Performance evaluation of adaptive introgression classification methods

Jules Romieu[1,2], Ghislain Camarata[1,2], Pierre-André Crochet[3], Miguel de Navascués[2], Raphaël Leblois[2], and François Rousset[1]

DOI not yet assigned

Abstract

Introgression, the incorporation of foreign variants through hybridization and repeated backcross, is increasingly being studied for its potential evolutionary consequences, one of which is adaptive introgression (AI). In recent years, several statistical methods have been proposed for the detection of loci that have undergone adaptive introgression. Most of these methods have been tested and developed to infer the presence of Neanderthal or Denisovan AI in humans. Currently, the behaviour of these methods when faced with genomic datasets from evolutionary scenarios other than the human lineage remains unknown. This study therefore focuses on testing the performance of the methods using test data sets simulated under an evolutionary scenario inspired by a complex of hybridising lizard species in the Iberian Peninsula: *Podarcis hispanicus*. Our tests focus on analysing the impact of variations in the strength of selection on the performance of three methods (`VolcanoFinder`, `Genomatnn` and `MaLAdapt`) and a standalone summary statistic (Q95). Furthermore, the hitchhiking effect of the adaptively introgressed mutation can have a strong impact on the flanking regions, and therefore on the best differentiation between the genomic windows classes (*i.e.* AI/non-AI). For this reason, three different types of non-AI windows are therefore taken into account in our analyses: independently simulated neutral introgression windows, windows adjacent to the window under AI and windows coming from a second neutral chromosome unlinked to the chromosome under AI. Our results highlight the importance of taking into account adjacent windows in the training data in order to correctly identify the window with the mutation under AI. Finally, our tests show that methods based on Q95 seem to be the most efficient for an exploratory study of AI (*i.e.* without any a priori on the presence of AI).

Keywords: Performance evaluation, simulations, adaptive introgression, selection, demographic scenario, *Podarcis*

[1]ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France, [2]CBGP, INRAE, CIRAD, IRD, Institut Agro, Univ Montpellier, Montpellier, France, [3]CEFE, CNRS, Univ Montpellier, EPHE, IRD, Montpellier, France

Correspondence
jules.romieu@umontpellier.fr

Introduction

Hybridisation is the reproduction of individuals belonging to differentiated gene pools. As long as reproductive isolation remains incomplete, hybridisation and backcrossing can lead to introgression, *i.e.* the incorporation of genetic material from a donor population into the genome of a recipient population (Anderson and Hubricht, 1938). Mallet estimated that hybridisation occurs in over 25% of plant species and 10% of animal species, and more evidence of interspecific hybridisation and introgression has emerged since then in plants and animals following the advent of high-throughput sequencing technologies (Taylor and Larson, 2019; Twyford and Ennos, 2012). Indeed, the barriers between species are not completely impermeable, allowing genetically distinct species to be maintained over time in the presence of gene flow (Harrison and Larson, 2014; Suarez-Gonzalez et al., 2018). While interspecific introgression is increasingly recognised as an important process in the evolution of biodiversity, its prevalence (in terms of the proportion of species involved) and importance (in terms of the proportion of the genome that is affected) remain poorly understood (Edelman and Mallet, 2021).

Hybridization and introgression have many potential evolutionary outcomes. They can lead ultimately to the accumulation of deleterious mutations (Adavoudi and Pilot, 2022), or to the loss of genetic diversity by the merging of the two parent species (Runemark et al., 2019), or even to the extinction of the parent species through genetic swamping (Todesco et al., 2016). Introgression can also allow the acquisition of alleles that increase the fitness of carrier individuals, a process known as adaptive introgression (AI; Burke and Arnold, 2001; Kim and Rieseberg, 1999). Although there are few examples to date, adaptive introgression is thought to occur in both animals and plants (e.g. Burgarella et al., 2019; Hedrick, 2013). In some cases, adaptive introgression could even play an important role in the speciation process as proposed in Darwin's finches (Grant and Grant, 2019) or butterflies of the genus *Heliconius* (Pardo-Diaz et al., 2012; Rosser et al., 2024).

Proving the presence of adaptive introgression in a species is challenging, because evidence is needed to support that the regions come from another species and that they provide a selective advantage on the fitness of the individuals carrying them or exhibit signs of positive selection in the recipient species (Burgarella et al., 2019; Caeiro-Dias et al., 2021). The first studies of adaptive introgression looked independently for introgression and selection signals in genetic sequences. However, the signals left by adaptive introgression in genetic sequences can be quite distinct from signals left by neutral introgression or by selective sweeps of non-introgressed variants (Burgarella et al., 2019; Racimo et al., 2015; Setter et al., 2020; Taylor and Larson, 2019). For this reason in recent years, classification methods have emerged to detect genomic signatures specific to adaptive introgression (Gower et al., 2021; Racimo et al., 2017; Setter et al., 2020; Zhang et al., 2023). These methods use various approaches to identify different genetic patterns specific to adaptive introgression, such as summary statistics focusing on excess of allele sharing between the introgressed population and the introgressing population (Racimo et al., 2017), genomic scans identifing patterns of diversity specific to adaptive introgression (Setter et al., 2020) or machine learning algorithms trained on data simulated under adaptive introgression scenarios (Gower et al., 2021; Zhang et al., 2023).

Most of these methods have been developed and tested on human genetic data to identify potentially adaptively introgressed regions caused by ancient hybridization among groups of humans: *Homo sapiens*, *Homo neanderthalensis* and Denisovans (Human evolutionary models

reviewed in Racimo et al., 2015). In addition, methods using machine learning algorithms have been only trained on genetic datasets simulated under human scenarios. The behaviour and performance of these methods on organisms with evolutionary scenarios distant from the human lineage are therefore unknown. Using a simulation-based inference method trained on a particular biological model to infer a biological process on an other model could degrade method's performances. The signals left by adaptive introgression depend on life history traits and evolutionary history of the biological model. The history of connections between populations, their sizes and divergence times, as well as the availability for sampling of introgressing, introgressed and non-introgressed populations, and the strength of selection are features specific to the biological model and the study, which influence the signature of AI in the genome and therefore the performance of the methods. In the case of non-model species, the knowledge of demographic history and genomic parameters such as mutation and recombination rates is often limited. For these species, it is difficult to choose the scenario under which genetic data should be simulated to train machine learning methods to detect AI, and more generally to define a score threshold differentiating regions under AI from other regions. Indeed, most current studies trying to detect AI in non-model species use `VolcanoFinder`(Setter et al., 2020), the only method that does not require the definition of an evolutionary model or knowledge of the introgressing species (Liu et al., 2022; Pawar et al., 2023; Wang et al., 2023) to be applied on a given data set. However, some authors have suggested that alternative simulation-based methods are robust to misspecification of demographic history (Gower et al., 2021; Zhang et al., 2023); it is thus important to test this robustness on evolutionary scenario distinct from the (human) ones considered in the design of these methods.

Furthermore, as these methods have been published recently, no study have compared their performance since the original articles. One aim of our work is therefore to compare the performance of published adaptive introgression classification methods in relation to variations in evolutionary parameters impacting the detection of adaptive introgression (divergence time between introgressed and introgressing populations, strength of selection on the introgressed allele). Last, each method has been evaluated against a different type of non-AI windows (genomic windows that do not carry AI alleles, see below). We thus also compared how existing methods fare to distinguish AI-windows from different types of non-AI windows in the test sets. In this study, we perform simulation tests of the performance of 4 different existing AI methods: `VolcanoFinder` (Setter et al., 2020), `MaLAdapt` (Zhang et al., 2023) and `Genomatnn` (Gower et al., 2021) and a method based on the summary statistic Q95(w, y) (Racimo et al., 2017).

To do so, we simulated genetic dataset under various evolutionary scenarios inspired by the life history traits of a complex of species of lizard of the genus *Podarcis* (Caeiro-Dias et al., 2023; Gaczorek et al., 2023). The *Podarcis hispanicus* complex includes at least height species that inhabit the Iberian Peninsula where they are sympatric or parapatric, meeting in narrow contact zones (Pinho et al., 2009; Renoult et al., 2009). Various levels of genomic and mitochondrial introgression have been identified between several species (Caeiro-Dias et al., 2021; Pinho et al., 2009; Renoult et al., 2009; Yang et al., 2021), although intrinsic barriers against introgression persist, confirming that they are in the late stages of speciation (Caeiro-Dias et al., 2023). For example, recent introgression between populations of *Podarcis carbonelli* and *Podarcis bocagei* is suggested by the discovery of first-generation hybrids and backcrosses (Caeiro-Dias et al., 2021; Pinho et al., 2009). In addition, adaptive introgression may potentially occur between these two

species, in view of the large number of MHC alleles shared between sympatric individuals of the two species compared to allopatric individuals (Caeiro-Dias et al., 2023; Gaczorek et al., 2023). The fact that these are extant species eases the sample collection and provides better quality genetic material than ancient DNA. Finally, their divergence times are in the order of a million generations (Kaliontzopoulou et al., 2011), much older than divergence between humans and neanderthals dating from less than 20,000 generations (Prüfer et al., 2014). Such a long-standing diversification between the donor and recipient species, increasing the differentiation between the two, might allow better identification of the regions of adaptive introgression (Setter et al., 2020) and thus lead to differences in classification performance of the tested methods compared to the study of the human lineages. This species complex is thus becoming a model for the study of introgression between highly differentiated species, and will be the focus of an empirical evaluation of the extant of AI in different ecological contexts based on whole genome sequence data that are currently being generated (authors' work in progress). For all those reasons, the *Podarcis hispanicus* complex seems to be an ideal system to test the existing methods designed to detect AI as a complement to previous similar studies carried out in the context of human evolution.

As briefly mentioned above, the type of non-AI windows that have be used to train the machine learning models differs between the simulation-based classification methods. Some methods use neutral introgression windows simulated independently of AI windows (Gower et al., 2021), while others use windows that do not carry the mutation under AI but share the same demographic and adaptive history (*i.e.* generated by the same simulation, Zhang et al., 2023), and that can even be physically linked to the AI-windows. As the aim of all these classification methods is to specifically identify windows carrying introgressed advantageous mutations, the types of non-AI windows used in the train set to achieve this objective is extremely important. Indeed, testing for the presence of AI in a genetic dataset composed of windows adjacent to the AI window using a method not trained with this type of window may result in the identification of false positives. Depending on the strength of the selection applied to the introgressed mutation, the impact of hitchhiking can lead to the identification of false AI signals in these windows. In addition, the other methods not requiring training, also use different non-AI window classes to test their methods. The impact of whether or not a non-AI window type is taken into account must therefore be evaluated.

In this study, we thus evaluate the performance of four AI classification methods to detect genomic regions containing an allele under adaptive introgression, as a function of the selection strength in an evolutionary context inspired by the *Podarcis hispanicus* complex. The objective of the methods evaluated is to detect the locus containing the causal variant under selection. However, because of hitchhiking, the signal of adaptive introgression can extend to large portions of the genome. Therefore, we also explore the impact of intra- and inter-chromosomal hitchhiking on performance, by taking into account in the test sets adjacent windows linked to the window carrying the mutation under AI, as well as a second chromosome harboring only neutral windows but potentially affected by hitchhiking. Indeed, although hitchhiking is mainly considered as the selective effect of an advantageous allele on genetic variation at adjacent neutral alleles linked to it (Smith and Haigh, 1974), which we call intra-chromosomal hitchhiking in this article, its effect extends to other chromosomes, an effect named here inter-chromosomal hitchhiking. Our results highlight the differences among the methods, provide a guideline on what method to use for the study of adaptive introgression. Our simulation tests suggest that the simple genome

scan method based on one of the Racimo et al. (2017) summary statistics has better performance in the classification of genomic regions under adaptive introgression than all other more complex methods.

Material and methods

Adaptive introgression model

We simulate genetic data under an adaptive introgression model where a beneficial allele is transferred by gene flow from a donor population to a recipient population. The full model is composed of five populations (Fig. 1): an outgroup population (O), a recipient population (R), a non-introgressed sister population of the recipient one (RS), a donor population (D), and a non-introgressing sister population of the donor one (DS). The donor's sister population is used in our study as a stand-in for the donor population to mimic empirical situations where the latter is unknown or where genomic data are not available from it (Setter et al., 2020). Genetic data sampled from different combinations of these populations are used by the evaluated methods.

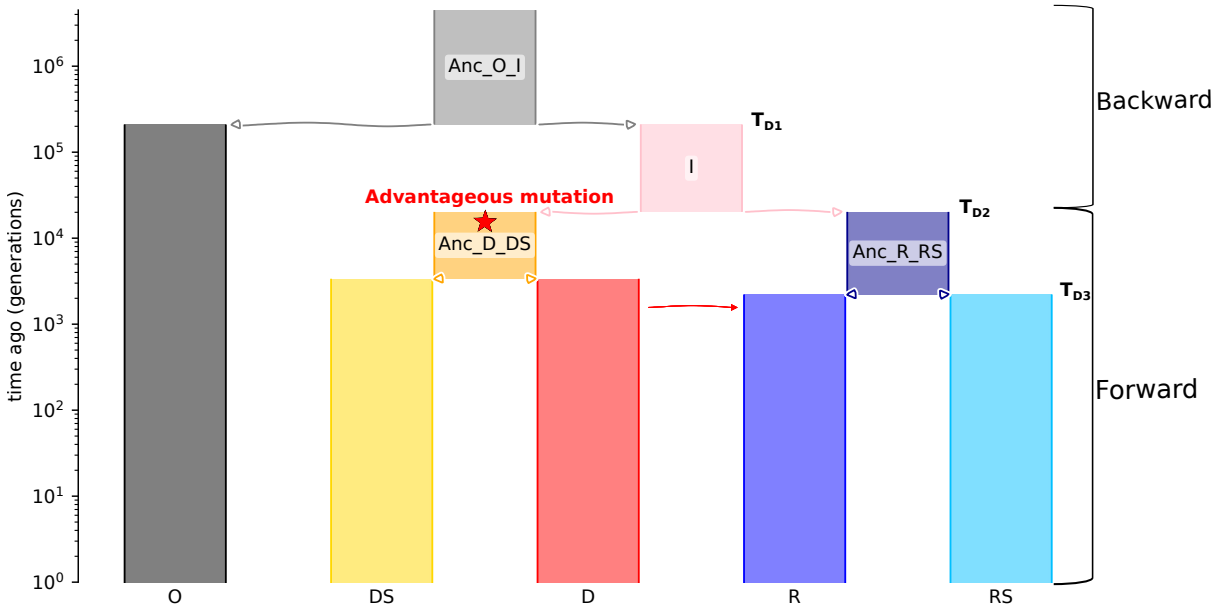


Figure 1 – Demographic model used in this study.
Anc_O_I = Outgroup and Ingroup ancestral population, **O** = Outgroup population, **I** = Ingroup population, **Anc_D_DS** = Donor and non-introgressing ancestral population, **Anc_R_RS** = Recipient and non-introgressed ancestral population. **DS** = sister to the donor population (non-introgressing population), **D** = Donor population (introgressing population), **R** = Recipient population (introgressed population), **RS** = Recipient sister population (non-recipient population), T_{D1} = Divergence time between **O** and **I**, T_{D2} = Divergence time between **Anc_D_DS** and **Anc_R_RS**, T_{D3} = Divergence time between **R** and **RS**.

Population sizes are constant throughout the duration of the simulation and fixed to a single value N for all populations. At the end of the simulation, a given number of individuals is sampled in each population, except the donor one. Except for reference scenarios of neutral introgression (NI) for which we consider a single chromosome, we simulate diploid genomes with two chromosomes, of length 1Mb each, where neutral mutations can occur at a mutation rate μ and (intra-chromosomic) recombination events at a rate r, both rates being expressed per base

per gamete (*i.e.* per generation). The different chromosomes are simulated by setting a recombination rate of 0.5 between the two genomic regions defining them. To simulate an adaptive introgression event, a single beneficial mutation (with selection coefficient s) is positioned at the center of one of the two haploid genomes of a single individual within the ancestral donor population (Anc_D_DS in Fig. 1). The beneficial mutation is then potentially transferred from the donor population to the recipient population during a single generation of migration, with an immigration rate m. The selection coefficient for the advantageous mutation does not change over time or between populations (*i.e.* s is the same for Anc_D_DS, D, DS and R). The simulation is discarded when this selected mutation is lost in AI scenarios, or when no genetic material is introgressed in NI scenarios.

Demographic parameter values are chosen to produce a demographic scenario whose divergence times in generations (all times are expressed backward from present) between donor and recipient populations are inspired by the evolutionary history of the genus *Podarcis* (table 1). In particular, we focus on the species pair *Podarcis carbonelli* (donor) and *Podarcis bocagei* (recipient; Caeiro-Dias et al., 2021, 2023; Gaczorek et al., 2023; Pinho et al., 2009) and set the generation time to 2.5 years (Pinho et al., 2007). These two species diverged from the outgroup, *Podarcis muralis*, 4 million generations ago (10 million years ago, Kaliontzopoulou et al., 2011). *Podarcis carbonelli* and *P. bocagei* diverged 2.24 million generations ago (5.6 million years ago, Kaliontzopoulou et al., 2011). As there is no available estimate of the intra-specific divergence times between populations of *P. bocagei* and *P. carbonelli*, chosen values correspond to the current state of our knowledge of their evolutionary history. The divergence time is set to 40,000 generation (100,000 years) among *P. carbonelli* populations and to 4,000 generations (10,000 years) among *P. bocagei* populations, reflecting, respectively, divergences before and after the Last Glacial Maximum (LGM, Clark et al., 2009) as suggested in Pinho et al. (2007) and Pinho et al. (2011) and Caeiro-Dias et al. (2021). The time of the migration event from *P. carbonelli* to *P. bocagei* is arbitrarily set to 2,000 generations, in-between present and the *P. bocagei* population divergence. As no information is available from the literature for some parameters, commonly accepted values for more or less closely related species are chosen: (1) μ is set to 1.2×10^{-8} mutation per base per gamete, the average reptile's mutation rate (Bergeron et al., 2023); (2) the recombination rate is set to the the same value as the mutation rate; and (3) m and N are set to 0.01 and $10,000$ respectively, values that facilitate comparison to previous work focused on adaptive introgression in humans (e.g. Racimo et al., 2017).

We use a combination of forward and backward (coalescent) simulators to generate the genetic data. First, `msprime` (Baumdicker et al., 2022; Kelleher et al., 2016) simulates backward in time the ancestral recombination graph (ARG) of the whole population from the divergence of populations ancestral to the current donor and recipient populations (*i.e.*, at time T_{D2} in Fig. 1), until the most recent common ancestor. `SLiM` (Haller and Messer, 2019) is then used to simulate, forward in time and starting from `msprime`'s ARG, the more recent section of the evolutionary scenario which includes selection, *i.e.* from T_{D2} until present. `SLiM`'s ARG of the individuals sampled in current populations is used by `msprime` to add neutral mutations to generate the sample polymorphism.

Sample sizes are set to 99, 108, 2 and 2 for the recipient population, the non-introgressed population, the sister population of the donor, and the outgroup population, respectively, corresponding to the values used by the tested simulation-based classification methods (`Genomatnn`

and `MaLAdapt`, Gower et al., 2021; Zhang et al., 2023). However, these two methods used samples from the donor population for their train and test sets where we used samples from the sister population of the donor. Haplotypic and genotypic matrices for all samples are then obtained from the ARG with mutations, using the `tskit` (Ralph et al., 2020) and `scikit-allel` packages (Miles et al., 2024). As forward simulations can be extremely time-consuming, a scaling factor of 10 was applied to all parameter values (except m) to drastically reduce simulation times: s, μ and r are multiplied by 10, and the times of events and N are divided by 10 (as in Gower et al., 2021; Hoggart et al., 2007; Zhang et al., 2023). Although the use of a scaling factor on selection coefficient and population size, can bias the forward simulation results (Dabi and Schrider, 2024), its use in our analysis is justified by the fact that simulating the test sets under the *Podarcis* scenario without it would be extremely time-consuming, but also because the simulation-based classification methods (`Genomatnn` and `MaLAdapt`) used it to generate their training set.

To assess the impact of selection strength, test sets are generated with different selection coefficients representing strong, moderate and weak selection scenarios (s values of 10^{-1}, 10^{-2} and 10^{-3}).

Table 1 – Parameter values for *Podarcis* scenarios

Parameters	Values
Ingroup/Outgroup divergence time (T_{D1}, generations)*	4000000
D/R divergence time (T_{D2}, generations)*	2240000
D/DS divergence time (generations)*	40000
R/RS divergence time (T_{D3}, generations)*	4000
Migration time (generations)*	2000
Neutral mutation rate (μ, per base per generation)*	1.2×10^{-8}
Advantageous mutation selection coefficient (s)*	$\{10^{-3}, 10^{-2}, 10^{-1}\}$
Recombination rate (r, per base per generation)*	1.2×10^{-8}
Immigration rate (m)	10^{-2}
Population size (N, diploid individuals)*	10000
Genome length (megabase pairs)	{1, 2}

*Parameters subject to scaling.

Methods to detect adaptive introgression

The performance of four adaptive introgression classification methods: `VolcanoFinder` (Setter et al., 2020), `MaLAdapt` (Zhang et al., 2023) and `Genomatnn` (Gower et al., 2021) and a method based on the summary statistic Q95(w, y) (Racimo et al., 2017) are evaluated in this work. All of them are classification methods whose objective is to detect regions of the genome that harbour a beneficial mutation originating from the donor population (AI region).

The Q95(w, y) statistic (Racimo et al., 2017) is one of the first summary statistics developed to specifically capture the signal of adaptive introgression genetic data. This summary statistic can be calculated for both phased and unphased genotypes. It gives the 95% quantile of the distribution of derived allele frequencies in the recipient population, for sites that are at higher or equal at a frequency y in the donor population and at frequency lower than w in the sister population of the recipient population. This summary statistic (named Q95 thereafter), with $w = 0.01$ and $y = 1$, and computed on 50kb non-overlapping windows, is used here as a score value to detect adaptive introgression.

`VolcanoFinder` (Setter et al., 2020) is the first adaptive introgression classification method to have been developed. This method aims to identify adaptive introgressed alleles recently fixed in the recipient population considering only the genetic information from the introgressed population. It uses an optional outgroup population to polarize the alleles and does not require phased haplotypes. For a given genomic position, `VolcanoFinder` calculates a composite likelihood ratio for the site frequency spectrum (SFS) around that focal site, comparing an AI model and a reference SFS acting as empirical null model. In our tests, we used the SFS of the whole chromosome as reference. To compute the composite likelihoods, `VolcanoFinder` can estimate the strength of selection and the genetic differentiation between the recipient and donor populations or use user-specified fixed values. We choose to let `VolcanoFinder` infer them, as recommended in the original publication. We set the interval between each tested site to 1000bp. The scores per position given by the method (*i.e.* the composite likelihood ratio) are further converted into a single score per 50kb window, corresponding to the highest score observed in the window considered. To obtain a relative score (*i.e.* between 0 and 1 as for the other methods), the score of each window is then divided by the maximum score observed among all windows and all analysed datasets for a given scenario.

`Genomatnn` (Gower et al., 2021) is a simulation-based classification method that uses a deep learning algorithm (convolutional neural network, CNN), trained on 100kb windows simulated independently under different scenarios with samples from the donor, recipient and recipient's sister populations, to detect adaptive introgression. `Genomatnn` gives a probability for each 100kb tested window to have been affected by an adaptive introgression event and can use both phased and unphased data. The training set includes sequences simulated under adaptive introgression, neutral introgression and classical selective sweep scenarios, and is used to classify genomic regions (windows) as AI or non-AI. We used two pre-trained CNNs (Nea_to_CEU_af-0.25_2918410235.hdf5 and Nea_to_CEU_af-0.05_2250018620.hdf5) described in the article presenting the method and corresponding to CNNs trained on simulations filtered so that the frequency of the beneficial allele in the recipient population was greater than or equal to 5% and 25% (respectively `Genomatnn_0.05` and `Genomatnn_0.25` thereafter; Gower et al., 2021). We use these originally trained CNNs to test their robustness when using test datasets simulated under a different demographic scenario (*Podarcis* scenario).

`MaLAdapt` (Zhang et al., 2023) is also a simulation-based classification method and uses a machine learning (ML) algorithm (extra tree classifier, ETC, Geurts et al., 2006), trained on simulations of a 5Mb chromosome phased genome under different scenarios with samples from the donor, recipient and recipient's sister populations, to classify genomic regions (50kb windows) as AI or non-AI. For each window, `MaLAdapt` computes a score corresponding to the probability that the window has experienced an AI event. Contrary to `Genomatnn`, `MaLAdapt` is not trained on raw sequences, but on a set of summary statistics calculated on each window. These summary statistics were chosen by Zhang et al. (2023) to capture patterns of genetic diversity, linkage disequilibrium and population differentiation, that are informative about introgression, selection and adaptive introgression. Contrarily to Zhang et al. (2023), we use non-overlapping windows and do not consider the presence of exons in the genome. The number of exons per window, a summary statistic used by Zhang et al. (2023), has been removed because this information is only available for annotated genome, which is rarely the case for non-model species. Furthermore, its use rests on simulations assuming that selection only takes place in exons, which is

not consistent with evidence of selection outside exons (Jo and Choi, 2015). Such assumption, should systematically lead to the classification of windows without exon as non-AI windows and thus artificially improve the method's performance. We also do not compute Kelly (1997)'s Z_{nS} summary statistic, because it is computationally intensive to calculate and because Z_{nS} does not carry pertinent information according to importance score reported in the publication describing `MaLAdapt` (Zhang et al., 2023). Re-using the original simulated training set from Zhang et al. (2023), we retrained two ML model without the two statistics mentioned above, one with the Q95$(0.01, 1.0)$ statistic and the other without (respectively `MaLAdapt_Q95` and `MaLAdapt_noQ95` thereafter), which are equivalent to models 4 and 6 from Zhang et al. (2023) but without Z_{nS} and number of exons. Summary statistics from train and test sets are standardised using their mean and variance computed on the training set (whereas Zhang et al., 2023, used the mean and variance computed on the analysed test set).

Inferences were thus carried out on non-overlapping 50kb windows (for `VolcanoFinder` `MaLAdapt`and the method based on Q95) or 100kb windows (for `Genomatnn`), and using phased genotypes of the donor's sister population, recipient and recipient's sister populations for `Genomatnn` and `MaLAdapt` or from the recipient population only for `VolcanoFinder`.

Method performance evaluation

The four methods can be used to identify candidate AI-windows as those whose score is higher than a given threshold. The threshold values need to be defined for each data set, as they depend on the sampling scheme and evolutionary history of the samples. Here, varying the score threshold value used to classify the windows as AI or non-AI, we compute true and false positive rates to generate receiver operating characteristic (ROC) curves for the different types of non-AI windows, and calculate the area under the ROC curve (AUC, Bradley, 1997). The use of these evaluation metrics (ROC and AUC) allows us to avoid defining score thresholds for each method that would only be relevant for a given scenario and sampling scheme.

For all methods tested, we define a region being under adaptive introgression as a window containing the advantageous mutation originating from the donor population (named AI window). But the definition of the "reference" non-adaptive introgression (non-AI) regions (*i.e.* the null model) differs for the different classification methods: (i) `VolcanoFinder` does not assume a specific null model in terms of evolutionary history but rather defines an empirical null model from the genome-wide SFS of the sampled individuals in the recipient population; (ii) `Genomatnn` uses a train set including, as non-AI windows, neutrally introgressed (NI) windows and non-introgressed windows with a beneficial site (classic sweep) with a class ratio of 1:1:1 (AI:NI:classic sweep), simulated independently of each other; and (iii) `MaLAdapt` uses (1) all windows of the 5Mb chromosome other than the one carrying the advantageous mutation (referred to here as adjacent windows) as non-AI windows, so that AI and adjacent non-AI windows are jointly simulated and can be physically linked, (2) selective sweep windows simulated independently of the AI windows. The AI:non-AI class ratio of the final training set is 1:2. By including adjacent windows as non-AI windows in its training sets, `MaLAdapt` treats as non-AI some windows that have undergone hitchhiking and therefore potentially do not have the same genetic diversity as neutral introgression windows simulated independently of the AI window. Indeed, depending on the strength of selection, recombination and demographic events, the hitchhiking effect of the advantageous mutation can have a strong impact on the evolution of the frequency and distribution of linked neutral alleles. In the context of adaptive introgression analysis, hitchhiking

certainly has an effect on inferences and may notably lead, for example, to the detection of false AI signals on windows adjacent to the beneficial allele. Furthermore, the effects of hitchhiking may extend to other chromosomes than the one carrying the mutation, as even free recombination ($r = 1/2$) will not immediately suppress the inter-chromosomal linkage disequilibrium in the first generations after the introgression of the beneficial haplotype. None of the AI classification methods tested here take this inter-chromosomal hitchhiking into account in the definition of its training set nor their sensitivity with respect to such hitchhicking in the (simulated) data has been tested.

To examine the effect of different forms of hitchhiking on the distribution of scores computed by each method we included neutral windows affected by intra- or inter-chromosomal hitchhicking in the test sets. The train sets were not modified in this respect relative to the original definitions of the methods (the definition of the `Genomatnn` train set ignores hitchiking; for `MaLAdapt`, it includes intra-chromosomal hitchhicking; the two other methods are not machine-learning methods with a train-set concept). The effect of inter-chromosomal hitchhiking is assessed by simulating genomes containing two chromosomes of 1Mb each, the first one carrying the introgressed adaptive mutation and the second one being fully neutral, and by comparing scores for windows from the second chromosome to scores within the first one. The effect of intra-chromosomal hitchhiking is assessed by examining the variation in distribution of scores for windows at different distance from the adaptive mutation on a 1Mb chromosome. Last, we also generated test sets of a 1Mb chromosome simulated without any selection (so containing only neutral introgression windows) to represent `Genomatnn` types of neutral non-AI windows.

For each scenario explored, we simulated one test set of 200 simulated data. We represent the distributions of scores over the 200 replicates as function of the distance to the beneficial mutation when relevant (*i.e.* not for the chromosome 2 and independent neutral chromosome, where hitchhiking affects the test sets).

Results

We compare the performance of the four methods described above in relation to two main factors of variation: the strength of selection and different types of non-IA windows (see above): windows simulated independently under an neutral introgression scenario, or neutral windows not physically linked to the beneficial mutation (*i.e.* on the second chromosome), or neutral windows on the same chromosome as the beneficial mutation (*i.e.* adjacent windows). The consideration of those three different classes of non-AI window classes allows to evaluate false positive rates in different regions of the genome. The ROC curves (presented only in the Supplementary Information, Fig. S4 and S5) and AUC express the expected performance of each method and ease the comparisons between them, whereas the plots of scores along the chromosomes are used to explore in more details the effects of hitchhiking on their performance. The performance of `MaLAdapt_noQ95` varies in the same way as `MaLAdapt_Q95`but with poorer AUC values (Fig. S1 to S3).

Intra-chromosomal hitchhiking effect — classification of adjacent windows

The scores of linked neutral windows (adjacent windows to the window under selection) are expected to decrease with increasing distance from the mutation due to the effect of hitchhiking. The adjacent windows closest to the window under AI, which potentially have the highest

scores, would therefore increase the FPRs. The AUCs can be used to highlight the differences between the scores for the different adjacent window classes. The pattern of decrease in score values is supposed to result in an increase in AUC values moving away from the window under AI. This pattern is expected for the combinations of methods/strengths of selection conditions where IA windows can be distinguished from independent neutral windows. This pattern of decrease in score values is apparent for most methods and s values, except for `VolcanoFinder` and `Genomatnn_0.25` in the case of weak and `VolcanoFinder` for the intermediate selection. For some methods, at a certain distance from the AI window the AI/non-AI classification even becomes perfect (meaning that AUC$= 1$). Interestingly, this pattern of score decrease varies according to the methods and the strength of the selection changes the maximum AUC values reached by each method considered.

For intermediate selection (Fig. 2 and Fig. S1), for the method based on Q95 and `MaLAdapt`, from the window under AI to the most distant adjacent windows, the scores decrease progressively until they reach a plateau at a certain distance (around 0.4 cM), which results in AUC values close to 1 for this summary statistics. Moreover, the Q95perform better than the other methods when the non-AI windows considered are close to the mutation under selection (from 0.06 to 0.3 cM). For `Genomatnn` on the other hand, the decline is much more extreme, with very high scores at windows close to the introgressed mutation, followed by an abrupt fall at around 0.36 cM (leading to an increase in AUC values approaching 1 from this distance). `VolcanoFinder` shows opposite patterns of variation to those observed with the other methods, with lowest average score values for windows close to the site with the beneficial mutation (even lower than those of the target site), and progressively increasing towards the chromosome edges. In addition, its performance for this selection strength is the lowest of all the methods, with a maximum $AUC = 0.75$ close to the introgressed window and falls below 0.5 at the edge of the chromosome.

For strong selection scenarios, there is a strong increase in the hitchhiking effect (Fig. 3 and Fig. S2) which increases the score distributions of all adjacent windows for the Q95 and `MaLAdapt` and `Genomatnn`. The decrease remains observable for the Q95 and `MaLAdapt` but the effect of hitchhiking increases the distance required for the methods to differentiate perfectly between AI and non-AI windows. However, for `Genomatnn`, AI windows and adjacent windows are no longer distinguishable on the basis of their scores. `Genomatnn`'s performance is greatly reduced by the close distributions between the scores of AI and adjacent windows (leading to a very small increase in AUC with distance and a low maximum value: AUC$= 0.56$ for `Genomatnn_0.25` at 0.6 cM). Unlike other selection regimes, for `VolcanoFinder` the pattern of score decrease is visible when the selection is strong. `VolcanoFinder` becomes the best performing method because the distance along the chromosome required to reach perfect discrimination (AUC value of 1) is shorter than for the Q95 (0.36 cM versus 0.54 cM).

In the case of weak selection (Fig. 4 and Fig. S3), there is little signal even for the AI window and signal decay with distance is barely visible for Q95, `MaLAdapt` and `Genomatnn_0.05` (Fig. S3). A weak selection leads the AUCs to never reach 1 (maximum value reached: AUC$= 0.87$), because of the proximity of the distributions of the scores of the windows under AI to the non-AI windows, even if the increase in the AUC with the distance of the non-AI windows is always visible. For `VolcanoFinder`, the pattern of inverted score values previously observed for intermediate selection is also found for weak selection. Its performance is even lower than that of

the intermediate selection, with a maximum AUC well below 0.5. Interestingly, this is the only selection force where `Genomatnn_0.05` outperforms `Genomatnn_0.25`. For `Genomatnn_0.25`, the distributions of scores for the AI windows and adjacent windows are almost identical, resulting in low AUC values (below 0.5).

Comparison with independent neutral introgression windows

For certain combinations of methods/selection strength, the distributions of the scores of the AI and independently simulated neutral introgression windows (NI windows) are disjoint, resulting in a perfect classification of the AI/non-AI windows ($AUC = 1$ for Q95, `MaLAdapt` and `Genomatnn_0.25` in the case of strong selection and Q95, and `MaLAdapt` for intermediate selection, Fig. 2 and 3). But, for some selection strengths and methods, the distributions of NI window scores overlap the distributions of AI windows. This is the case for all methods in the presence of weak selection (Fig. 3 and Fig. S2); this results in lower AUC values than for higher selection coefficients. For `VolcanoFinder` this overlap in AI/NI score distributions can be observed whatever the selection strength (in some cases the highest NI window scores are even higher than the highest AI scores). Finally, this overlap is also present, to a lesser extent, for `Genomatnn` (except for `Genomatnn_0.25` in the presence of strong selection).

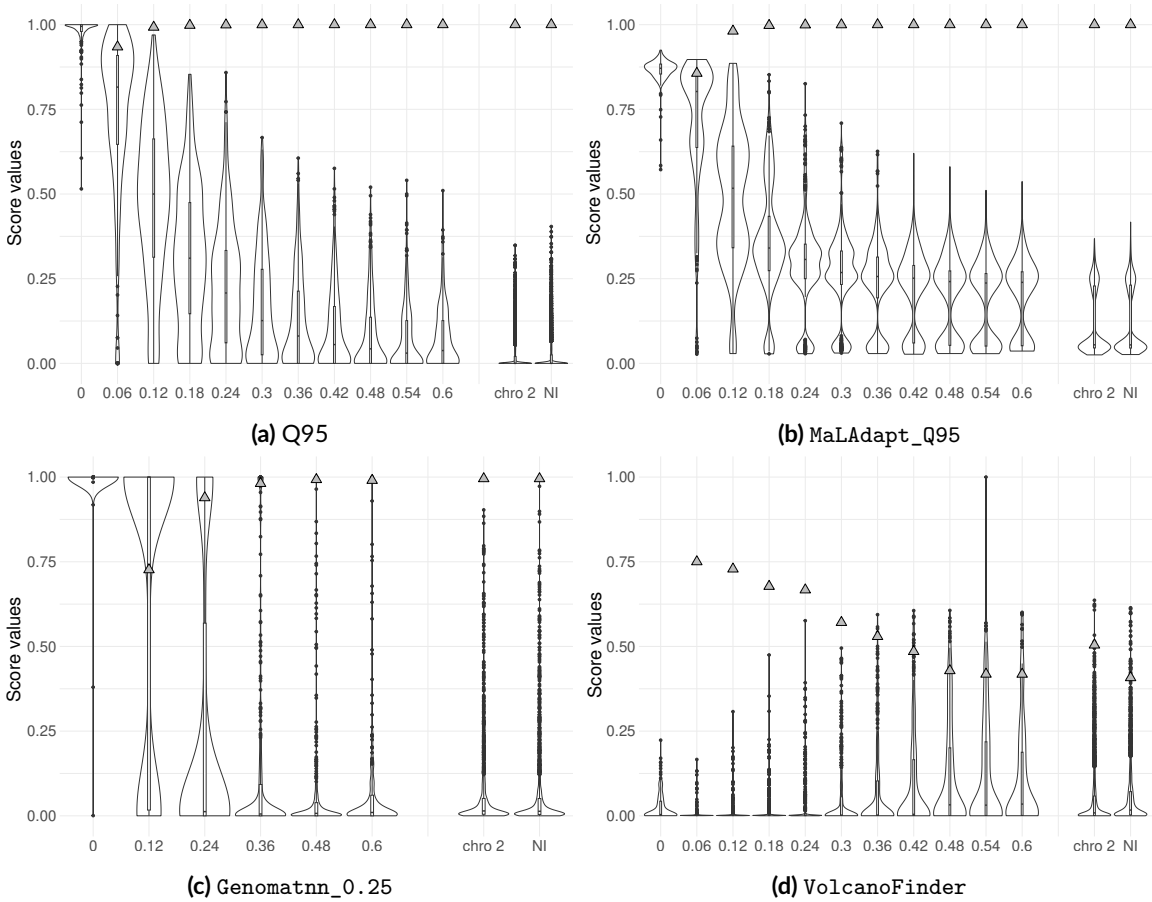
The scores for NI windows may be distinguishable or not from those of adjacent windows, when this is the case this should lead to an increase in FPR in presence of adjacent windows for methods trained on NI as non-AI window. The effect of hitchhiking is still visible in several conditions of simulation for adjacent windows, even for the most distant ones, causing these windows to have higher score values than the NI windows. For example, in case of strong selection (Fig. 3 and Fig. S2), the scores of adjacent windows are much higher than those of the windows under NI (except for `VolcanoFinder`), this leads to a drop in `Genomatnn`'s AUC values for adjacent windows compared to the NI windows (maximum AUC$= 0.55$ versus 1). On the other hand, for intermediate selection (Fig. 2 and Fig. S1), the effect of hitchhiking on the scores of adjacent windows is weaker and lead their score distributions to be close to the NI windows (for the adjacent windows on the edge of the chromosome). This lead methods in presence of distant adjacent windows to have similarly performance to those with NI windows.

Inter-chromosomal hitchhiking effect — classification of windows from the second chromosome

Comparing the scores from the second chromosome with scores along the first chromosome and from totally independent NI simulations shows that inter-chromosomal hitchhiking sometimes has a detectable but weak effect, only under strong selection. For weak and intermediate selection, and for all methods, scores are similar (and even sometimes slightly lower) for windows on the second chromosome than for totally independent NI windows, showing that the signals of AI due to the selection at the first chromosome do not extend to other chromosomes. This result is confirmed by the similar AUC values observed between test sets with the second chromosome and those with independent NI for the same combination of methods/selection strength.

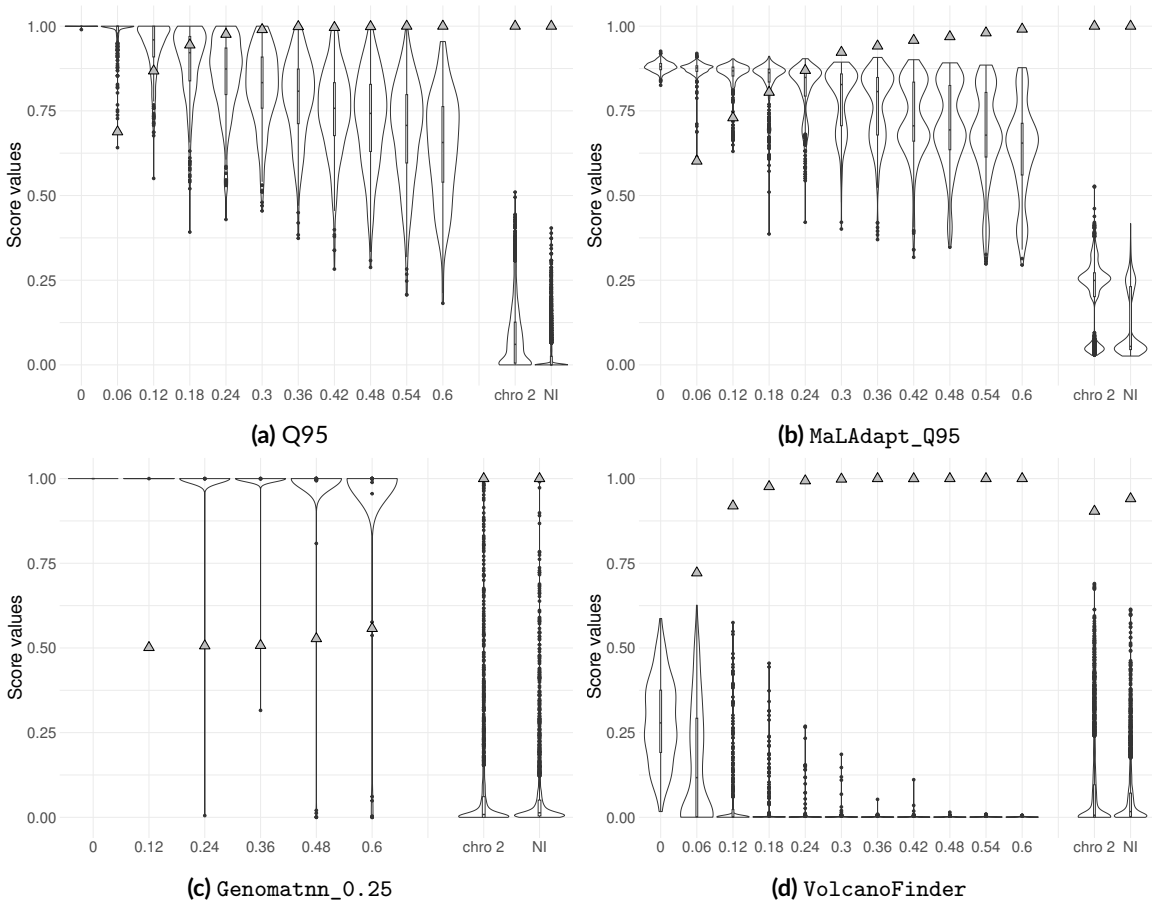
Under strong selection (Fig. 3 and Fig. S2), inter-chromosomal hitchhiking leads to increased score values for windows on the second chromosome for all methods. The strength of this increase and its impact on the method's performances varies according to the method. For `Genomatnn_0.05`, which appears to be the most affected (Fig. S2), many windows on the second

chromosome harbor high scores typical of AI windows rather than of independent NI windows. Those high score windows greatly reduce the performance of `Genomatnn_0.05` in the presence of second chromosome windows, compared to the case with NI windows only (AUC= 0.66 against 0.91). A reduction of performance due to higher score for the windows on the second chromosome is also observed for `VolcanoFinder`, but to a much lesser extent (0.94 against 0.90). For `MaLAdapt` and Q95, the maximum scores on the second chromosome are also higher than the maximum score for NI windows but still much lower than typical scores of AI windows. This weaker inter-chromosomal hitchhiking effect has no impact on their performances, which remains perfect in both cases (AUC = 1). For `Genomatnn_0.25`ś scores on the second chromosome are not affected by inter-chromosomal hitchhiking, showing similar scores distribution and AUC value between second chromosome windows and NI windows.



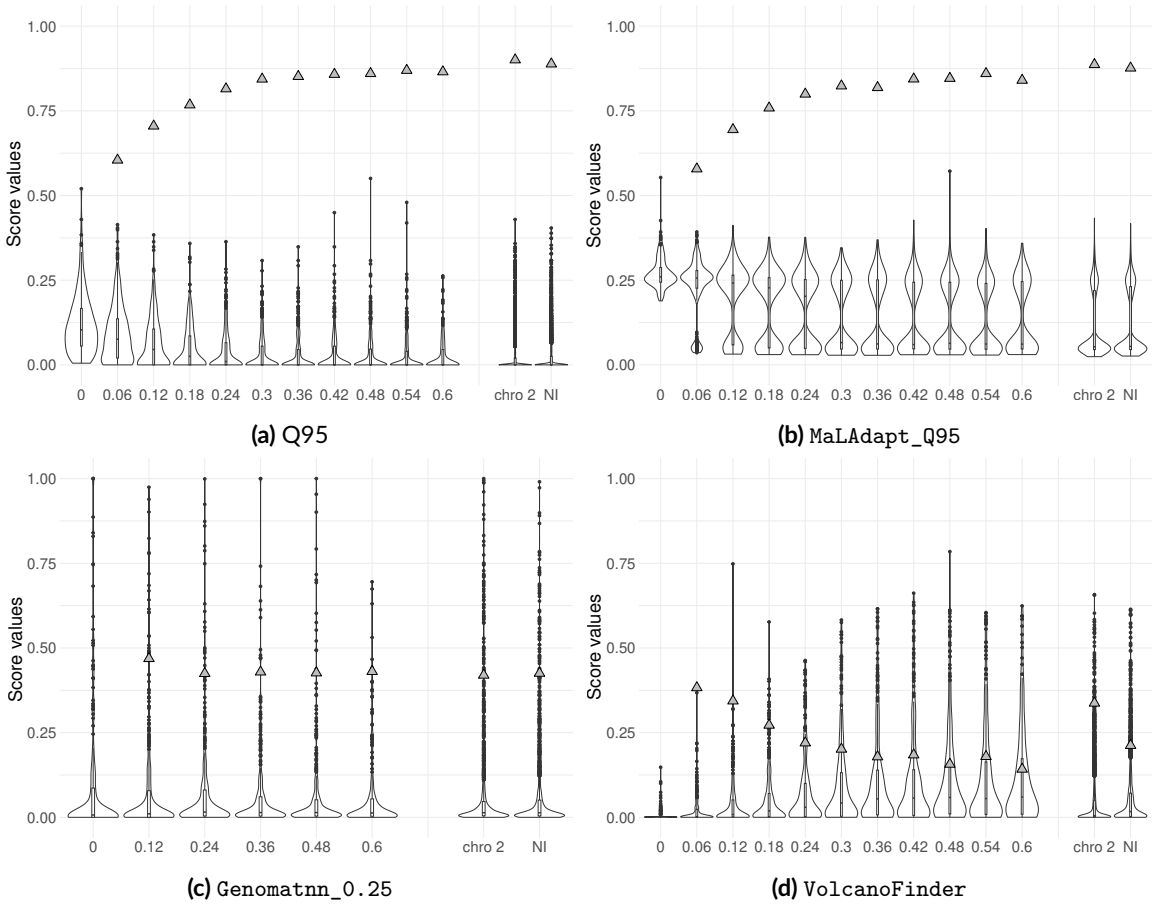
(a) Q95 **(b)** `MaLAdapt_Q95` **(c)** `Genomatnn_0.25` **(d)** `VolcanoFinder`

Figure 2 – Score values for the different windows considered under a scenario with intermediate selection strength ($s = 10^{-2}$). Windows from the first chromosome are represented in function to their distance (cM) to the site under selection (0 indicates the window containing it) followed by the windows from the second neutral chromosome ("Chro 2") and from a independently simulated neutral chromosome (NI) for reference. The areas under the ROC curves (AUC) for each test set containing AI windows and each class of non-AI windows are represented by the grey triangles.



(a) Q95 **(b)** `MaLAdapt_Q95` **(c)** `Genomatnn_0.25` **(d)** `VolcanoFinder`

Figure 3 – Same as figure 2 but for scenario with strong selection ($s = 10^{-1}$).



(a) Q95 (b) MaLAdapt_Q95 (c) Genomatnn_0.25 (d) VolcanoFinder

Figure 4 – Same as figure 2 but for a scenario with weak selection ($s = 10^{-3}$).

Discussion

Given our results, the use of Q95 as a genome scan statistic seems the most discriminating method, with high power for low false positive rates, even for the weakest strength of selection considered here. It also leads to the narrowest regions of adjacent windows giving positive scores due to hitchhiking, allowing to better pinpoint the window under selection. To the extent that this result is robust, it also appears quite convenient as this summary statistics is easy to compute without training a machine learning algorithm, in contrast to simulation-based classification methods. Moreover, this summary statistic does not require the use of phased data, unlike `MaLAdapt`, for example.

Our results are generally congruent with the results of the previous evaluations of these methods. Although `VolcanoFinder` can be very powerful to detect AI when the mutation under AI has recently reached fixation Setter et al., 2020, it showed little power in scenarios with incomplete sweep (Gower et al., 2021; Zhang et al., 2023). In our simulated test data, the mutation is generally not fixed under weak and intermediate selection, in which case we also find poor performance. The presence of spurious peaks on neutral regions near chromosome ends was also described in the original publication. The `Genomatnn` and `MaLAdapt` simulation-based methods showed better performance than `VolcanoFinder` in our study and in previous evaluations, with `MaLAdapt` being better than `Genomatnn` with lower false positive rate and more robust results

regarding hitchhiking (Zhang et al., 2023). Finally, Q95 also showed similar or even better performance than the other tested methods in previous comparisons (Gower et al., 2021; Racimo et al., 2017).

The mains differences between our results and previous works is that Zhang et al. (2023), and to a lesser extent, Gower et al. (2021) reported better performance for `MaLAdapt` (AUC values, Zhang et al., 2023) and `Genomatnn` (mcccf1 curve, Gower et al., 2021) compared to Q95. The performance of the two simulation-based classification methods, `MaLAdapt` and `Genomatnn`, by definition depends on the training set. The differences in performance observed between our tests and those in the articles describing these methods can thus be partly explained by the fact that we used the pre-trained ML models of the original publications, rather than ML models specificaly re-trained under the *Podarcis* scenario. Our test set generated under the *Podarcis* evolutionary scenario thus corresponds to a case of misspecification of the `MaLAdapt` and `Genomatnn`ś models. Gower et al. (2021) showed that the performance of `Genomatnn` decreases when using test sets from a different demographic model than the one considered for training, and Q95 performed better than their method in this case. `MaLAdapt`ś performance also dropped when the simulated test data were produced under parameter values of the `Genomatnn` train set. Comparison between our results and their results is difficult because Zhang et al. (2023) averaged the performance of the methods (*i.e.* the false positive rates) over different scenarios, making it impossible to tease apart the effect of different factors (strength of selection, hitchhiking, etc.). In addition, it is also difficult to identify the exact reason behind this discrepancy from our results because the models used by Zhang et al. (2023) to generate test sets differ substantially from ours. Indeed, their model includes past changes in population size, occurrence of deleterious mutations and genetic structure (intro/exon). Finally, Another possible cause, presented in the `MaLAdapt` tests, as a source of increased RPFs, is that `MaLAdapt` was trained on lower selection values than that corresponding to our strong selection scenario ($s \in [10^{-4}, 10^{-2}]$ and $s \in [10^{-3}, 10^{-1}]$ for `MaLAdapt` and our tests respectively), which might potentially explain the observed drop in its performance for non-AI adjacent windows with $s = 10^{-1}$.

Our results show that false positive rates increase when considering non-AI window types not taken into account in the train set. The original train set of `Genomatnn` considers only independent neutral introgressed windows. Accordingly, our results on independent NI windows show coherent low false positive rates, at levels similar or only slightly higher than for Q95, but considering adjacent windows leads to higher false positive rate. Unlike `Genomatnn`, `MaLAdapt` took into account adjacent windows in its train set. This may explain, in part, why `MaLAdapt`performs better than `Genomatnn` whatever the type of non-IA window considered.

Our results show the importance of considering different types of neutral potentially introgressed windows during training. The aim of all those classification methods is to specifically identify the window carrying the mutation under AI. It is therefore necessary to take into account windows that have undergone evolutionary processes that could lead to genetic signals similar to those of adaptive introgression. The evolutionary processes often taken into account in train sets are classic selective sweep and independent neutral introgression, but these processes are not the only ones that can leave confounding signals with AI in genomes. Depending on the strength of the selection applied to the mutation and the number of migrants, the effect of hitchhiking can leave an AI signal on adjacent windows linked to the mutation. As seen in our results, the effect of hitchhiking can even extend, to a lesser extent, to other chromosomes, *i.e.*

not carrying the mutation under AI. Other processes not included in the published methods and our tests, as for example the presence of structured populations in demographic model could lead to false introgression signals in the genome due to incomplete lineage sorting (Tournebize and Chikhi, 2023). It is therefore important to take these confounding processes into account in the train set or during robustness tests in order to quantify and possibly limit their impact on the classification performances.

The use of classification methods implies the need to define a threshold for classification scores. This threshold can be defined for a given train set but can be difficult to define when trying to control the error rate when proportions of positive cases in the analyzed set potentially differ from those in the train set. In this respect, binary classification problems for multiple windows in a genome are comparable to multiple hypothesis testing problems. The hypothesis tested here for each window is that there is no adaptive introgression within it. For genomic scan on multiple windows, it may be more specifically relevant to control the false discovery rate (FDR), *i.e.* the proportion of false positives among all positives detected.

Various strategies have been developed to control the FDR, such as methods based on the availability of p-values for each window (Benjamini and Hochberg, 1995), or empirical Bayes procedures requiring the distribution of some "score" for each window under the null hypothesis (Efron, 2010; Storey, 2003). All these methods require the distribution of classification scores (or p-values) under the null model to be known, so it may be necessary to run simulations under the null model, even for scoring methods that do not require training (such as Q95). Some of the FDR methods require independence among windows scores, which is not the case for our inferences because the presence of AI windows affects the distribution of scores in adjacent non-AI windows through hitchhiking. For this reason, an FDR method that takes data dependency into account would be required (Heller and Rosset, 2021; Tang and Zhang, 2007). Here it might be further necessary to control the error rate among non-AI windows even in the presence of some AI window(s) in a genome, which may raise distinct problems in controlling error rates.

For such reasons, it might be necessary to first infer a single measure of AI at the genomic level (rather that at the level of individual windows), such as an overall adaptive effect of AI representing a cumulative effect of adaptive mutations in different windows, and then to use windows-based criteria measures of adaptive effect to identify the windows likely to contribute to the genomic level measure of AI. It is possibly in this context that the windows-based methods evaluated in this paper could be best used.

Acknowledgements

We would like to thank Graham Gower for his availability, help and advice in using `Genomatnn`. We would also like to thank Xinjun Zhang for giving us permission to use her scripts and for answering all our questions about `MaLAdapt`. Finally, we would also like to thank Sylvain Mousset for his help and advice on using `VolcanoFinder`.

Fundings

This work was funded by the Agence Nationale de la Recherche (project INTROSPEC ANR-19-CE02-0011), by the Occitanie Regional Council's program "Key challenge BiodivOc" managed by the University of Montpellier (DevOCGen project), and by recurrent funding from INRAE

and CNRS. Part of this work was carried out by using the resources of the national INRAE MIGALE (Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) and GENOTOUL (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) bioinformatics HPC platforms, as well as the local Montpellier Bioinformatics Biodiversity (MBB, supported by the LabEx CeMEB ANR-10-LABX-04-01) and CBGP HPC platform services. This work also benefited from an ERJ (junior research team) grant to J.R. and G.C. by the LabEx CeMEB ANR-10-LABX-04-01.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. The authors declare the following non-financial conflict of interest: M.dN., R. L. and F.R. are recommenders for PCI Evol. Biol.

Data, script, code, and supplementary information availability

Script and codes are available online (https://doi.org/10.5281/zenodo.11448725)
Supplementary information is available online (https://doi.org/10.5281/zenodo.11520880)

References

Adavoudi R, Pilot M (2022). Consequences of Hybridization in Mammals: A Systematic Review. *Genes* **13**. 50. https://doi.org/10.3390/genes13010050.

Anderson E, Hubricht L (1938). Hybridization in Tradescantia. III. the Evidence for Introgressive Hybridization. *American Journal of Botany* **25**. 396–402. https://doi.org/10.1002/j.1537-2197.1938.tb09237.x.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, Gladstein AL, Gorjanc G, Guo B, Jeffery B, Kretzschumar WW, Lohse K, Matschiner M, Nelson D, Pope NS, Quinto-Cortés CD, et al. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* **220**, iyab229. https://doi.org/10.1093/genetics/iyab229.

Benjamini Y, Hochberg Y (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**. 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bergeron LA, Besenbacher S, Zheng J, Li P, Bertelsen MF, Quintard B, Hoffman JI, Li Z, St. Leger J, Shao C, Stiller J, Gilbert MTP, Schierup MH, Zhang G (2023). Evolution of the germline mutation rate across vertebrates. *Nature* **615**. 285–291. https://doi.org/10.1038/s41586-023-05752-y.

Bradley AP (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **30**, 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Burgarella C, Barnaud A, Kane NA, Jankowski F, Scarcelli N, Billot C, Vigouroux Y, Berthouly-Salazar C (2019). Adaptive Introgression: An Untapped Evolutionary Mechanism for Crop Adaptation. *Frontiers in Plant Science* **10**. 4. https://doi.org/10.3389/fpls.2019.00004.

Burke JM, Arnold ML (2001). Genetics and the Fitness of Hybrids. *Annual Review of Genetics* **35**. 31–52. https://doi.org/10.1146/annurev.genet.35.102401.085719.

Caeiro-Dias G, Brelsford A, Kaliontzopoulou A, Meneses-Ribeiro M, Crochet PA, Pinho C (2021). Variable levels of introgression between the endangered *Podarcis carbonelli* and highly divergent congeneric species. *Heredity* **126**. 463–476. https://doi.org/10.1038/s41437-020-00386-6.

Caeiro-Dias G, Brelsford A, Meneses-Ribeiro M, Crochet PA, Pinho C (2023). Hybridization in late stages of speciation: Strong but incomplete genome-wide reproductive isolation and 'large Z-effect' in a moving hybrid zone. *Molecular Ecology* **32**. 4362–4380. https://doi.org/10.1111/mec.17035.

Clark PU, Dyke AS, Shakun JD, Carlson AE, Clark J, Wohlfarth B, Mitrovica JX, Hostetler SW, McCabe AM (2009). The Last Glacial Maximum. *Science* **325**. 710–714. https://doi.org/10.1126/science.1172873.

Dabi A, Schrider D (2024). Population size rescaling significantly biases outcomes of forward-in-time population genetic simulations. https://doi.org/10.1101/2024.04.07.588318.

Edelman NB, Mallet J (2021). Prevalence and Adaptive Impact of Introgression. *Annual Review of Genetics* **55**. 265–283. https://doi.org/10.1146/annurev-genet-021821-020805.

Efron B (2010). Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. In: Institute of Mathematical Statistics monographs. Cambridge University Press, p. 276. https://doi.org/https://doi.org/10.1017/CBO9780511761362.

Gaczorek TS, Chechetkin M, Dudek K, Caeiro-Dias G, Crochet PA, Geniez P, Pinho C, Babik W (2023). Widespread introgression of MHC genes in Iberian *Podarcis* lizards. *Molecular Ecology* **32**. 4003–4017. https://doi.org/10.1111/mec.16974.

Geurts P, Ernst D, Wehenkel L (2006). Extremely randomized trees. *Machine Learning* **63**, 3–42. https://doi.org/10.1007/s10994-006-6226-1.

Gower G, Picazo PI, Fumagalli M, Racimo F (2021). Detecting adaptive introgression in human evolution using convolutional neural networks. *eLife* **10**. e64669. https://doi.org/10.7554/eLife.64669.

Grant PR, Grant BR (2019). Hybridization increases population variation during adaptive radiation. *Proceedings of the National Academy of Sciences* **116**. 23216–23224. https://doi.org/10.1073/pnas.1913534116.

Haller BC, Messer PW (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution* **36**, 632–637. https://doi.org/10.1093/molbev/msy228.

Harrison RG, Larson EL (2014). Hybridization, Introgression, and the Nature of Species Boundaries. *Journal of Heredity* **105**, 795–809. https://doi.org/10.1093/jhered/esu033.

Hedrick PW (2013). Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology* **22**. 4606–4618. https://doi.org/10.1111/mec.12415.

Heller R, Rosset S (2021). Optimal Control of False Discovery Criteria in the Two-Group Model. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **83**, 133–155. https://doi.org/10.1111/rssb.12403.

Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ (2007). Sequence-Level Population Simulations Over Large Genomic Regions. *Genetics* **177**, 1725–1731. https://doi.org/10.1534/genetics.106.069088.

Jo BS, Choi SS (2015). Introns: The Functional Benefits of Introns in Genomes. *Genomics & Informatics* **13**. 112–118. https://doi.org/10.5808/GI.2015.13.4.112.

Kaliontzopoulou A, Phino C, Harris DJ, Carretero MA (2011). When cryptic diversity blurs the picture: a cautionary tale from Iberian and North African *Podarcis* wall lizards. *Biological Journal of the Linnean Society* **103**, 779–800. https://doi.org/10.1111/j.1095-8312.2011.01703.x.

Kelleher J, Etheridge AM, McVean G (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology* **12**. e1004842. https://doi.org/10.1371/journal.pcbi.1004842.

Kelly JK (1997). A Test of Neutrality Based on Interlocus Associations. *Genetics* **146**, 1197–1206. https://doi.org/10.1093/genetics/146.3.1197.

Kim SC, Rieseberg LH (1999). Genetic Architecture of Species Differences in Annual Sunflowers: Implications for Adaptive Trait Introgression. *Genetics* **153**, 965–977. https://doi.org/10.1093/genetics/153.2.965.

Liu S, Zhang L, Sang Y, Lai Q, Zhang X, Jia C, Long Z, Wu J, Ma T, Mao K, Street NR, Ingvarsson PK, Liu J, Wang J (2022). Demographic History and Natural Selection Shape Patterns of Deleterious Mutation Load and Barriers to Introgression across Populus Genome. *Molecular Biology and Evolution* **39**, msac008. https://doi.org/10.1093/molbev/msac008.

Mallet J (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution* **20**, 229–237. https://doi.org/10.1016/j.tree.2005.02.010.

Miles A, Rodrigues MF, Ralph P, Kelleher J, Schelker M, Pisupati R, Rae S, Millar T (2024). cggh/scikit-allel: v1.3.8. https://doi.org/10.5281/zenodo.10876220.

Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, Joron M, McMillan WO, Jiggins CD (2012). Adaptive Introgression across Species Boundaries in *Heliconius* Butterflies. *PLOS Genetics* **8**. e1002752. https://doi.org/10.1371/journal.pgen.1002752.

Pawar H, Rymbekova A, Cuadros-Espinoza S, Huang X, Manuel M, Valk T, Lobon I, Alvarez-Estape M, Haber M, Dolgova O, Han S, Esteller-Cucala P, Juan D, Ayub Q, Bautista R, Kelley JL, Cornejo OE, Lao O, Andrés AM, Guschanski K, et al. (2023). Ghost admixture in eastern gorillas. *Nature Ecology & Evolution* **7**. 1503–1514. https://doi.org/10.1038/s41559-023-02145-2.

Pinho C, Harris DJ, Ferrand N (2007). Contrasting patterns of population subdivision and historical demography in three western Mediterranean lizard species inferred from mitochondrial DNA variation. *Molecular Ecology* **16**. 1191–1205. https://doi.org/10.1111/j.1365-294X.2007.03230.x.

Pinho C, Kaliontzopoulou A, Carretero MA, Harris DJ, Ferrand N (2009). Genetic admixture between the Iberian endemic lizards *Podarcis bocagei* and *Podarcis carbonelli*: evidence for limited natural hybridization and a bimodal hybrid zone. *Journal of Zoological Systematics and Evolutionary Research* **47**. 368–377. https://doi.org/10.1111/j.1439-0469.2009.00532.x.

Pinho C, Kaliontzopoulou A, Harris DJ, Ferrand N (2011). Recent evolutionary history of the Iberian endemic lizards *Podarcis bocagei* (Seoane, 1884) and *Podarcis carbonelli* Pérez-Mellado, 1981 (Squamata: Lacertidae) revealed by allozyme and microsatellite markers. *Zoological Journal of the Linnean Society* **162**, 184–200. https://doi.org/10.1111/j.1096-3642.2010.00669.x.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**. 43–49. https://doi.org/10.1038/nature12886.

Racimo F, Marnetto D, Huerta-Sánchez E (2017). Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. *Molecular Biology and Evolution* **34**, 296–317. https://doi.org/10.1093/molbev/msw216.

Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics* **16**. 359–371. https://doi.org/10.1038/nrg3936.

Ralph P, Thornton K, Kelleher J (2020). Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. *Genetics* **215**, 779–797. https://doi.org/10.1534/genetics.120.303253.

Renoult JP, Geniez P, Bacquet P, Benoit L, Crochet PA (2009). Morphology and nuclear markers reveal extensive mitochondrial introgressions in the Iberian Wall Lizard species complex. *Molecular Ecology* **18**. 4298–4315. https://doi.org/10.1111/j.1365-294X.2009.04351.x.

Rosser N, Seixas F, Queste LM, Cama B, Mori-Pezo R, Kryvokhyzha D, Nelson M, Waite-Hudson R, Goringe M, Costa M, Elias M, Mendes Eleres de Figueiredo C, Freitas AVL, Joron M, Kozak K, Lamas G, Martins ARP, McMillan WO, Ready J, Rueda-Muñoz N, et al. (2024). Hybrid speciation driven by multilocus introgression of ecological traits. *Nature* **628**. 811–817. https://doi.org/10.1038/s41586-024-07263-w.

Runemark A, Vallejo-Marin M, Meier JI (2019). Eukaryote hybrid genomes. *PLOS Genetics* **15**. e1008404. https://doi.org/10.1371/journal.pgen.1008404.

Setter D, Mousset S, Cheng X, Nielsen R, DeGiorgio M, Hermisson J (2020). VolcanoFinder: Genomic scans for adaptive introgression. *PLOS Genetics* **16**. e1008867. https://doi.org/10.1371/journal.pgen.1008867.

Smith JM, Haigh J (1974). The hitch-hiking effect of a favourable gene. *Genetics Research* **23**, 23–35. https://doi.org/10.1017/S0016672300014634.

Storey JD (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* **31**. 2013–2035. https://doi.org/10.1214/aos/1074290335.

Suarez-Gonzalez A, Lexer C, Cronk QCB (2018). Adaptive introgression: a plant perspective. *Biology Letters* **14**. 20170688. https://doi.org/10.1098/rsbl.2017.0688.

Tang W, Zhang CH (2007). Empirical Bayes methods for controlling the false discovery rate with dependent data. In: *Institute of Mathematical Statistics Lecture Notes - Monograph Series*. Institute of Mathematical Statistics, pp. 151–160. https://doi.org/10.1214/074921707000000111.

Taylor SA, Larson EL (2019). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology & Evolution* **3**. 170–177. https://doi.org/10.1038/s41559-018-0777-y.

Todesco M, Pascual MA, Owens GL, Ostevik KL, Moyers BT, Hübner S, Heredia SM, Hahn MA, Caseys C, Bock DG, Rieseberg LH (2016). Hybridization and extinction. *Evolutionary Applications* **9**. 892–908. https://doi.org/10.1111/eva.12367.

Tournebize R, Chikhi L (2023). Questioning Neanderthal admixture: on models, robustness and consensus in human evolution. https://doi.org/10.1101/2023.04.05.535686.

Twyford AD, Ennos RA (2012). Next-generation hybridization and introgression. *Heredity* **108**. 179–189. https://doi.org/10.1038/hdy.2011.68.

Wang Y, Wang Y, Cheng X, Ding Y, Wang C, Merilä J, Guo B (2023). Prevalent Introgression Underlies Convergent Evolution in the Diversification of Pungitius Sticklebacks. *Molecular Biology and Evolution* **40**, msad026. https://doi.org/10.1093/molbev/msad026.

Yang W, Feiner N, Pinho C, While GM, Kaliontzopoulou A, Harris DJ, Salvi D, Uller T (2021). Extensive introgression and mosaic genomes of Mediterranean endemic lizards. *Nature Communications* **12**. 2762. https://doi.org/10.1038/s41467-021-22949-9.

Zhang X, Kim B, Singh A, Sankararaman S, Durvasula A, Lohmueller KE (2023). MaLAdapt Reveals Novel Targets of Adaptive Introgression From Neanderthals and Denisovans in Worldwide Human Populations. *Molecular Biology and Evolution* **40**, msad001. https://doi.org/10.1093/molbev/msad001.