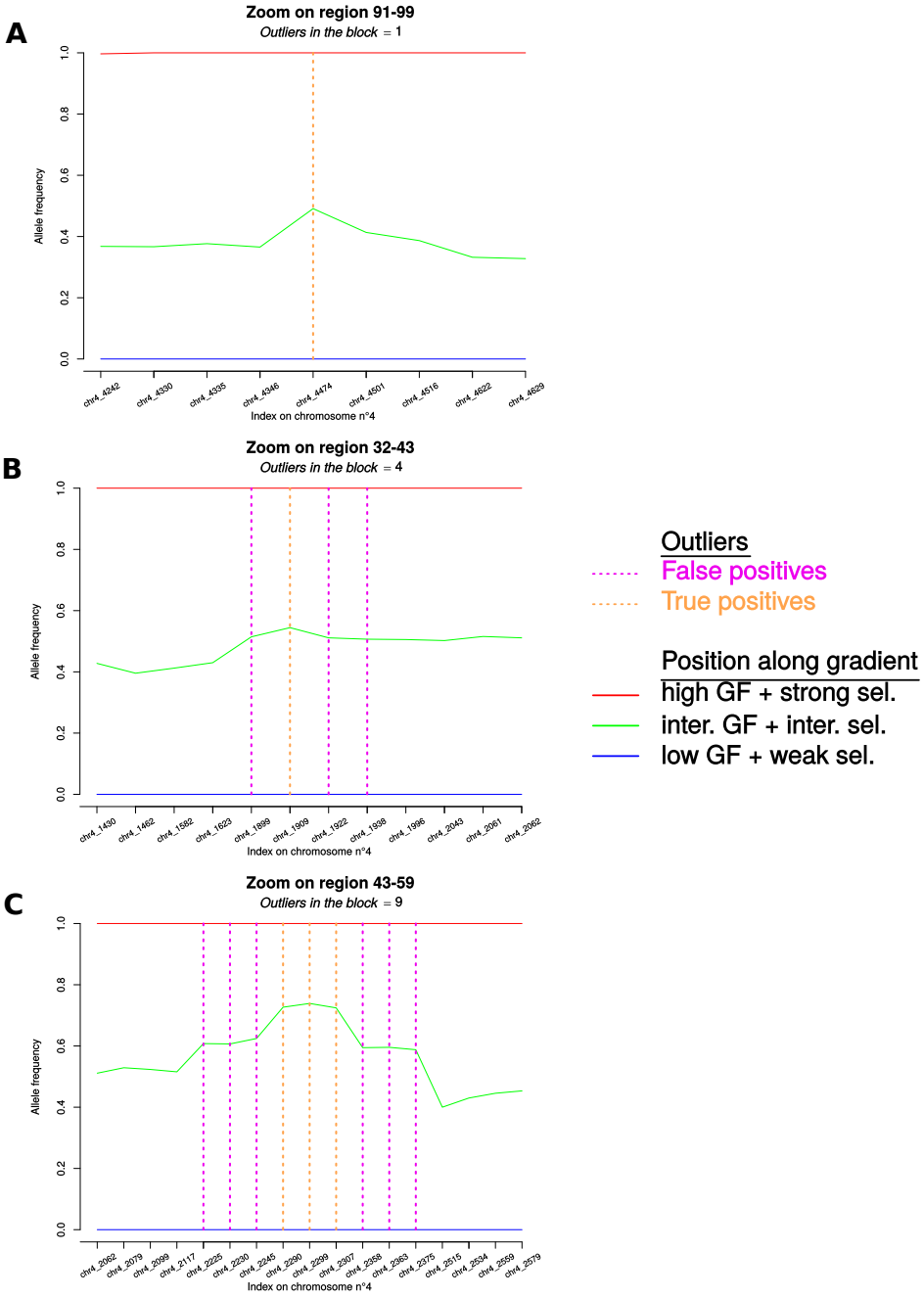

Figure 7: Worked example: close-up on introgressed blocks. Focus on some introgressed blocks of Figure 6 (cluster of adjacent outliers as detected at the 0.90-quantile threshold for the "Smooth AUC") of different size along the fourth chromosome (discrete locus positions on the x-axis). Horizontal curves represent the frequencies of introgressed alleles (y-axis) for the model 4 (var ω + var GF), under strong (red curve), intermediate ("inter."; green curve) and low selection pressure ("sel.") and gene flow ("GF") (inversely proportional to the position along the gradient). Candidate loci identified as outliers are either genes under selection that tend to occupy a central position in the introgressed blocks (true positives; orange vertical line), or surrounding neutral loci (false positives; magenta vertical line). **A)** One gene (true positive) isolated without neutral markers detected as outliers (false positives) in the vicinity. **B)** One gene (true positive) surrounded by three linked neutral markers identified as outliers (false positives). **C)** Three neighbouring genes (true positives) surrounded by six neutral markers mistakenly characterized as outliers (false positives).

Discussion

Here, we proposed and evaluated a strategy to detect adaptive loci in admixed populations. Our AUC-based approach is designed for clinal populations and takes advantage of the use of independent biological replicates. We showed that the signature of selection is well captured by frequencies of introgressed alleles regressed against an environmental gradient. Logically, we observed that adaptive alleles reached higher frequencies than the neutral ones. This is expected, except in cases where an adaptive locus is linked to detrimental genes or located in regions with low recombination rate (Ortíz-Barrientos et al. 2002; Kulathinal et al. 2009), but such situations were not considered here. Therefore, it appears that the area under the curve (AUC) metric might appropriately be used to discriminate adaptive from neutral loci.

At a quantile threshold of 0.90, our AUC-based approach leads to more satisfying results than *pcadapt* for all four benchmark metrics (FNR, FPR, TNR, TPR). In particular, this feature was striking with respect to sensitivity (FNR and TPR). To overcome the arbitrary choice of the quantile cut-off, we then assessed the performance of both methods with the Youden's J and F1 score by iterating over the quantile threshold values. The result of this procedure confirmed that the AUC-based method performance clearly outstrips that of *pcadapt*. The lower performance of *pcadapt* might be explained by the sampling scheme. Indeed, this software is not intended to take the environmental gradient directly into consideration. Nonetheless, it infers population structure on the basis of principal components, and the first of these tightly correlates with the environmental gradient in our simulations (Fig. S1). Conversely, the AUC measure integrates the positions along the gradient, which makes it particularly adequate for our simulated set-up. And although the AUC does not explicitly control for population structure, this does not seem to impact outlier detection. Incidentally, it has been previously highlighted that sampling along geographical transects allows to mitigate the confounding effects of the demography (Adrion et al. 2015). Another explanation for the discrepancy between the two outlier tests might lie in the way *pcadapt* detects outliers. In fact, these latter correspond to the markers "for which the vector of z-scores does not follow the distribution of the main bulk of points" (Luu et al. 2016), as assessed using the Mahalanobis distance. As in our simulations gene flow occurs in the same direction as selection pressure, it is possible that the distribution of the neutral loci does not sufficiently differ from that of those under selection to be discriminated by *pcadapt* (K. Luu, personal communication). To clarify this point, it would be informative to perform simulations for a null model where gene flow only occurs in the first generations.

At this point, based on the performance of outlier detection, our study might also provide some indications about the sampling design. Surprisingly, the lower false negative and higher true positive rates (Fig. 2C/D, 3B/C) suggest that selecting a field site where variable gene flow is at work could facilitate the detection of outliers with the AUC-based method. Indeed, this finding was not expected as it has been shown that asymmetrical gene flow can interfere with adaptation processes (Sexton et al. 2013) and impede the detection of adaptive loci (Manel et al. 2009). It should however be stated that this finding possibly arises from an artifact of our implementation of selection strength and gene flow rather than from biological processes. In fact, in the way we defined it, the gene flow under variable gene flow models (var GF) is higher than the gene flow under uniform gene flow models (unif GF) at closest positions, and lower at distant positions. As the selection pressure occurs in the same direction, this could exacerbate the effect of

selection, and facilitate the outlier detection (Fig. S6). Besides the choice of the statistical method *per se*, we also demonstrated that the use of independent biological replicates could decrease the false positive rate, which is recognized to be particularly high in genome scans for local adaptation (e.g. François et al. 2016). However, adding more than two additional replicates does not necessarily improve - or even worsen - the general performance of the analysis. This non-monotonic behaviour results from a trade-off between sensitivity (TPR) and specificity (TNR) and is apparent in Figures 3C/D and 4C/D. Combining results of independent experimental replicates has been suggested in previous simulation studies (e.g. Pérez-Figueroa et al. 2010). As for empirical studies focusing on local adaptation, the use of biological replication remains scarce in practice, but received some interest these last years (e.g. Perrier et al. 2013; Zulliger et al. 2013 ("cross-species" replication); Foll et al. 2014; Hand et al. 2016). For example, Berthouly-Salazar and co-workers (2016) used aridity gradients in two different countries to identify loci involved in climate adaptation in wild pearl millet. Previously, different studies focusing on *Arabis alpina* in the Alps (Poncet et al. 2010; Buehler 2013, 2014) have been conducted to study local adaptation in geographically independent populations. A narrow subset of outlier loci (four) were common to French and Swiss populations (Poncet et al. 2010), and a single candidate locus was detected among Swiss populations (Buehler et al. 2013) - but this latter could not be validated with an additional dataset (Buehler et al. 2014). According to the authors, failure of replication in this latter study might be caused by population structure specific to geographical location that interferes with outlier detection or the lack of convergent selection pressure. In general, from the aforementioned studies and in accordance with our simulations, it arises that cross-checking detected outliers enables to narrow down the set of adaptive candidates. However, it is also apparent that the intersect between replicates might be drastically shrunk, so that false negative rate possibly becomes unaffordably high in return. Such undesirable loss of sensitivity will inevitably occur if variance among replicates increases. Among others, such variance could arise from different histories of the hybrid zones (timing of the hybridization events, demographic and environmental contexts, etc.). Accordingly, our simulated biological replicates share identical and time invariant population and genetic structures (exactly identical gene flow, selection strength, number of generations, etc.), which is not necessarily realistic. In addition, we made the strong assumption that the same genes underly the adaptive introgression of a given phenotype in all biological replicates. However, parallel genetic bases are conceivable *in natura* (Elmer and Meyer 2011; Yeaman et al. 2016). In the end, the number of replicates to consider for taking the intersection of candidate loci will depend on the aim of the study, and also on the similarity, hence comparability, of the sampled transects.

Finally, the last step of our strategy consists in closing up on introgressed regions to investigate linkage relationship among loci. Consistent with a previous study (Vilas et al. 2012), we showed that neutral markers erroneously detected as adaptive tended to cluster around true positive loci. As the positive selection on the gene influences the evolutionary trajectory of the neighbouring loci with which it is in linkage disequilibrium (Hill and Robertson 1966; Comeron et al. 2008), genome scans for adaptive loci can hardly get rid of those false positives (e.g. McVean 2007; Hohenlohe et al. 2010; Pardo-Diaz et al. 2015). In fact, the chromosomal proximity imposes a physical constraint that in our case only recombination, which breaks association between alleles, could eliminate (either through more generations or higher recombination rate; e.g. Meuwissen and Goddard 2000). Notably, the limited number of outlier loci identified in real-life experiments (according to the aforementioned field studies) makes it possible to perform a visual inspection

of the introgressed blocks to distinguish selected genes (local maximum, central position) from their neutral hitchhikers.

Clearly, our work is not free from criticisms. First of all, for reasons of time, we restricted our comparison to *pcadapt*. It would also have been worth including other outlier detection methods recently developed, for example those built upon the framework of genome-wide association studies (e.g. EigenGWAS of Chen et al. 2016) or that are based on phylogenetic models (e.g. Liu et al. 2014 or Hejase and Liu 2016). In particular, a method that retrieves association with environmental gradient - such as LFMM (Frichot et al. 2013) or BayeScEnv (de Villemereuil and Gaggiotti 2015) - might be more comparable with our AUC-based method, that captures environmental variation along a gradient.

Next, our simulations suffer from substantial limitations and further work will be required to investigate scenarios with more complex genetic and demographic structures. Among others, we assumed that all genes equally contribute to the individual fitness, in the absence of epistatic interactions or pleiotropy. However, more complex genetic architectures can be expected (e.g. Carlborg and Haley 2004; Holland 2007; Shao et al. 2008; Taylor and Ehrenreich 2015). Also, a possible effect of the genomic background of the recipient genome was disregarded, thus overlooking potential endogenous genetic barriers (Bierne et al. 2011). Furthermore, we assumed that both parental species harbor the same number of chromosomes, each of them undergoing recombination events at identical frequency. Then, the gradient was considered as an appropriate proxy for the fitness, thereby neglecting landscape heterogeneity. Moreover, migration between hybrid populations was excluded from our scenarios, and migration rates were constant, as were the carrying capacities. Finally, a further shortcoming of our work lies in the fact that the outlier detection method we propose is a rank-based procedure. This implies that "outliers" would be identified, even in the total absence of selection.

In conclusion, we presented in this article an approach for the detection of outlier loci in the context of adaptive introgression along clinal gradients and validated it with simulated data. Remarkably, this three-step procedure relies on only two requirements nowadays commonly available: genome-wide data and marker chromosomal positions. Obviously, for our strategy to be efficient, replication sites should be carefully selected with respect of the environmental gradient investigated. An opening question that requires additional research is the automation of the choice of an appropriate quantile distribution threshold for the definition of outlierness. Eventually, an interesting outlook will be to deploy the AUC approach on field datasets.

Acknowledgments

The authors thank Jérôme Goudet for advice and fruitful discussions. We are also grateful to Rongfeng Cui and Keurcien Luu for providing assistance for data simulation and analysis. This research is funded by an SNSF Ambizione research grant (PZ00P3_148224) to NA.

References

Adrion, J. R., M. W. Hahn, and B. S Cooper (2015). "Revisiting classic clines in *Drosophila melanogaster* in the age of genomics". *Trends in genetics* 31 (8), pp. 434–444.

Baack, E. J. and L. H. Rieseberg (2007). "A genomic view of introgression and hybrid speciation". *Current Opinion in Genetics and Development* 17 (6), pp. 513–518.

Barrett, R. D. H. and D. Schluter (2007). "Adaptation from standing genetic variation". *TRENDS in Ecology & Evolution* 23 (1), pp. 38–44.

Barton, N. H. (2001). "The role of hybridization in evolution". *Molecular Ecology* 10 (3), pp. 551–568.

Beaumont, M. A. and R. A. Nichols (1996). "Evaluating loci for use in the genetic analysis of population structure". *Proceedings of the Royal Society B* 263 (1377), pp. 1619–1626.

Berthouly-Salazar, C. et al. (2016). "Genome scan reveals selection acting on genes linked to stress response in wild pearl millet". *Molecular Ecology* 25 (21), pp. 5500–5512.

Bierne, N., J. Welch, E. Loire, F. Bonhomme, and P. David (2011). "The coupling hypothesis: why genome scans may fail to map local adaptation genes". *Molecular Ecology* 20 (10), pp. 2044–2072.

Buehler, D., R. Holderegger, S. Brodbeck, E. Schnyder, and F. Gugerli (2014). "Validation of outlier loci through replication in independent data sets: a test on *Arabis alpina*". *Ecology and Evolution* 4 (22), pp. 4296–4306.

Buehler, D., B. N. Poncet, R. Holderegger, S. Manel, P. Taberlet, and F. Gugerli (2013). "An outlier locus relevant in habitat-mediated selection in an alpine plant across independent regional replicates". *Evolutionary Ecology* 27 (2), pp. 285–300.

Carlborg, O. and C. S. Haley (2004). "Epistasis: too often neglected in complex trait studies?" *Nature Review Genetics* 5 (8), pp. 618–625.

Chen, G.-B., S. H. Lee, Z.-X. Zhu, and B. Benyamin (2016). "EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations". *Heredity* 117 (1), pp. 51–61.

Comeron, J. M., A. Williford, and R. M. Kliman (2008). "The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations". *Heredity* 100 (1), pp. 19–31.

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard (2010). "Using environmental correlations to identify loci underlying local adaptation". *Genetics* 185 (4), pp. 1411–1423.

Crawford, J. E. and R. Nielsen (2013). "Detecting adaptive trait loci in nonmodel systems: divergence or admixture mapping?" *Molecular Ecology* 22 (24), pp. 6131–6148.

Cronn, R. and J. F. Wendel (2003). "Cryptic trysts, genomic mergers, and plant speciation". *New Phytologist* 161, pp. 133–142.

Cui, R., M. Schumer, and G. G. Rosenthal (2016). "Admix'em : a flexible framework for forward-time simulations of hybrid populations with selection and mate choice". *Bioinformatics* 32 (7), pp. 1103–1105.

Darvasi, A. and S. Shifman (2005). "The beauty of admixture." *Nature genetics* 37 (2), pp. 118–119.

De Mita, S. et al. (2013). "Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations". *Molecular Ecology* 22 (5), pp. 1383–1399.

De Villemereuil, P., E. Frichot, E. 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 (8), pp. 2006–2019.

De Villemereuil, P. and O. E. Gaggiotti (2015). "A new F_{ST}-based method to uncover local adaptation using environmental variables". *Methods in Ecology and Evolution* 6 (11), pp. 1248–1258.

Earley, E. J. and C. D. Jones (2011). "Next-generation mapping of complex traits with phenotype-based selection and introgression". *Genetics* 189 (4), pp. 1203–1209.

Elmer, K. R. and A. Meyer (2011). "Adaptation in the age of ecological genomics: insights from parallelism and convergence". *Trends in Ecology and Evolution* 26, pp. 298–306.

Excoffier, L. and H. E. L. Lischer (2010). "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows". *Molecular Ecology Resources* 10 (3), pp. 564–567.

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 (2), pp. 977–993.

Foll, M., O. E. Gaggiotti, J. T. Daub, A. Vatsiou, and L. Excoffier (2014). "Widespread signals of convergent adaptation to high altitude in Asia and America". *The American Journal of Human Genetics* 95 (4), pp. 394–407.

François, O., H. Martins, K. Caye, and S. D. Schoville (2016). "Controlling false discoveries in genome scans for selection". *Molecular Ecology* 25 (2), pp. 454–469.

Frichot, E., S. D. Schoville, G. Bouchard, and O. François (2013). "Testing for associations between loci and environmental gradients using latent factor mixed models". *Molecular Biology and Evolution* 30 (7), pp. 1687–1699.

Hamilton, J. A., C. Lexer, and S. N. Aitken (2013). "Differential introgression reveals candidate genes for selection across a spruce (*Picea sitchensis* × *P. glauca*) hybrid zone". *New Phytologist* 197 (3), pp. 927–938.

Hand, B. K. et al. (2016). "Climate variables explain neutral and adaptive variation within salmonid metapopulations: the importance of replication in landscape genetics". *Molecular Ecology* 25 (3), pp. 689–705.

Hedrick, P. W. (2013). "Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation". *Molecular Ecology* 22 (18), pp. 4606–4618.

Hejase, H. A. and K. J. Liu (2016). "Mapping the genomic architecture of adaptive traits with interspecific introgressive origin: a coalescent-based approach". *BMC Genomics* 17 (8), pp. 1–17.

Hill, W. G. and A. Robertson (1966). "The effect of linkage on limits to artificial selection". *Genetics research* 8 (2), pp. 269–294.

Hohenlohe, P. A., P. C. Phillips, and W. A. Cresko (2010). "Using population genomics to detect selection in natural populations: key concepts and methodological considerations". *International Journal of Plant Sciences* 171 (9), pp. 1059–1071.

Holland, J. B. (2007). "Genetic architecture of complex traits in plants". *Current Opinion in Plant Biology* 10 (2), pp. 156–161.

Joly, S., P. A. McLenachan, and P. J. Lockhart (2009). "A statistical approach for distinguishing hybridization and incomplete lineage sorting". *The American Naturalist* 174 (2), E54–E70.

Jones, M. R. et al. (2013). "Integrating landscape genomics and spatially explicit approaches to detect loci under selection in clinal populations". *Evolution* 67 (12), pp. 3455–3468.

Jurasinski, G., K. Franziska, G. Anke, and B. Sascha (2014). *flux: flux rate calculation from dynamic closed chamber measurements*. R package version 0.3-0.

Kawecki, T. J. and D. Ebert (2004). "Conceptual issues in local adaptation". *Ecology Letters* 7 (12), pp. 1225–1241.

Kim, M. et al. (2008). "Regulatory genes control a key morphological and ecological trait transferred between species". *Science* 322 (5904), pp. 1116–1119.

Kulathinal, R. J., L. S. Stevison, and M. A. F. Noor (2009). "The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing". *PLoS Genetics* 5 (7), pp. 1–7.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush (2012). "Inference of population structure using dense haplotype data". *PLoS Genetics* 8 (1), pp. 11–17.

Lexer, C., B. Heinze, R. Alia, and L. H. Rieseberg (2004). "Hybrid zones as a tool for identifying adaptive genetic variation in outbreeding forest trees: lessons from wild annual sunflowers (*Helianthus* spp.)" *Forest Ecology and Management* 197 (1-3), pp. 49–64.

Liu, K. J., J. Dai, K. Truong, Y. Song, M. H. Kohn, and L. Nakhleh (2014). "An HMM-based comparative genomic framework for detecting introgression in Eukaryotes". *PLoS Computational Biology* 10 (6), pp. 1–13.

Lotterhos, K. E. and M. C. Whitlock (2015). "The relative power of genome scans to detect local adaptation depends on sampling design and statistical method". *Molecular Ecology* 24 (5), pp. 1031–1046.

Luu, K., E. Bazin, and M. G. B. Blum (2016). "*pcadapt*: an R package to perform genome scans for selection based on principal component analysis". *Molecular Ecology Resources* 33, pp. 1–11.

Lynch, M. (2007). "The evolution of genetic networks by non-adaptive processes". *Nature Reviews Genetics* 8 (10), pp. 803–813.

Mallet, J., N. Besansky, and M. W. Hahn (2015). "How reticulated are species?" *BioEssays* 38 (2), pp. 140–149.

Manel, S., C. Conord, and L. Després (2009). "Genome scan to assess the respective role of host-plant and environmental constraints on the adaptation of a widespread insect". *BMC Evolutionary Biology* 9 (288), pp. 1–10.

Maronna, R. A. and R. H. Zamar (2002). "Robust estimates of location and dispersion for high-dimensional datasets". *Technometrics* 44 (4), pp. 307–317.

Martin, N. H., A. C. Bouck, and M. L. Arnold (2006). "Detecting adaptive trait introgression between *Iris fulva* and *I. brevicaulis* in highly selective field conditions". *Genetics* 172 (4), pp. 2481–2489.

McKeigue, P. M. (2005). "Prospects for admixture mapping of complex traits". *American Journal of Human Genetics* 76 (1), pp. 1–7.

McVean, G. (2007). "The structure of linkage disequilibrium around a selective sweep". *Genetics* 175 (3), pp. 1395–1406.

Meuwissen, T. H. E. and M. E. Goddard (2000). "Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci". *Genetics* 155 (1), pp. 421–430.

Nadeau, N. J. et al. (2012). "Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing". *Philosophical Transactions of the Royal Society B*, pp. 343–353.

Narum, S. R. and J. E. Hess (2011). "Comparison of F_{ST} outlier tests for SNP loci under selection". *Molecular Ecology Resources* 11 (Suppl. 1), pp. 184–194.

Ortíz-Barrientos, D., J. Reiland, J. Hey, and M. A. F. Noor (2002). "Recombination and the divergence of hybridizing species". *Genetica* 116 (2), pp. 167–178.

Pardo-Diaz, C., C. Salazar, and C. D. Jiggins (2015). "Towards the identification of the loci of adaptive evolution". *Methods in Ecology and Evolution* 6 (4), pp. 445–464.

Pérez-Figueroa, A., M. J. García-Pereira, M. Saura, E. Rolán-Alvarez, and A. Caballero (2010). "Comparing three different methods to detect selective loci using dominant markers". *Journal of Evolutionary Biology* 23 (10), pp. 2267–2276.

Perrier, C., V. Bourret, M. P. Kent, and L. Bernatchez (2013). "Parallel and nonparallel genome-wide divergence among replicate population pairs of freshwater and anadromous Atlantic salmon". *Molecular Ecology* 22 (22), pp. 5577–5593.

Poncet, B. N. et al. (2010). "Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*". *Molecular Ecology* 19 (14), pp. 2896–2907.

R Core Team (2016). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rhoné, B., J.-T. Brandenburg, and F. Austerlitz (2011). "Impact of selection on genes involved in regulatory network: a modelling study". *Journal of Evolutionary Biology* 24 (10), pp. 2087–2098.

Rieseberg, L. H. (2009). "Evolution: replacing genes and traits through hybridization". *Current Biology* 19 (3), pp. 119–122.

Rosenzweig, B. K., J. B. Pease, N. J. Besansky, and M. W. Hahn (2016). "Powerful methods for detecting introgressed regions from population genomic data". *Molecular Ecology* 25 (11), pp. 2387–2397.

Scascitelli, M., K. D. Whitney, R. A. Randell, M. King, C. A. Buerkle, and L. H. Rieseberg (2010). "Genome scan of hybridizing sunflowers from Texas (*Helianthus annuus* and *H. debilis*) reveals asymmetric patterns of introgression and small islands of genomic differentiation". *Molecular Ecology* 19 (3), pp. 521–541.

Seldin, M. F. (2007). "Admixture mapping as a tool in gene discovery". *Current Opinion in Genetics and Development* 17 (3), pp. 177–181.

Sexton, J. P., S. B. Hangartner, and A. A. Hoffmann (2013). "Genetic isolation by environment or distance: which pattern of gene flow is most common?" *Evolution* 68 (1), pp. 1–15.

Shao, H. et al. (2008). "Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis." *PNAS* 105 (50), pp. 19910–19914.

Smith, M. W. and S. J. O'Brien (2005). "Mapping by admixture linkage disequilibrium: advances, limitations and guidelines". *Nature Reviews Genetics* 6 (8), pp. 623–632.

Storey, J. D (2015). *qvalue: q-value estimation for false discovery rate control*. R package version 2.2.2.

Taylor, M. B. and I. M. Ehrenreich (2015). "Higher-order genetic interactions and their contribution to complex traits". *Trends in Genetics* 31 (1), pp. 34–40.

Tigano, A. and V. L. Friesen (2016). "Genomics of local adaptation with gene flow". *Molecular Ecology* 25 (10), pp. 2144–2164.

Twyford, A. D. and R. A. Ennos (2012). "Next-generation hybridization and introgression". *Heredity* 108 (3), pp. 179–189.

van Rijsbergen, C. (1979). *Information Retrieval*. ButterworthHeinemann: Boston, 2nd edition. 224 p.

Verity, R., C. Collins, D. C. Card, S. M. Schaal, L. Wang, and K. E. Lotterhos (2016). "MINOTAUR : a platform for the analysis and visualization of multivariate results from genome scans with R Shiny". *Molecular Ecology Resources* 17 (1).

Vilas, A., A. Pérez-Figueroa, and A. Caballero (2012). "A simulation study on the performance of differentiation-based methods to detect selected loci using linked neutral markers". *Journal of Evolutionary Biology* 25 (7), pp. 1364–1376.

Vitalis, R., K. Dawson, P. Boursot, and K. Belkhir (2003). "DetSel 1.0: a computer program to detect markers responding to selection". *Journal of Heredity* 94 (5), pp. 429–431.

Ward, B. J. and C. van Oosterhout (2016). "HYBRIDCHECK: software for the rapid detection, visualization and dating of recombinant regions in genome sequence data". *Molecular Ecology Resources* 16 (2), pp. 534–539.

Whitney, K. D., R. A. Randell, and L. H. Rieseberg (2010). "Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*". *New Phytologist* 187 (1), pp. 230–239.

Yeaman, S. et al. (2016). "Convergent local adaptation to climate in distantly related conifers". *Science* 353 (6306), pp. 23–26.

Youden, W. J. (1950). "Index for rating diagnostic tests". *Cancer* 3 (1), pp. 32–35.

Zulliger, D., E. Schnyder, and F. Gugerli (2013). "Are adaptive loci transferable across genomes of related species? Outlier and environmental association analyses in Alpine Brassicaceae species". *Molecular Ecology* 22 (6), pp. 1626–1639.

Supplementary Materials

A strategy to detect adaptive introgression in clinal populations

Marie Zufferey, Nils Arrigo

Correlation between the positions along the gradient and the coordinates on the first principal component

The software *pcadapt* does not directly use the information of the environmental gradient. In fact, it rather infers population structure on the basis of the eigenaxes of the principal component analysis (Luu et al. 2016). However, from Figure S1, it is clearly apparent that the first principal component (PC1) reflects the position along the gradient, as the coordinates along PC1 strongly correlate with the positions along the gradient.

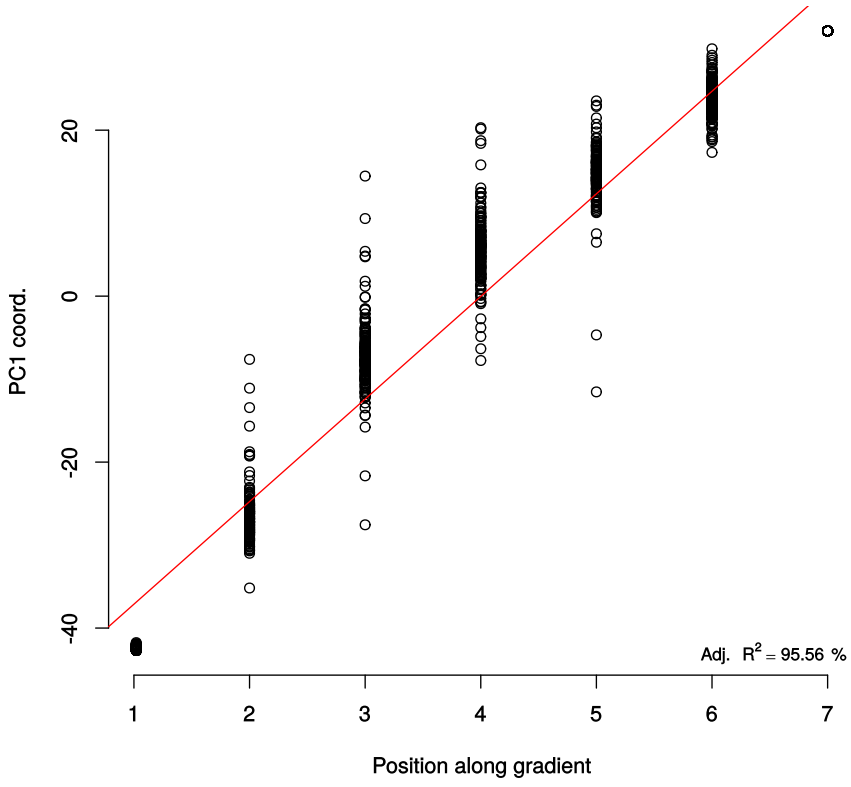

Figure S1: Correlation between PC1 coordinates and positions along the gradient. The positions along the gradient (x-axis) strongly correlate with the coordinates along the first principal component (y-axis).

Example of scree plot obtained with *pcadapt*

For the analysis with *pcadapt*, the user has to define the number of eigenaxes to retain for the outlier detection. This can be done using a scree plot (as explained in the vignette of the R package; Luu et al. 2016). Such plot indicates how much variance is explained by each principal component. An example of such plot is given in Figure S2, for one simulation under the model 4 (var GF + var ω; similar results are obtained for the other models). Normally, it should be retained as many axes they are on the left of the almost-straight line (i.e. the principal components corresponding to the steep curve). As we could not check this scree plot for all runs of Admix'em, we chose $K = 10$ to ensure retaining enough principal axes. In fact, as illustrated by Figure S2, the line is almost straight already after the second principal component.

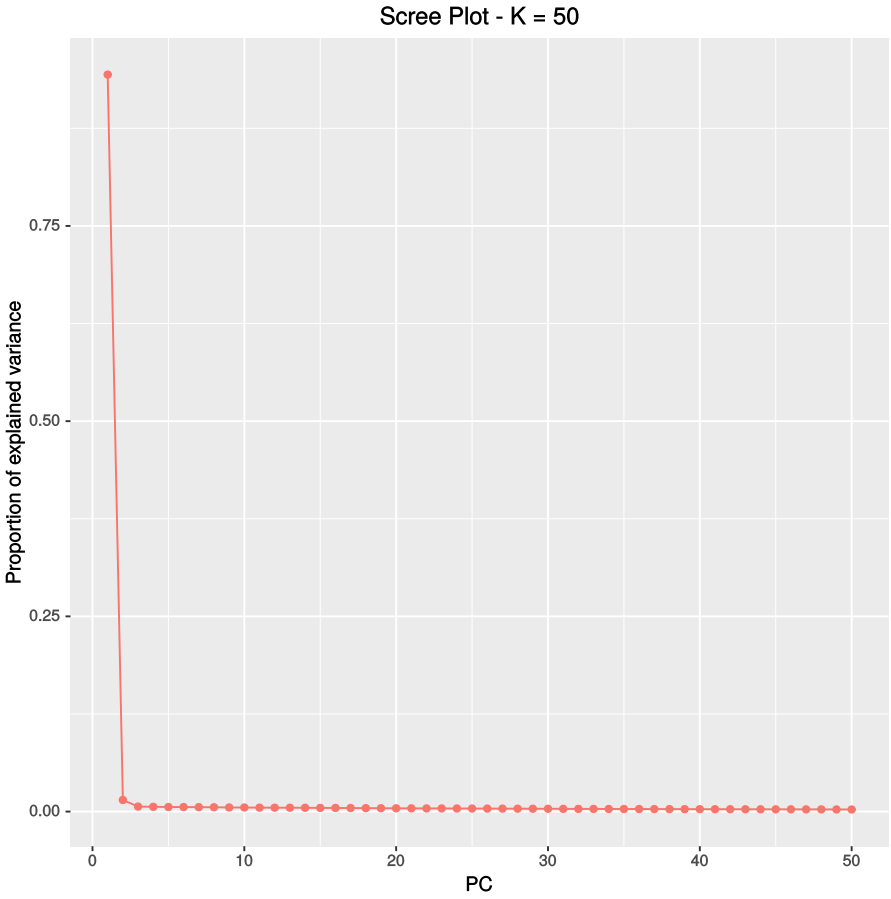

Figure S2: Scree plot obtained with *pcadapt*. This scree plot was obtained from the data of one simulation under the model 4 (var GF + var ω). The y-axis represents the variance explained by each principal component (x-axis). The drop in variance explained is already noticeable after the first principal component.

Details about Admix'em settings

Settings for Admix'em (Cui et al. 2016) have been thoroughly described in the main text. Some additional technical details are provided here below (Tables S1/2/3; see captions for the explanations).

Parameter	*Value*
Nbr. males sampled by female	50
Avg. female gametes	10
Nbr. of generations	200
Carrying capacity hyb. pop.	1000
Carrying capacity pop1 and pop2	1000
Carrying capacity hybFoo	500
pop1 initial size	500
pop2 initial size	500
hybFoo → hyb. pop. (unif GF)	0.0286
hybFoo → hyb. pop. (var GF, P1)	0.0570
hybFoo → hyb. pop. (var GF, P2)	0.0477
hybFoo → hyb. pop. (var GF, P3)	0.0381
hybFoo → hyb. pop. (var GF, P4)	0.0286
hybFoo → hyb. pop. (var GF, P5)	0.0191
hybFoo → hyb. pop. (var GF, P6)	0.0095
hybFoo → hyb. pop. (var GF, P7)	0.00
pop1 → hyb. pop. (unif GF)	0.0229
pop1 → hyb. pop. (var GF)	0.0200
Function kids per female	Poisson
pop1 male ratio	0.5
pop1/pop2 → hybFoo	0.1

Table S1: Settings of the configuration file for Admix'em. Some details of the configuration file required for running Admix'em. Migration is represented by the → symbol. In the case where gene flow was uniform, the migration from hybFoo was the same for the 35 demes. Conversely, when it was set to be variable, it was decreasing along the gradient (P1 represents the closest position, and P7 the position at the distal extremity of the gradient). Abbreviations: Nbr. = number, Avg. = average, hyb. pop. = hybrid populations (correspond to the 35 demes described in the main text), pop1 = resident species, pop2 = external species, hybFoo = hybrid pool.

Phenotypes	*Formula*
Sex	chr1_68
Pheno0	if(chr2_2177==2,1,if(chr2_2177==1,0.5,if(chr2_2177==0,0,0)))
Pheno1	if(chr2_3223==2,1,if(chr2_3223==1,0.5,if(chr2_3223==0,0,0)))
Pheno2	if(chr2_3286==2,1,if(chr2_3286==1,0.5,if(chr2_3286==0,0,0)))
...	...
Pheno49	if(chr10_4772==2,1,if(chr10_4772==1,0.5,if(chr10_4772==0,0,0)))

Table S2: Phenotype configuration file for Admix'em. This file allows to define the 50 alleles under positive selection. In addition, Admix'em requires to select arbitrarily a gene for sex determination (not under selection in our study). All loci were randomly selected.

Population	*Gen*	*Selection*
pop1	-1	1
pop2	-1	1
hybFoo	-1	1
hyb1	-1	exp(-pow(Pheno0+Pheno1+...+Pheno49-50, 2)/pow(0.04, 2))
...	...	...
hyb7e	-1	exp(-pow(Pheno0+Pheno1+...+Pheno49-50, 2)/pow(100, 2))

Table S3: Natural selection for Admix'em. This file allows to define natural selection. The fitness value is given by the function explained in the main text. The adaptive trait is determined by the additive effect of the 50 genes under selection.

Performance assessed with F1 score

In addition to the Youden's J, we also used the F1 score to assess the performance of the adaptive locus detection with the different methods (Fig. S3). Both metrics lead to similar results (see main text for the result discussion).

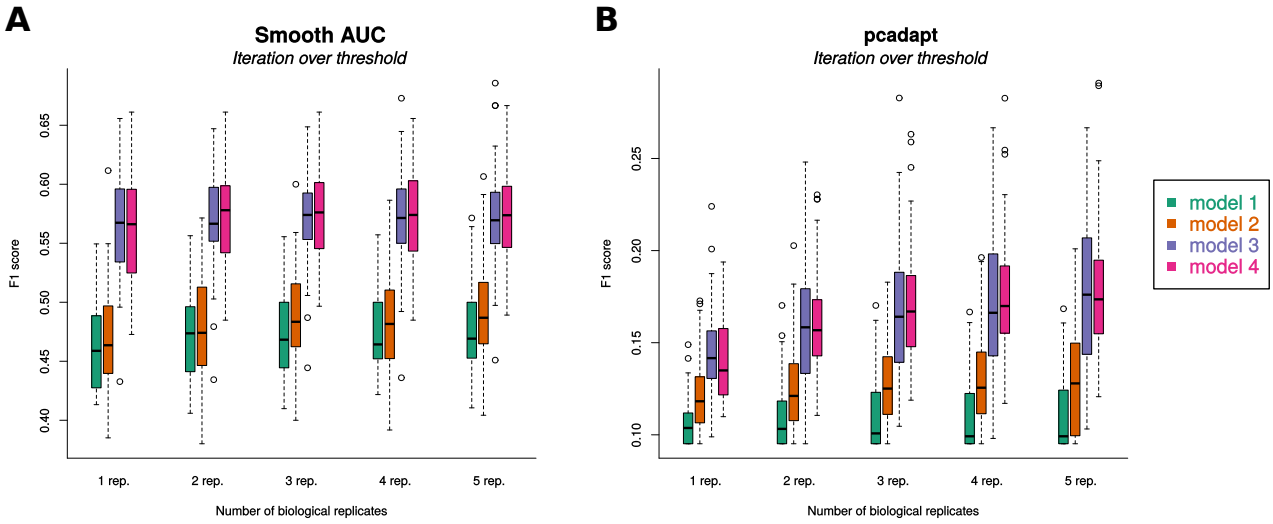

Figure S3: Performance of the outlier detection with increasing number of biological replicates. 35 demes of hybrid populations arranged in seven classes of positions along an environmental gradient were simulated under four different scenarios (model 1: unif GF + unif ω; model 2: unif GF + var ω; model 3: var GF + var ω; model 4: var GF + var ω; see main text for comprehensive model descriptions). The performance of the outlier detection is assessed using the F1 score. Detection of adaptive loci was performed using: **A)** our "Smooth AUC" method and **B)** *pcadapt*, as described in Material and methods section of the main text. For the "Smooth AUC" method, the candidate loci are those for which the AUC is above a given threshold, defined as a quantile of the AUC estimates distribution. In the case of *pcadapt*, were considered as outliers the loci with a q-value smaller than α, where $\alpha = 1 - threshold$. When several biological replicates were included in the analysis (x-axis), candidate loci correspond to the outliers that are shared across the replicates (i.e. the intersect of all the respective lists of outliers). To overcome the arbitrary choice of the cut-off, "Smooth AUC" and *pcadapt* methods were evaluated in their respective best case scenario, by iterating over threshold values ranging from 0 to 1 and retaining the best (i.e. closest to one) F1 score.

Youden's J mean values

Tables S4/5 provide the mean values and standard deviations of Youden's J for the different models with increasing number of biological replicates (related to Figure 5 of the the main text).

	r1 mean	*r1 sd*	*r2 mean*	*r2 sd*	*r3 mean*	*r3 sd*	*r4 mean*	*r4 sd*	*r5 mean*	*r5 sd*
m1	0.791	0.0330	0.803	0.0294	0.809	0.0290	0.812	0.0254	0.814	0.0260
m2	0.808	0.0246	0.821	0.0251	0.825	0.0251	0.829	0.0264	0.832	0.0245
m3	0.873	0.0208	0.882	0.0194	0.886	0.0191	0.887	0.0178	0.889	0.0171
m4	0.868	0.0196	0.882	0.0188	0.886	0.0200	0.886	0.0215	0.888	0.0225

Table S4: Mean values and standard deviations for the Youden's J of "Smooth AUC" outlier detection. Values derived from the data used for the construction of the boxplots presented in the main text (Fig. 5A). Abbreviations: m = model, r = replicate. See Figure S3 caption for model description.

	r1 mean	*r1 sd*	*r2 mean*	*r2 sd*	*r3 mean*	*r3 sd*	*r4 mean*	*r4 sd*	*r5 mean*	*r5 sd*
m1	0.0489	0.0472	0.0570	0.0471	0.0518	0.0521	0.0479	0.0538	0.0387	0.0494
m2	0.0246	0.0319	0.0255	0.0312	0.0164	0.0255	0.0108	0.0166	0.0136	0.0184
m3	0.0793	0.0592	0.0741	0.0525	0.0792	0.0446	0.0767	0.0494	0.0640	0.0528
m4	0.0299	0.0415	0.0267	0.0379	0.0309	0.0422	0.0306	0.0407	0.0265	0.0399

Table S5: Mean values and standard deviations for the Youden's J of *pcadapt* outlier detection. Values derived from the data used for the construction of the boxplots presented in the main text (Fig. 5B). As Table S4, but for *pcadapt*.

Results for the "GLM AUC" method

As explained in the main text, we computed the AUC estimates using two different methods, "Smooth AUC" and "GLM AUC". Both methods lead to highly similar results. In the main text, we reported and discussed the results for the "Smooth AUC". Here, we presented the corresponding plots for the "GLM AUC" (Fig. S4).

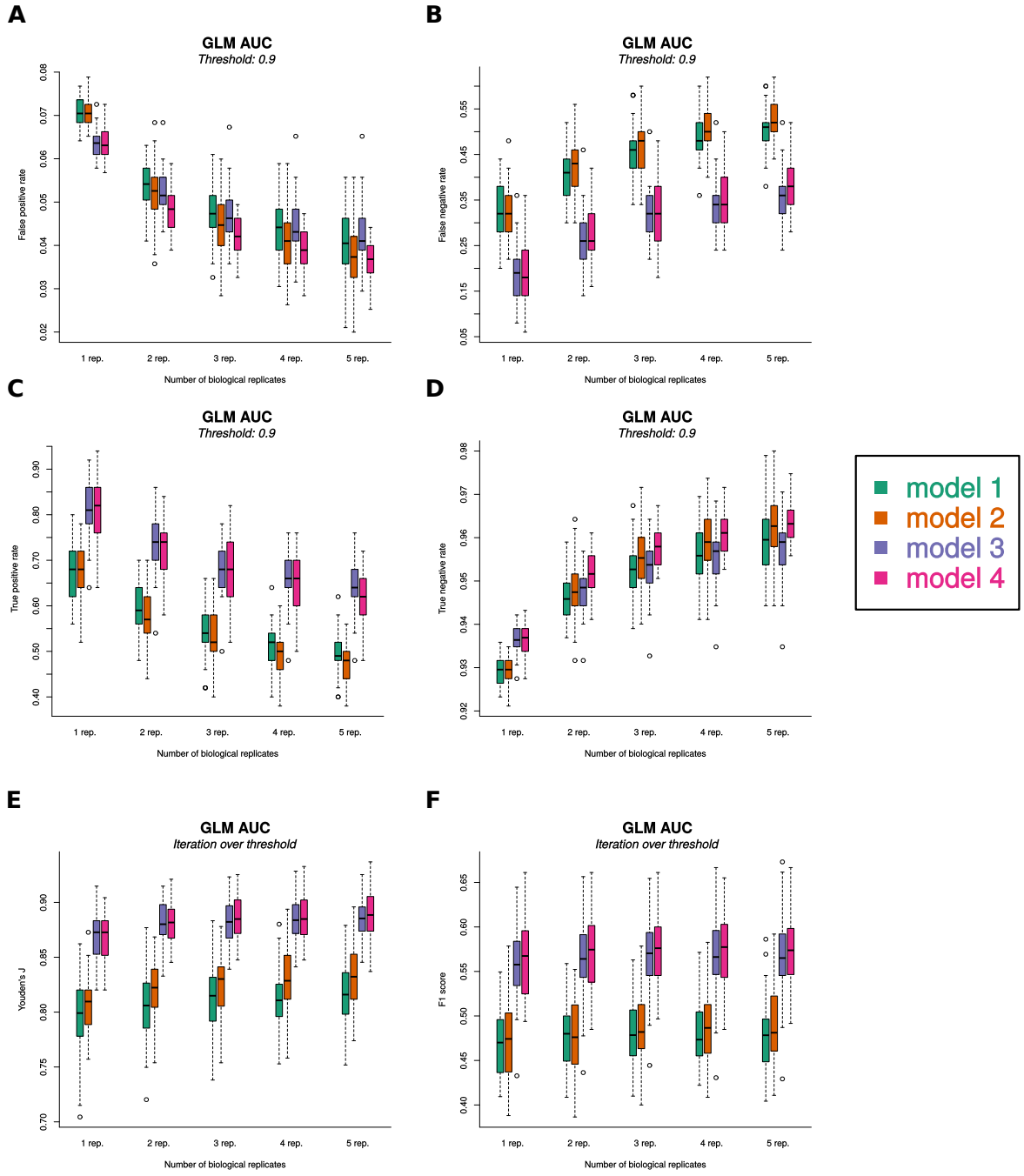

Figure S4: Results obtained with the "GLM AUC" method. The "GLM AUC" method leads to highly similar results than the "Smooth AUC". These latter were presented in the main text, and their "GLM AUC" equivalents are presented here below: **A)** false positive rate, **B)** false negative rate, **C)** true positive rate, **D)** true negative rate, **E)** Youden's J, and **F)** F1 score. See the corresponding "Smooth AUC" plots for comprehensive figure descriptions (Fig. 3/5A/S3A).

AUC estimates and distance to closest selected gene

By relating the AUC estimates for all loci against the distance to the nearest selected gene, we observed that the false negatives are almost always located in close proximity to a selected gene (Fig. S5; plot for one simulation under model 4; similar results are obtained for the other models). This is an indication that linkage disequilibrium impedes accurate detection of adaptive loci, as the neutral markers physically linked to the targeted genes will often result in false negatives (see Discussion in the main text).

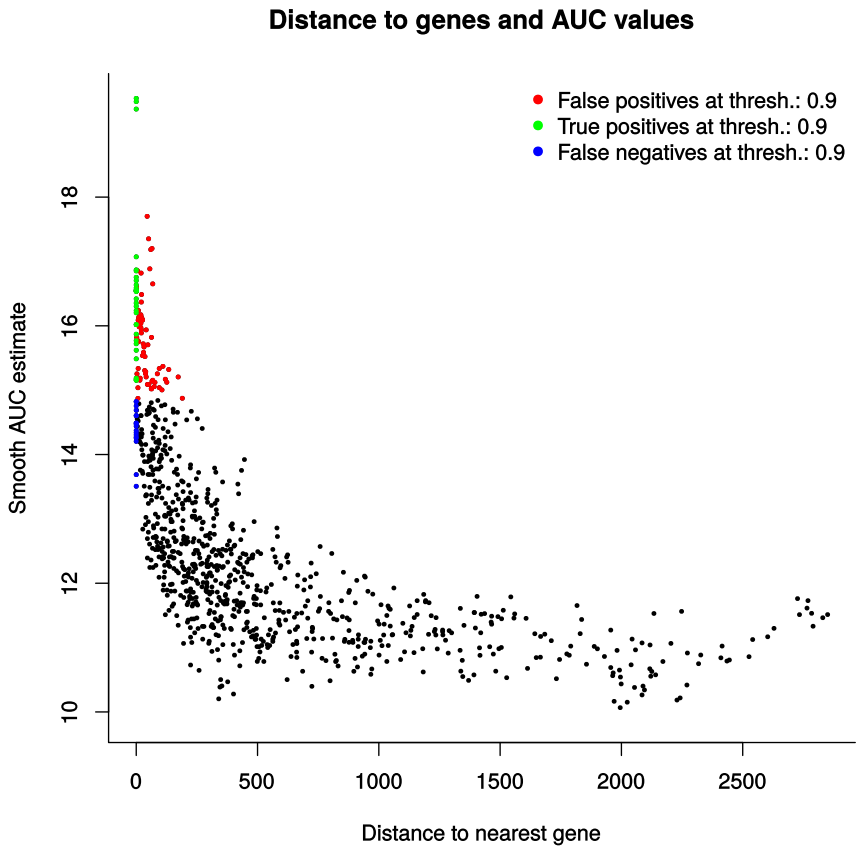

Figure S5: AUC estimates and distance to nearest gene. AUC estimates (computed with the "Smooth AUC" method) in function of the distance to the closest gene under selection. Colours indicate the results of the "Smooth AUC" outlier detection at a quantile threshold of 0.90: red dots indicate neutral markers identified as adaptive genes (false positives); green dots the correctly identified selected genes (true positives); blue dots the genes under selection that are missed (false negatives); and black dots the neutral markers correctly characterized (true negatives).

Performance by subgroups of positions along the gradient

As discussed in the main text, we observed that the outlier detection was most effective under models 3 (var GF + unif ω) and 4 (var GF + var ω). By decomposing the gradient in subgroups of positions, we noted that the performance was better under the model 4 at closest positions, and under model 3 at the distal extremity of the gradient (Fig. S6 for the two first (A) and last (B) positions; see Discussion in the main text).

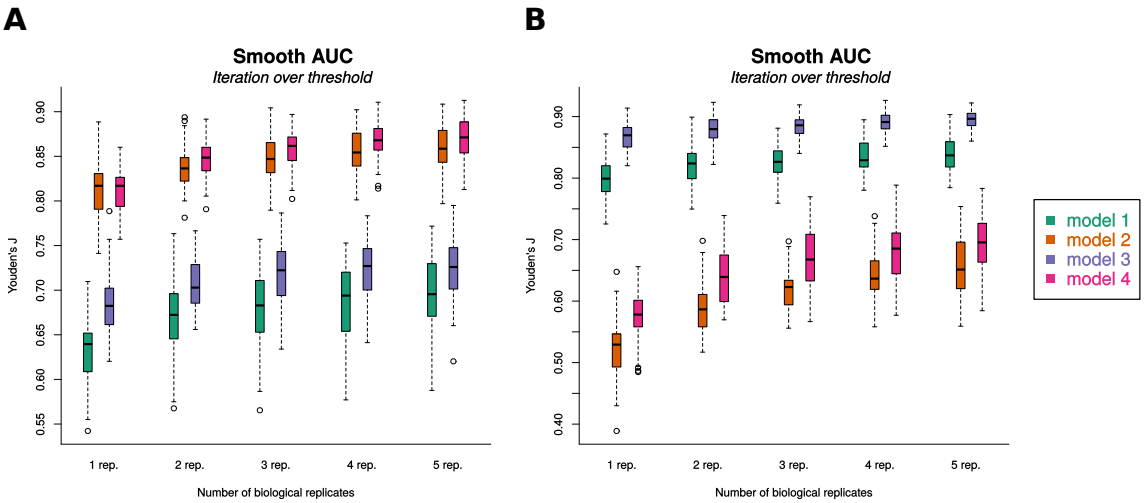

Figure S6: Youden's J for the "Smooth AUC" outlier detection for subgroups of positions. Similar to Figure 5A of the main text, but **A)** for the first two positions along the gradient (where gene flow and/or selection pressure are strong), **B)** for the last two positions along the gradient (where gene flow and/or selection pressure are weak). Refer to Figure 5A for comprehensive figure description.

References

Cui, R., M. Schumer, and G. G. Rosenthal (2016). "Admix'em : a flexible framework for forward-time simulations of hybrid populations with selection and mate choice". *Bioinformatics* 32.7, pp. 1103–1105.

Luu, K., E. Bazin, and M. G. B. Blum (2016). "*pcadapt*: an R package to perform genome scans for selection based on principal component analysis". *Molecular Ecology Resources* 33.