Received: 4 March 2020 | Revised: 5 October 2020 | Accepted: 9 October 2020

DOI: 10.1111/1755-0998.13286

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

# Biased assessment of ongoing admixture using STRUCTURE in the absence of reference samples

Sara Ravagni | Ines Sanchez-Donoso | Carles Vilà

Conservation and Evolutionary Genetics Group, Doñana Biological Station (EBD-CSIC), Seville, Spain

**Correspondence**
Carles Vilà, Conservation and Evolutionary Genetics Group, Doñana Biological Station (EBD-CSIC), Avd Americo Vespucio 26, 41092 Seville, Spain.
Email: carles.vila@ebd.csic.es

**Funding information**
Ministerio de Economía, Industria y Competitividad, Gobierno de España, Grant/Award Number: BES-2017-081291 and CGL2016-75227-P

## Abstract

Detection of hybridization and introgression is important in ecological research as in conservation and evolutionary biology. STRUCTURE is one of the most popular software to study introgression and allows estimating what proportion of the genome of each individual belongs to each ancestral population, even in cases where no reference sample from the ancestral nonadmixed populations is previously identified. In spite of its frequent use, some studies have indicated that ancestry estimates may not always be reliable. We simulated population data under different conditions with regard to the genetic differentiation between ancestral populations, number of loci considered, number of alleles per marker and hybridization rate, and analysed data with STRUCTURE. When reference samples were not included, the comparison of the known degree of admixture for each simulated individual and the value estimated with STRUCTURE revealed a strong underestimation of the level of introgression, classifying many admixed individuals as nonadmixed. This derives from an inaccurate estimation of the ancestral allele frequencies. When samples from the nonadmixed ancestral population were included as reference in the analyses, the bias in the estimations was reduced. The most accurate estimates were obtained when potentially admixed samples were few in relation to reference samples. Thus, whenever possible, a very large proportion of nonadmixed reference samples should be included in admixture assessments and different approaches should be combined. The misestimate of the amount of introgression can impair our understanding of the evolutionary history of species and misguide conservation efforts.

KEYWORDS

Bayesian clustering, hybrid zone, hybridization, introgression, population admixture, simulation

## 1 | INTRODUCTION

Hybridization is a major concern in conservation biology (Rhymer & Simberloff, 1996) but also a source of evolutionary innovation (Abbott et al., 2013; Arnold, 2015) and adaptive introgression (the introduction of genes from one evolutionary lineage into the gene pool of another) has been shown to play an important role in the evolution and diversification of different clades (Burgarella et al., 2019; Hamilton & Miller, 2016; Oziolor et al., 2019; Suarez-Gonzalez et al., 2018). Thus, the identification of admixed individuals and the assessment of the levels of introgression are fundamental steps to study the effect of hybridization on populations. However, the detection of these admixed individuals can be problematic when the parental species show limited phenotypic differentiation (Randi, 2008). The ineffectiveness of morphological criteria in differentiating cryptic hybrids or admixed individuals (Oliveira et al., 2008) promoted the use of highly polymorphic genetic markers, such as microsatellites or SNPs. The development of software programs based

  

on the analysis of panels of genetic markers and the implementation of Bayesian clustering models has facilitated the detection of hybrids and the assessment of introgression of genes of one species into the gene pool of another.

One of the most widely used programs for these analyses is STRUCTURE (Pritchard et al., 2000) which allows identification of the origin of the genome of each individual. After choosing a K value, i.e., the number of populations, STRUCTURE subdivides the sample into K different clusters - trying to minimize departures from Hardy-Weinberg equilibrium and linkage disequilibrium - and estimates for each individual the proportion of the genome that could originate from each cluster (Barilani et al., 2007). In this way, STRUCTURE is very efficient at separating populations and identifying admixed individuals. A Scopus search (on 26 June 2019) for articles citing the original paper of Pritchard et al. (2000) and containing the words 'hybrid*' or 'introgression' in the title, abstract or keywords returned a total of 3,446 articles, highlighting the popularity of the software to detect hybridization and population admixture.

However, the power of STRUCTURE to accurately estimate the amount of introgression has been questioned as it could underestimate the number of admixed individuals (Randi, 2008; Sanchez-Donoso et al., 2014; Sanz et al., 2009). The main limitation seems to be related to the detection of old admixture as opposed to F1 hybrids and first generations backcrosses (Oliveira et al., 2008). It has been suggested that estimates of the amount of introgression could be improved by increasing the number of loci (Pritchard et al., 2000; Randi, 2008; Vähä & Primmer, 2006) and using both linked and unlinked markers (Lecis et al., 2006), as this could enable a better assignment of admixed individuals in separate genotypic classes and the identification of past events of hybridization. The use of high number of loci derived from genome-wide data increases resolution in the detection of varying levels of introgression by identifying regions of the genome of different origin (Gómez-Sánchez et al., 2018). Unfortunately, this is not always feasible due to the lack of a reference genome, or due to a trade-off between costs and number of specimens to analyse. In order to characterize population variability, large numbers of samples may need to be studied, making genomic approaches unaffordable and, despite technical advances, studies with reduced number of loci continue to be frequent in day-to-day studies of hybridization and admixture (for example, see Alacs et al., 2010; Arias et al., 2019; De Barba et al., 2017; Sujii et al., 2019). In addition, large numbers of unmapped markers are not necessarily an advantage for the study of ongoing hybridization because they could yield redundant information if they are in complete linkage.

Another factor affecting the performance of STRUCTURE is the sampling scheme, i.e. whether or not samples of each parental population or species are similar in size, and the phylogenetic relationships between the two species (Neophytou, 2014; Puechmaille, 2016). A previous study using empirical data of populations with known ancestry showed that STRUCTURE outperforms other common approaches in the identification of admixed individuals (Bohling et al., 2013). However, this study emphasizes the importance of including in the analyses a portion of individuals that can be a priori diagnosed as nonadmixed for the two parental classes and this may not always be feasible. In many studies genetic analyses are carried out to identify potentially admixed individuals without previously defining reference samples (for example, see Godinho et al., 2011; Muñoz-Fuentes et al., 2007; Oliveira et al., 2008; Ortego et al., 2017; Trigo et al., 2013). This can be particularly important in cases with high levels of hybridization and persistent introgression throughout the distribution range, or in cases where introgression takes place in geographically structured populations for which reference samples from another population may not be appropriate (for example, see Glover et al., 2017; Lavretsky et al., 2019; Sullivan et al., 2016).

In this study we used simulations to determine the accuracy of STRUCTURE in the estimation of the individual level of introgression when nonadmixed reference individuals were not available. We assessed the importance of the number and polymorphism of the loci used, the divergence between the ancestral populations and different rates of hybridization between them. We also studied how adding appropriate reference samples impacts the reliability of the estimates. Our goal was to identify ways to improve the accuracy of the estimates of the degree of introgression.

## 2 | MATERIALS AND METHODS

We carried out simulations of asymmetric gene flow from one population to another to reduce complexity and to facilitate analyses because the allele frequencies for one of the populations would not change over time. However, this is not an uncommon situation. Some examples of this kind of gene flow are the hybridization between domestic and vulnerable wild species (Godinho et al., 2011), admixture between a rare species and an abundant one (An et al., 2017), restocking of game species with farmed animals of alien origin (Sanchez-Donoso et al., 2014), or directional gene flow as a result of the biology of the species hybridizing (Muñoz-Fuentes et al., 2007).

### 2.1 | Simulation of ancestral populations

We used the software EASYPOP v. 2.0.1 (Balloux, 2001) to simulate pairs of diploid random mating populations, genotyped for 200 unlinked loci assuming different levels of polymorphism, with (a maximum of) either two, five or 10 alleles per locus. Mutation rate for markers with five or 10 alleles was set at $10^{-3}$ and for markers with two alleles at $10^{-8}$, as commonly assumed for microsatellites and SNPs (Drake et al., 1998; Ellegren, 2004; Payseur & Nachman, 2000). We assumed a mutation model (K-allele model: KAM) for markers with two alleles: each allele had the same probability to mutate to the other allelic state. We used a mixed model including single step mutation model (SSM) for markers with up to five or 10 alleles, with a proportion of 0.3 of KAM events (Ellegren, 2004). We generated between 700 and 5,000 individuals per generation and ran the simulations for at least 1,000 generations to assume a long time of separation between the populations and approach mutation-drift

equilibrium and stable differentiation. We generated 100 pairs of populations for each kind of marker and with genetic differentiation, measured as $F_{ST}$, around 0.05, 0.1 and 0.2. Thus, we simulated 900 pairs of populations genotyped for 200 loci each (3 kinds of markers × 3 levels of differentiation × 100 replicates = 900 pairs of populations). Parameters used for the runs are reported in Table S1. For details of the simulations see Supporting Information, Figures S1 and S2.

## 2.2 | Simulation of allele introgression

We wrote a script in Python 2.7 to simulate different hybridization rates and to subsample a random subset of individuals from each population to be analysed with STRUCTURE (see Figure 1 for a schematic representation of the simulations). Population A was where admixture took place due to some individuals arriving from population B. The goal of the analyses was to assess if it was possible to correctly estimate the degree of introgression in individuals sampled from the population A using STRUCTURE. Hybridization rate, that is the proportion of breeders originating from population B that contributed to the offspring of population A every generation, was around 1% (0.01) or 5% (0.05). Although high, rates of hybridization this high have been described for diverse taxa (for example, see Lavretsky et al., 2019; Muñoz-Fuentes et al., 2007; Nussberger et al., 2014). A key factor here is that we assume recurrent hybridization every generation and that admixed individuals do not have a reduced fecundity so that they are able to freely interbreed with other individuals in population A. To initiate the simulations (at generation 0), the program randomly selected 1,000 individuals from populations A and B from one of the 100 initial pairs of populations generated with Easypop for a given combination of $F_{ST}$ and type of marker.

For 10 generations, the programme generated the same number of individuals (1,000) by selecting two parents from the previous

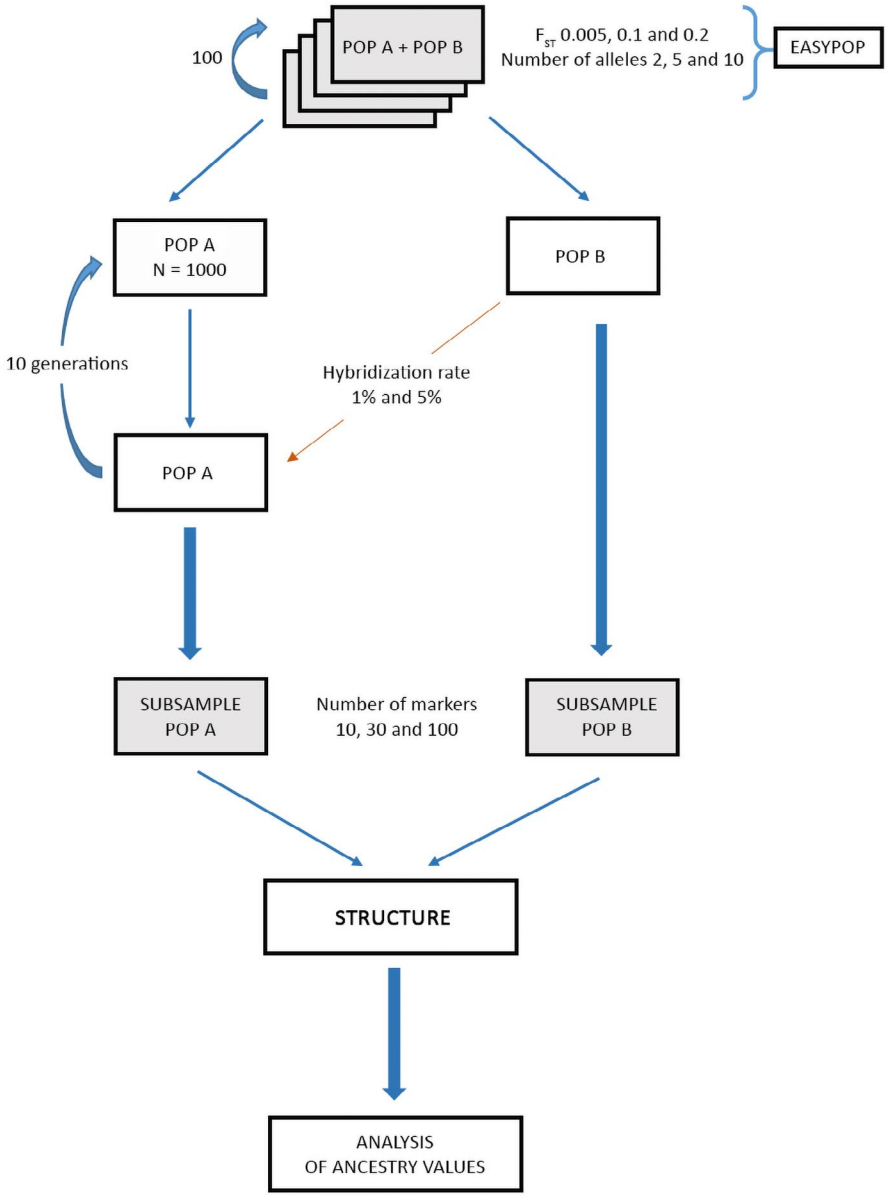


**FIGURE 1** Pipeline to simulate introgression and subsampling for STRUCTURE analyses. First, 100 pairs of populations for each set of cases (pairwise $F_{ST}$ = 0.05, 0.1 or 0.2, typed at 200 marker loci with two, five or 10 alleles) were simulated with Easypop. For 10 generations in population A, we randomly selected pairs of genotypes from the previous generation in population A or from population B according to the hybridization rate, and generated genotypes of 1,000 offspring. Finally, we randomly subsampled 100 individuals from each population A and B (with genotypes for 10, 30 or 100 randomly chosen loci) and genotypes were analysed in STRUCTURE. From the output of this program, we extracted estimates of ancestry and compared to the real values derived from the proportion of ancestry from population A in the simulations (see text) [Colour figure can be viewed at wileyonlinelibrary.com]

generation. Each parent was randomly selected from population A with a probability of (1 − m), where m was the hybridization rate (0.01 or 0.05); otherwise, the parent was randomly chosen from population B (geneflow was unidirectional from B to A). The genotype of the offspring was obtained by randomly selecting one of the two alleles from each parent at each locus with equal probability; loci were independent. For each individual, we calculated the proportion of the genome belonging to population A ($q_{real}$): for individuals originating from population B, this proportion was $q_{real} = 0.0$; for individuals from population A at generation 0, this proportion was $q_{real} = 1.0$; for subsequent generations, the proportion was the average of the values of the two parents. To confirm that the script was working as expected, we calculated the level of introgression expected after 10 generations in the different scenarios as in Verdu and Rosenberg (2011). Indeed, the value corresponded to the average level of introgression per individual in population A.

At the tenth generation, a random subsample of 100 individual genotypes was taken both from population B and from the introgressed population A to be analysed with STRUCTURE. In these subsamples, the number of loci was reduced to 10, 30 or 100 loci chosen at random. The data were used to generate an input file for STRUCTURE. As a result, 5,400 runs of STRUCTURE were carried out with 200 genotypes each (100 from the admixed population A and 100 from population B): three different values of $F_{ST}$ (0.05, 0.1 and 0.2) × three kinds of markers (two, five or 10 alleles) × three numbers of loci (10, 30 and 100) × two hybridization rates (1% and 5%) × 100 replicates (100 pairs of populations simulated in Easypop for each set of conditions with regard to $F_{ST}$ and type and number of loci).

The script for these simulations is available at https://github.com/sararvg/introgression_structure. This script was slightly modified for subsequent analyses.

## 2.3 | Analysis of simulated data sets

The simulated data were analysed in STRUCTURE v. 2.3.4 under the admixture model, as each individual may have ancestry in both initial populations, with correlated allele frequencies. We also carried out about 15% of the STRUCTURE runs under the independent allele frequency model but results were practically identical (data not shown) and we decided to focus on the first model. Analyses were run without population or location information, i.e., with the options USEPOPINFO and LOCPRIOR set to 0, in order to allow assignments based only on genetic information. INFERALPHA was set to 1 to let STRUCTURE infer α (the relative admixture levels between populations) from the data. K was set to 2 to try to separate the two initial populations. After visually confirming that this was enough for convergence, runs were carried out using 30,000 burnin steps followed by 100,000 iterations of MCMC, with only one replicate for each data set. We prepared a script in Python to extract the estimated proportion of the individual's genome corresponding to population A ($q_{STRUCTURE}$) from the STRUCTURE output.

We graphically compared $q_{real}$ and $q_{STRUCTURE}$ with the package ggplot2 (Wickham, 2009) using R v. 4.0.2 (R Core Team, 2020) in RSTUDIO v. 1.3.959 (RStudio Team, 2020). All statistical analyses were carried out using the same versions of R and RSTUDIO. We tested if $q_{STRUCTURE}$ estimates were significantly higher than $q_{real}$ values by performing a Wilcoxon signed rank test after excluding individuals from population B, with the function wilcox.test() from the coin package (v.1.3.1, Hothorn et al., 2008). We tested for a linear relationship between $q_{real}$ and $q_{STRUCTURE}$ for individuals sampled in population A through linear regressions with the function lm(). We visually examined the normal distribution of the residuals of the regressions and only reported results for the cases in which this requirement was fulfilled. In these cases, we used generalized linear models to test if the absolute difference between $q_{real}$ and $q_{STRUCTURE}$ depended on $F_{ST}$ and on the number of alleles, both included in the model as explanatory variables. We run these models under a beta distribution with the function betareg() from the betareg package (v. 3.1.2, Zeileis et al., 2012). As the response variable included 0 and 1, we applied the transformation suggested by Smithson and Verkuilen (2006): $y' = [(y * (n - 1) * 0.5)/n]$, where $n$ is the sample size. We visually checked the models for homoscedasticity and normality of the residuals.

Potential over- or underestimation of the proportion of genome belonging to population A could be related to an inaccurate estimation of the ancestral allele frequencies for populations A and B. To visualize changes in the allele frequency estimates, we carried out 100 additional runs of STRUCTURE sampling individuals at generations 0, three, six and 10 of the simulations (this was done for a single case, i.e., $F_{ST} = 0.1$, 30 loci with 10 alleles and hybridization rate of 5%). From the output, we extracted allele frequencies estimated for all loci for the ancestral populations A and B. The matrices corresponding to both populations were compared to the true allele frequencies calculated from the populations generated by Easypop, before admixture started. For the comparison we used a distance calculated as the sum of the squared differences between each pair of allele frequencies (frequency for one allele at one locus in the ancestral population minus the frequency estimated by STRUCTURE), divided by the number of loci.

To evaluate if ancestry estimates obtained with STRUCTURE improved with the inclusion of reference samples, we carried out additional simulations for four cases ($F_{ST} = 0.1$, 30 loci with five and 10 alleles, and both hybridization rates) but now the sample for population A included 10% or 30% of individuals from the ancestral population (generation 0). We run the analyses without providing information about the locality or providing this information (USEPOPINFO option, PopFlag was set to 1 only for individuals belonging to the ancestral population A to use them as reference). For the same cases we also tested the effect of activating the POPALPHAS option to infer α for each population separately, which is suggested in cases of strong asymmetric admixture in the STRUCTURE manual. We tested with generalized linear models how the differences between $q_{real}$ and $q_{STRUCTURE}$ for the samples of the admixed population A were affected by the proportion of samples

used as a reference (0%, 10% or 30%), number of alleles (five or 10) and the use of the USEPOPINFO and POPALPHAS options.

We also assessed the reliability of estimates obtained with STRUCTURE when almost all samples included in the run were used as reference and the introgression was assessed in just a few target individuals. We selected 100 individuals representing the full range of $q_{real}$ values (evenly distributed from 0 to 1). Twenty groups of five of these individuals were randomly selected without replacement and analysed together with 100 samples from A before admixture and 100 samples from B (for $F_{ST} = 0.1$) using 30 or 100 markers of five or 10 alleles. This process was repeated 10 times. The values of $q_{real}$ and $q_{STRUCTURE}$ were then compared for the target individuals.

To confirm the importance of using appropriate reference samples, we analysed data from a natural population. We used a data set of wild common quails (*Coturnix coturnix*) and game farm quails from Sanchez-Donoso et al., (2014), genotyped at nine autosomal microsatellite loci. A previous study had shown that game farm quails used for restocking were a genetically diagnosable mix of common and Japanese quails (*C. japonica*; Sanchez-Donoso et al., 2012). We randomly selected genotypes from 100 wild quails from NE Spain obtained from 2007–2010 and analysed them in STRUCTURE together with 52 quails from game farms to assess the impact of the restocking on the natural populations. Afterwards, 10 samples collected in the same area in 1996–1997, before most releases of farm quails for hunting, were added as reference common quails free of introgression. The analyses were conducted using POPINFO and POPALPHAS. We tested if STRUCTURE suggested lower degree of admixture in the wild population in the absence of reference samples by performing a Wilcoxon signed rank test after excluding farm and reference individuals.

Finally, in order to assess if the biases detected when analysing the data with STRUCTURE were common to other programs also used to assess introgression, we also compared $q_{real}$ to estimates of $q$ obtained with ADMIXTURE v. 1.3.0 (Alexander & Lange, 2011), OHANA v.1.0 (Cheng et al., 2017) and SNMF v. 2.0 (Frichot et al., 2014). We used default parameters and K was set to 2. For Ohana, the maximum number of steps was set to 130,000 to simulate the iterations used in STRUCTURE and for sNMF alpha was set to 0.5. Since some of these programs are designed for analyses of markers with two alleles (SNPs), the comparisons were restricted to the cases of $F_{ST} = 0.1$, 100 loci with two alleles and hybridization rates of 1% and 5%.

## 3 | RESULTS

We graphically compared the proportion of the genome coming from population A as estimated by STRUCTURE ($q_{STRUCTURE}$) with the real values ($q_{real}$, Figure 2). Each plot represents 100 runs of STRUCTURE with 200 genotypes, resulting in 20,000 pairs of values of $q_{STRUCTURE}$ and $q_{real}$. Ideally, if the estimates of $q$ by STRUCTURE precisely corresponded to the real values, all points should fall on the diagonal of the diagrams. As expected, increasing the number of loci and the number of alleles per marker improved precision (reduction in the variance) in the estimates of $q_{STRUCTURE}$ and, to a lesser degree, it also improved accuracy (similarity between $q_{STRUCTURE}$ and $q_{real}$ values). At the same time, comparing cases with the same number of loci and alleles per marker, $q_{STRUCTURE}$ values showed lower variance as the divergence level between the hybridizing populations increased (e.g., Figures S3a versus S4a). However, the comparison of $q_{STRUCTURE}$ and $q_{real}$ revealed systematic biases.

Considering a hybridization rate of 1% (Figure 2a, S3a and S4a), for markers with two alleles, $q_{STRUCTURE}$ estimates tended to be quite independent from $q_{real}$, forming scattered clouds of points (meaning that STRUCTURE provided a poor assessment of the ancestry), except when the number of loci was 100 and the genetic differentiation was strong (Figure S4a). When the number of alleles per locus was 5, estimations using 10 or 30 loci were not reliable either and had a large variance. The consistency of the estimates increased notably with the use of 100 loci when $F_{ST}$ was 0.1 or 0.2 (Figure 2a and S4a). In the cases of markers with 10 alleles, in general, estimations started to improve from 30 loci for all values of $F_{ST}$, with reduced variance in $q_{STRUCTURE}$ values. However, even in the cases with the lowest variance, $q_{STRUCTURE}$ tended to be larger than $q_{real}$ (Figure 2a).

With a hybridization rate of 5% (Figure 2b, S3b and S4b), the entire population A was admixed and no pure individuals were left, i.e., no individuals whose genomes derived solely from the ancestral population; the population had turned into a hybrid swarm. Although $q_{real}$ values were always smaller than 0.85, $q_{STRUCTURE}$ values tended to be higher, identifying many admixed individuals as nonadmixed ($q_{STRUCTURE}$ close to 1) and therefore underestimating the magnitude of the introgression in the population (Figure 2b, S3b and S4b).

The Wilcoxon signed rank test confirmed that $q_{STRUCTURE}$ values were significantly higher than the corresponding $q_{real}$ ($p < 10^{-16}$) for individuals from population A, except in five cases in which $q_{STRUCTURE}$ was practically uninformative (Figure 2a: $F_{ST} = 0.1$, hybridization rate of 1%, 10 markers with two alleles; Figure S3a: $F_{ST} = 0.05$, hybridization rate of 1%, 10 markers with two and five alleles, as well as 30 markers with two alleles; Figure S3b: $F_{ST} = 0.05$, hybridization rate of 5%, 10 markers with two alleles). This implies that there was a tendency to overestimate the proportion of the genome from the ancestral population in practically all cases ($q_{STRUCTURE} > q_{real}$). A significant linear relationship between $q_{real}$ and $q_{STRUCTURE}$ was found in nine cases, when 100 loci with five alleles (only for a high hybridization rate of 5%) or 10 alleles were used (Figure S5), and in all nine cases the regression line was above the diagonal (intercept significantly higher than 0, $p < 10^{-16}$). The fact that in the other cases no relationship could be found between $q_{STRUCTURE}$ and $q_{real}$ highlights the limited power of analyses with reduced number of loci and alleles.

We used generalized linear models to assess both the effect of the degree of differentiation between the ancestral populations ($F_{ST}$) and the number of alleles per marker on the absolute difference between $q_{STRUCTURE}$ and $q_{real}$ (the bias in the inference of $q$) for the two hybridization rates and considering 100 loci. Both variables proved to have a highly significant effect ($p < 10^{-16}$). The increase in genetic differentiation showed the stronger effect in reducing the

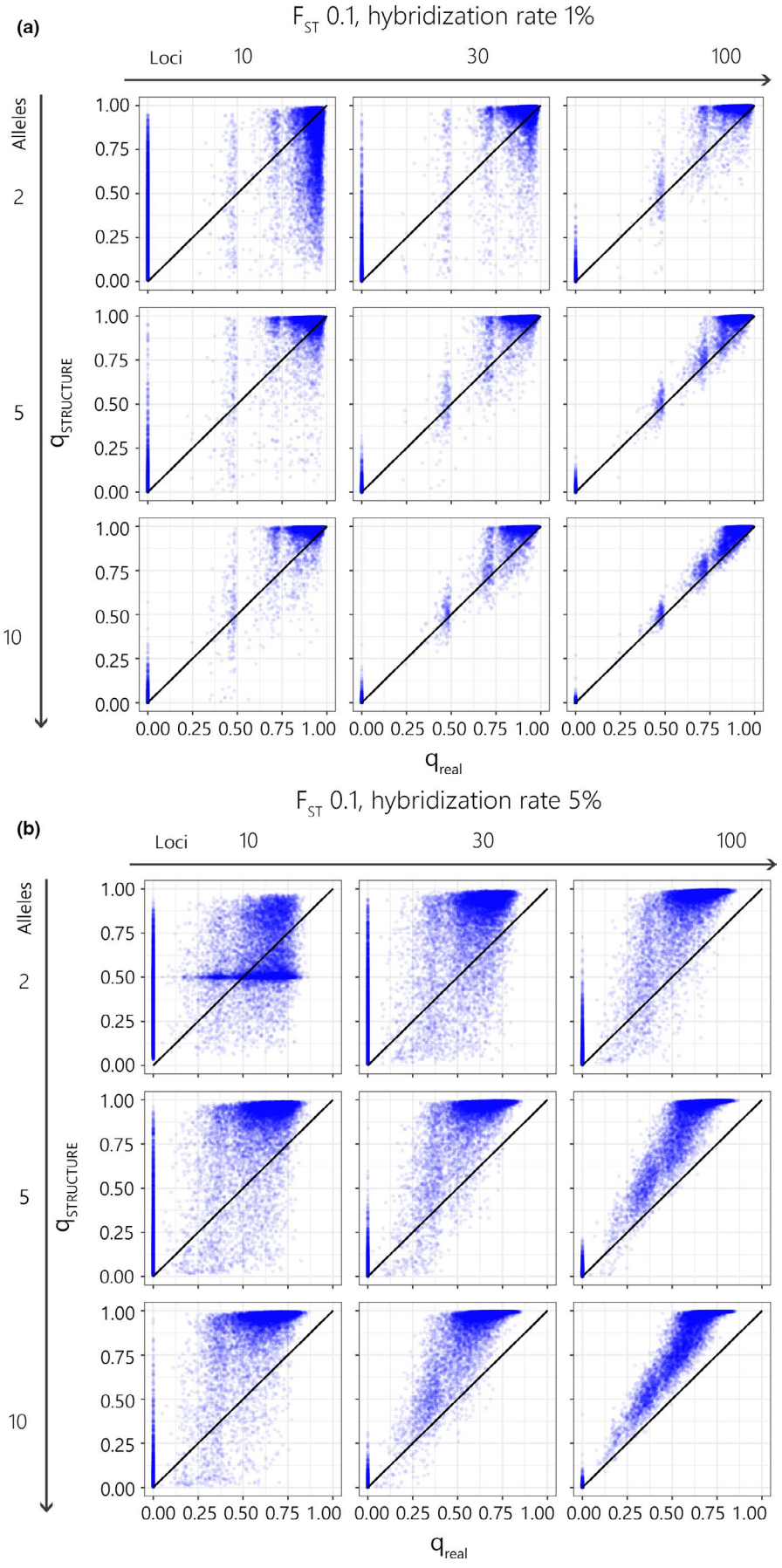


**FIGURE 2** Individual proportion of genome belonging to population A estimated with STRUCTURE ($q_{STRUCTURE}$) compared to the real proportion ($q_{real}$) calculated during the simulations. Simulations of admixture were conducted for two populations differentiated with $F_{ST} = 0.1$. Each panel represents 20,000 pairs of values, 10,000 pairs originating from population A and 10,000 from individuals from population B (100 runs with 100 individuals from each one of the two populations). If the estimates of STRUCTURE precisely corresponded to real values, points should lay on the diagonal. (a) hybridization rate of 1% per generation; (b) hybridization rate of 5% [Colour figure can be viewed at wileyonlinelibrary.com]

**TABLE 1** Generalized linear models testing the effect of $F_{ST}$ and number of alleles over the absolute difference between $q_{real}$ and $q_{STRUCTURE}$ for samples from the admixed population A

| Response variable | Hybridization rate | Explanatory variables | Estimates | *z* | *p* |
|---|---|---|---|---|---|
| $\lvert q_{STRUCTURE} - q_{real}\rvert$ | 1% | (intercept) | −2.541 | −300.71 | $<2e^{-16}$ |
| | | $F_{ST}$ | −1.007 | −26.93 | $<2e^{-16}$ |
| | | Number of alleles | −0.014 | −14.72 | $<2e^{-16}$ |
| $\lvert q_{STRUCTURE} - q_{real}\rvert$ | 5% | (intercept) | −0.838 | −109.26 | $<2e^{-16}$ |
| | | $F_{ST}$ | −0.943 | −27.98 | $<2e^{-16}$ |
| | | Number of alleles | −0.0129 | −15.4 | $<2e^{-16}$ |

*Note:* The two factors have a significant effect. We used 100 markers with five or 10 alleles and the two rates of hybridization.

difference between the two *q* values (Table 1), therefore improving STRUCTURE estimates, but having markers with more alleles also helped.

The degree of overestimation of *q* by STRUCTURE can be better appreciated in the distribution of $q_{STRUCTURE} - q_{real}$ for population A (Figure 3, S6 and S7). While the overestimation was limited when the hybridization rate was 1%, it dramatically increased when the rate was 5%. We compared the distribution of $q_{real}$ and $q_{STRUCTURE}$ for an example case ($F_{ST} = 0.1$, 30 markers with 10 alleles) for both hybridization rates (Figure 4). Although the two hybridization rates resulted in very different populations with regard to the level of introgression (see $q_{real}$ in Figure 4), the results obtained with STRUCTURE were almost identical (see $q_{STRUCTURE}$), suggestive of a relatively low introgression, and showing that STRUCTURE was unable to differentiate the two cases.

The allele frequencies estimated by STRUCTURE for population A were progressively diverging from the ancestral as the number of generations of introgression increased (Figure 5). At generation 0, before any admixture, the estimates of allele frequencies for populations A and B were similar to the frequencies calculated from the ancestral populations. The distances in this case were not 0 because the estimates obtained by STRUCTURE were based on subsamples of the populations, resulting in sampling errors. As introgression increased the number of alleles from population B into population A in subsequent generations, STRUCTURE estimates of the allele frequencies in population A diverged more strongly from the ancestral ones, while estimates of the allele frequencies for population B remained unaffected. As estimations of *q* are associated to the inferred allele frequencies, the inaccuracy in the ancestral allele frequency estimates could lead to the observed overestimations in $q_{STRUCTURE}$ for individuals originating from population A.

A possible solution to improve accuracy in the estimates of the allele frequencies could be the inclusion of reference individuals from the ancestral population A. Generalized linear models showed that (with a hybridization rate of 5%) the accuracy of $q_{STRUCTURE}$ estimates improved by increasing the proportion of reference individuals and the number of alleles per marker, as well as activating the POPALPHAS option and marking reference individuals with USEPOPINFO (Table 2; Figure 6 and S8). However, $q_{real}$ and $q_{STRUCTURE}$ were still very different (Figure 6). When the hybridization rate was lower (1%) the effect of including reference samples was not obvious, and the generalized linear model could not be fitted because the model assumptions were not fulfilled.

Despite the improvement in the estimates with the inclusion of reference individuals, the biases were still very apparent even in the case when 30% of the individuals from A corresponded to reference samples (Figure 6) and $q_{STRUCTURE}$ values were still significantly higher than the corresponding $q_{real}$ ($p < 10^{-16}$). This could be due to inherent biases in STRUCTURE or difficulties in the inference of the ancestral frequencies when a large proportion of the samples derived from the admixed population. To investigate which was the case, we compared $q_{real}$ and $q_{STRUCTURE}$ obtained when analysing small sets of 5 admixed individuals with 100 samples from A before admixture and 100 samples from B. The results (Figure 7) show that the biases in the estimates of $q_{STRUCTURE}$ practically disappeared. Variance of $q_{STRUCTURE}$ estimates was quite large when using 30 markers but centred around the corresponding $q_{real}$ values (along the diagonal in the figures), but still $q_{STRUCTURE}$ values were significantly higher than $q_{real}$ ($p < .003$). As expected, using 100 markers greatly reduced this variance and with 10 alleles $q_{STRUCTURE}$ and $q_{real}$ values were not significantly different (*p*-value = .510). These results imply that STRUCTURE estimates were not intrinsically biased and including a very large number of reference samples and high number of markers helped to reduce biases in the estimates.

The analysis of a data set from a natural population of common quails that experienced introgression from farm quails showed the same pattern when we added reference samples belonging to the same population but from before most of the restocking campaigns (Figure S9). The *q* values estimated after adding reference showed more admixture than those estimated without adding a proper reference for the wild population (Wilcoxon signed rank test, $p < 10^{-6}$). This result confirmed the pattern observed in the simulations.

Given the biases observed with STRUCTURE, we carried out additional analyses with simulated data corresponding to ancestral populations differing by $F_{ST} = 0.1$ and 100 biallelic loci with ADMIXTURE, Ohana and sNMF to assess if the same biases were present in all cases. We graphically compared results from the four programs and STRUCTURE exhibited the worst performance under the set of conditions that we evaluated (Figure S10). The Wilcoxon signed rank test confirmed that q values were not overestimated

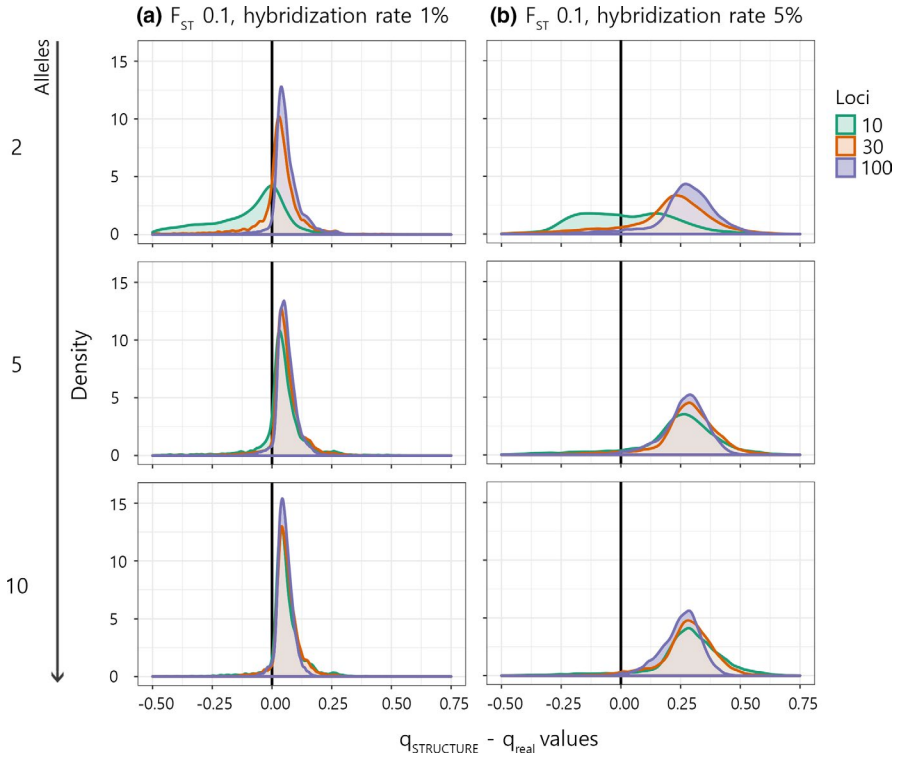


**FIGURE 3** Density plots of $q_{STRUCTURE} - q_{real}$ for individuals from population A. Simulations were carried for a differentiation of $F_{ST} = 0.1$ between the ancestral populations. (a) hybridization rate of 1%; (b) 5%. Data should ideally distribute around value 0 (marked with a vertical line). However, STRUCTURE analyses tended to overestimate the proportion of genome belonging to population A ($q_{STRUCTURE} - q_{real}$), being the overestimation bigger with higher hybridization rate [Colour figure can be viewed at wileyonlinelibrary.com]

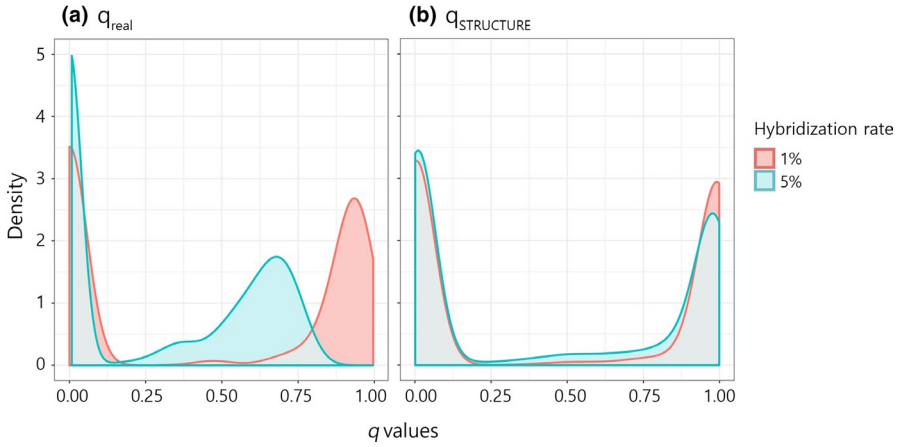


**FIGURE 4** Density plots for $q_{real}$ (a) and $q_{STRUCTURE}$ (b) resulting from hybridization rates of 1% and 5%). Simulations were carried for a degree of differentiation between the ancestral populations ($F_{ST}$) of 0.1 and 30 loci with 10 alleles. The values around $q = 0$ corresponded to individuals from population B, while the rest reflect the admixed population A. Although $q_{real}$ indicated different biological situations for the two hybridization rates (a), with very different number of nonadmixed A individuals ($q$ close to 1), $q_{STRUCTURE}$ values suggested that the two situations resulted in similar introgression (b) [Colour figure can be viewed at wileyonlinelibrary.com]

with ADMIXTURE, Ohana and sNMF when the hybridization rate was 1%.

## 4 | DISCUSSION

Our results show that when appropriate reference samples are not included in the analyses, ancestry estimates provided by STRUCTURE can be very biased. In fact, the results provided by this software can be very similar even when comparing populations with completely different levels of introgression (Figure 4).

When populations experience hybridization and admixture during multiple generations, the proportion of the genome deriving from the ancestral local population tends to be overestimated by STRUCTURE, leading to an underestimation of the real degree of introgression and of the number of admixed individuals. Although the precision of the ancestry estimates provided by STRUCTURE tends to improve with higher number of markers and alleles (as suggested by previous studies; McFarlane & Pemberton, 2019; Vähä & Primmer, 2006), this increase in precision does not correspond to an increase in accuracy and similar biases are observed using a small or a larger number of markers. The overestimate is particularly extreme in scenarios of hybrid swarms,

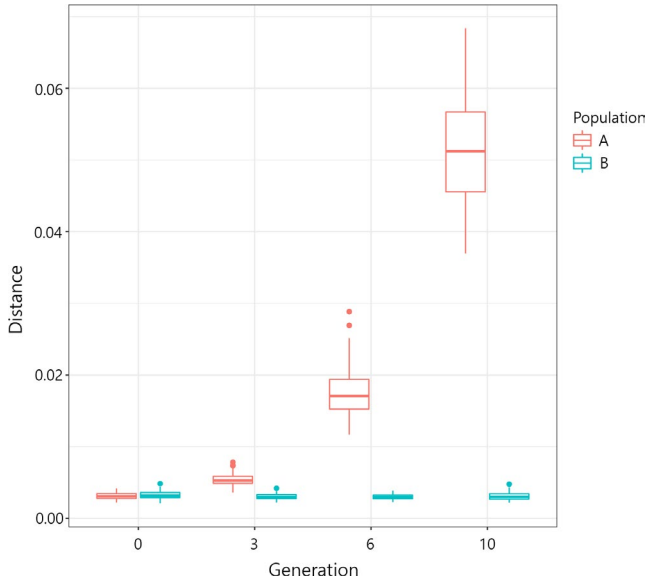


**FIGURE 5** Distance between estimates of ancestral allele frequencies obtained with STRUCTURE and their real values for populations A and B as introgression of alleles from B to A advanced. A total of 100 STRUCTURE estimates of ancestral allele frequencies were obtained at generations 0, three, six and 10. In population B, estimates corresponded closely to the real values in all generations. For the admixed population A, divergence from the ancestral allele frequencies increased with introgression showing that STRUCTURE did not correctly estimate ancestral allele frequencies [Colour figure can be viewed at wileyonlinelibrary.com]

as simulated with a hybridization rate of 5%: while none of the individuals was free of introgression, the estimation offered by the program suggested that purebred individuals were majority in the sample.

The poor performance of STRUCTURE under the simulated scenarios could impact the interpretation of admixture patterns in deeply introgressed populations, affected by various generations of hybridization, as in contact zones (Baldassarre et al., 2014; Johnson et al., 2015; Ortego et al., 2018). Extensive admixture could also occur in other cases, such as in the intercrossing between wild and domestic species, which have been coexisting for centuries or decades, or between different wild species that become in contact after a long time of evolution in isolation (Beaumont et al., 2001; Burgarella et al., 2018; Glover et al., 2017; Mckelvey et al., 2016; Scarcelli et al., 2017). It may seem that a hybridization rate of 5%, as in our models, is unlikely to exist in nature. However, similar values have been reported in the literature based on the identification of F1 hybrids (Lorenzini et al., 2014; Muñoz-Fuentes et al., 2007; Pacheco et al., 2017; Sullivan et al., 2016).

The limitations and the poor performance of STRUCTURE in admixed populations were already in part highlighted in the recommendations of Pritchard et al. (2000) about a proper utilization of the software to obtain reliable ancestry estimates. These authors indicate that, in cases of extensive admixture, STRUCTURE cannot estimate ancestral allele frequencies and it cannot give accurate estimates of $q$ because of the high variance in how many of the individual's alleles derive from one or the other population. Our results confirm this and show that introgression leads to poor estimates of the ancestral allele frequencies, as we observe how the allele frequencies estimated for the ancestral population were increasingly different from the real ancestral frequencies (Figure 5), reflecting the allele frequencies for an already admixed population. The biases in the estimates of ancestry persist even in the cases with 100 loci of high polymorphism (Figure 2), suggesting that the estimation of ancestral allele frequencies may not be corrected just by increasing the number of unlinked loci. After repeated introgression during several generations, STRUCTURE may be unable to reliably reconstruct ancestral allele frequencies. Including purebred reference individuals in the data set resulted in a larger improvement in the estimates of $q$ than just increasing the number of markers. However, even replacing 30% of the individuals from the admixed population A by reference individuals was not enough to completely remove the bias (Figure 6) and the best estimates were obtained when analysing just a few target individuals together with many reference individuals (Figure 7).

In order to develop cost-effective genetic tools for the assessment of introgression, efforts have been concentrating in the identification of the most suitable combination of markers. A reduced number of informative loci with high diagnostic power could be as effective as a high number of less informative loci, indicating that the discriminating power could be more important than their number (Oliveira et al., 2015; Randi et al., 2014). However, it is not completely clear what would be the best strategy to identify the most informative loci without previous data on marker variability across populations. On the other hand, in our simulations we considered

**TABLE 2** Generalized linear models explaining the differences between $q_{real}$ and $q_{STRUCTURE}$ for samples from the admixed population A

| Response variable | Hybridization rate | Explanatory variables | Estimates | $z$ | $p$ |
|---|---|---|---|---|---|
| $\lvert q_{STRUCTURE} - q_{real}\rvert$ | 5% | (intercept) | −0.863 | −140.89 | $<2e^{-16}$ |
| | | Reference | −3.432 | −205.56 | $<2e^{-16}$ |
| | | Number of alleles | −0.014 | −19.32 | $<2e^{-16}$ |
| | | USEPOPINFO (activated) | −0.133 | −33.92 | $<2e^{-16}$ |
| | | POPALPHAS (activated) | −0.499 | −140.03 | $<2e^{-16}$ |

*Note:* The explanatory variables were the proportion of samples used as a reference (0%, 10% or 30%), number of alleles of the markers (five or 10) and use of the USEPOPINFO and POPALPHAS options in STRUCTURE for simulations using 30 loci and a hybridization rate of 5%. The strongest effect is associated to the proportion of individuals used as reference (an increase in the proportion of individuals used as reference leads to a decrease in the difference between the estimates).

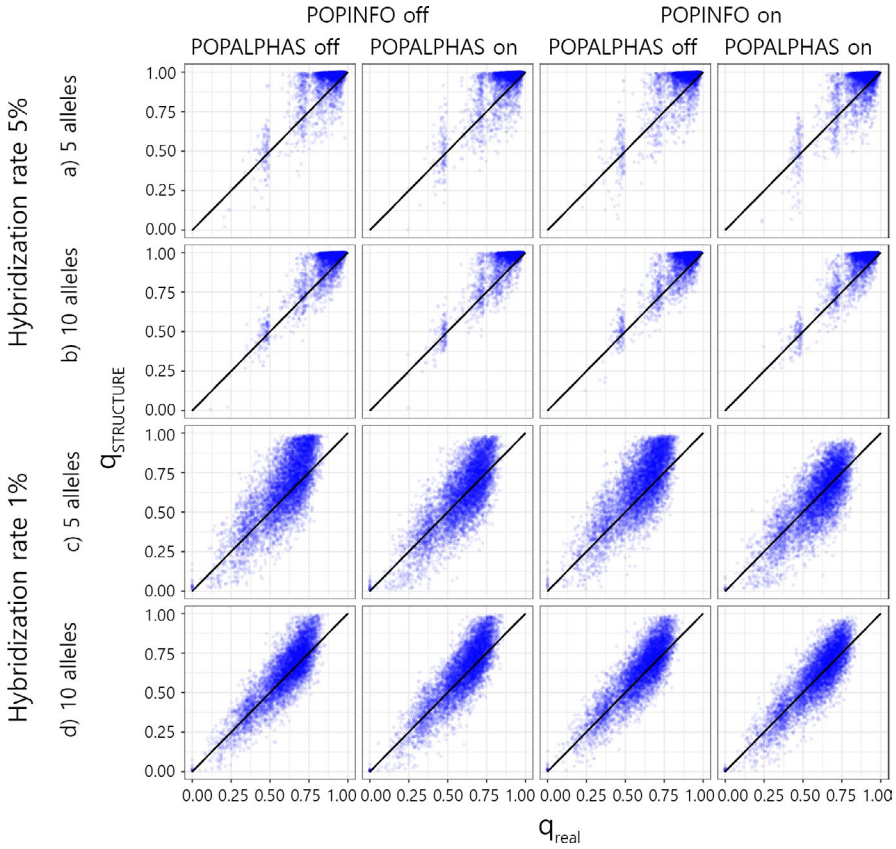


**FIGURE 6** Comparison of $q_{STRUCTURE}$ and $q_{real}$ when 30% of the individuals from the target population are sampled from the ancestral population and are used as reference. Individuals used as reference and those belonging to population B were excluded from the plots. Simulations were carried out with $F_{ST} = 0.1$ and 30 loci, varying the number of alleles per marker and the hybridization rate. The accuracy of $q_{STRUCTURE}$ estimates improved notably with the inclusion of reference individuals (compare to Figure 2b) and when activating the POPALPHAS option, especially for the cases with higher hybridization rate, where no pure individuals remained in the admixed population [Colour figure can be viewed at wileyonlinelibrary.com]

all loci to be unlinked. The use of linked markers could also improve the identification of older admixture between populations (Falush et al., 2003; Lecis et al., 2006) because introgression can differently affect regions of the genome (for example, see Anderson et al., 2009), but may be less suitable to study ongoing hybridization and introgression.

Our study highlights the importance of carrying out simulations in each study case to assess the reliability of estimates. The level of introgression can be assessed combining different approaches and simulations trying to properly reflect the functioning of the study system (McFarlane & Pemberton, 2019) should be carried out to test the power of the analysis, determine accuracy and assess possible biases (for example, see Oliveira et al., 2008; Randi et al., 2014; Sanchez-Donoso et al., 2014 Sanz et al., 2009).

The estimates by STRUCTURE show a remarkable increase in accuracy when many nonadmixed reference individuals are included in the analysis (Figure 6). Therefore, we stress the importance of including samples from reference nonadmixed populations (for example, museum specimens dating from before an admixture event; see also Sanchez-Donoso et al., 2014) to increase the reliability of STRUCTURE analyses. The availability of these samples can be limited, and historical DNA extraction and amplification can be costly and effort-demanding, but the accuracy in the analysis improves notably, increasing the reliability of the results. When only a limited number of reference samples is available, it could be useful to carry on multiple STRUCTURE analyses including only a small proportion of those samples whose ancestry is unknown (Figure 7), using USEPOPINFO to define the reference samples and comparing the results obtained with different test sample sets. Also, the use of the option POPALPHAS can improve the analyses when source populations are unequally represented and if there is unbalanced sampling (see Wang, 2017). The use of a small number of markers is also known to influence the results of STRUCTURE (Toyama et al., 2020). Nevertheless, Lawson et al. (2018) have shown that different demographic histories can lead to identical results suggesting admixture in STRUCTURE and emphasize the importance of combining analytical approaches to obtain a more robust analysis of recent demographic history. Our comparison of the results provided by different programs showed that not all of them suffer the same biases or to the same degree. Consequently, we strongly suggest to combine STRUCTURE with other approaches and simulations, and evaluate the consistency in the results, especially when it is not possible to include suitable reference samples in the analyses. This is especially important, for example, in the cases where hybridization and introgression can be relevant in the design of management and conservation plans.

In ecology, conservation and evolutionary biology, it is important to efficiently identify hybrids and admixed individuals, as well as to determine gene flow among different populations. STRUCTURE analyses showed a tendency to classify admixed individuals as nonadmixed when reference samples were not included. The misidentification of the degree of introgression can impact our understanding about the evolutionary history of species or the risks of genetic homogenization and extinction, and therefore its implications for management and conservation plans should be carefully considered.

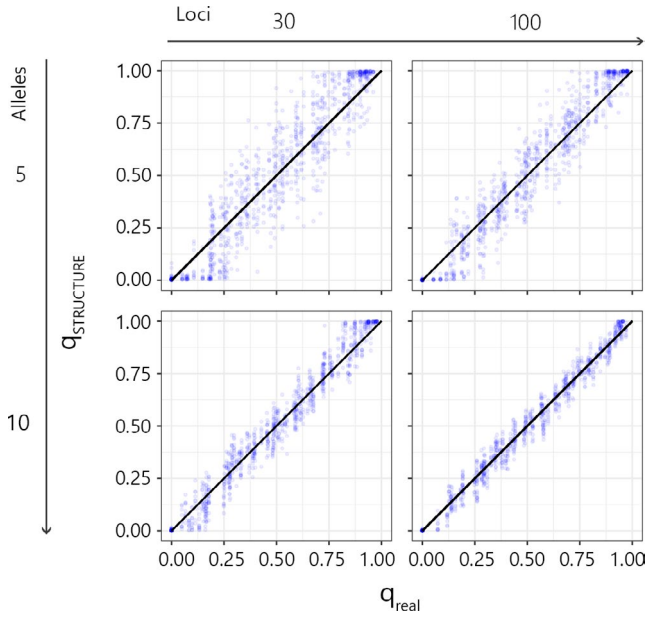


**FIGURE 7** Comparison of $q_{STRUCTURE}$ and $q_{real}$ when most of the analysed individuals were used as reference and admixture was evaluated in just a few individuals. Simulations were carried out for a differentiation of $F_{ST} = 0.1$ between the ancestral populations. Each panel represents 10 independent $q_{STRUCTURE}$ estimates for 100 admixed individuals representing the full range of $q_{real}$ values. In each STRUCTURE run, 100 individuals from the ancestral population A and 100 from population B were analysed together with five admixed samples. The points lie around the diagonal showing that $q_{real}$ and $q_{STRUCTURE}$ tended to be similar [Colour figure can be viewed at wileyonlinelibrary.com]

## ACKNOWLEDGEMENTS

We thank Carlos Rafii for providing access to servers to run simulations and analyses. Part of the computer work was carried out in Genomics servers of the Doñana's Singular Scientic-Technical Infrastructure (ICTS-RBD). The Conservation and Evolutionary Genetics Group at the Estación Biológica de Doñana provided valuable support and comments on the manuscript and Dr Jennifer Leonard also helped us to improve the English. The study is supported by project CGL2016-75227-P to CV from the Spanish Government and an FPI (Formación de Personal Investigador) fellowship (BES-2017-081291) to SR. We are also grateful to the anonymous referees and associate editors for their constructive comments.

## AUTHOR CONTRIBUTIONS

C.V., and I.S.-D. designed the study, S.R. carried out the project and wrote the scripts, S.R., and I.S.-D. analysed the data, S.R. wrote the first draft of the manuscript and all authors contributed to the text. This study was initiated as part of the Master in Biodiversity and Conservation Biology of S.R. at the University Pablo de Olavide (Seville, Spain).

## DATA AVAILABILITY STATEMENT

Genotypes corresponding to simulated starting populations, scripts used to simulate introgression and sampling, and output of these scripts that were used for subsequent analyses (as well as R scripts used to analyse and plot data) are available at https://github.com/sararvg/introgression_structure. Quail data is available at the original publication (Sanchez-Donoso et al., 2012).

## ORCID

*Sara Ravagni* https://orcid.org/0000-0003-0320-3447
*Ines Sanchez-Donoso* https://orcid.org/0000-0003-2773-9844
*Carles Vilà* https://orcid.org/0000-0002-4206-5246

## REFERENCES

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., Butlin, R. K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S. H., Hermansen, J. S., Hewitt, G., Hudson, A. G., Jiggins, C., … Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*(2), 229–246. https://doi.org/10.1111/j.1420-9101.2012.02599.x

Alacs, E. A., Georges, A., FitzSimmons, N. N., & Robertson, J. (2010). DNA detective: A review of molecular approaches to wildlife forensics. *Forensic Science, Medicine, and Pathology*, *6*, 180–194. https://doi.org/10.1007/s12024-009-9131-7

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*(1), 246. https://doi.org/10.1186/1471-2105-12-246

An, M., Deng, M., Zheng, S.-S., Jiang, X.-L., & Song, Y.-G. (2017). Introgression threatens the genetic diversity of *Quercus austrocochinchinensis* (Fagaceae), an endangered oak: A case inferred by molecular markers. *Frontiers in Plant Science*, *8*, 229. https://doi.org/10.3389/fpls.2017.00229

Arias, O., Cordeiro, E., Corrêa, A. S., Domingues, F. A., Guidolin, A. S., & Omoto, C. (2019). Population genetic structure and demographic history of *Spodoptera frugiperda* (Lepidoptera: Noctuidae): Implications for insect resistance management programs. *Pest Management Science*, *75*(11), 2948–2957. https://doi.org/10.1002/ps.5407

Arnold, M. L. (2015). Divergence with genetic exchange. *Divergence with Genetic Exchange*. https://doi.org/10.1093/acprof:oso/9780198726029.001.0001

Anderson, T. M., Bridgett, M., Candille, S. I., Musiani, M., Stahler, D. R., Smith, D. W., Padhukasahasram, B., Randi, E., Leonard, J. A., Bustamante, C. D., Ostrander, E. A., Tang, H., Wayne, R. K., & Barsh, G. S. (2009). Molecular and eovlutional history of melanism in North American Gray Wolves. *Science*, *323*(5919), 1339–1343. https://doi.org/10.1126/science.1165448

Baldassarre, D. T., White, T. A., Karubian, J., & Webster, M. S. (2014). Genomic and morphological analysis of a semipermeable avian hybrid zone suggests asymmetrical introgression of a sexual signal. *Evolution*, *68*(9), 2644–2657. https://doi.org/10.1111/evo.12457

Balloux, F. (2001). EASYPOP (Version 1.7): A computer program for population genetics simulations. *Journal of Heredity*, *92*(3), 301–302. https://doi.org/10.1093/jhered/92.3.301

Barilani, M., Bernard-Laurent, A., Mucci, N., Tabarroni, C., Kark, S., Perez Garrido, J. A., & Randi, E. (2007). Hybridisation with introduced chukars (*Alectoris chukar*) threatens the gene pool integrity of native rock (*A. graeca*) and red-legged (*A. rufa*) partridge populations. *Biological Conservation*, *137*(1), 57–69. https://doi.org/10.1016/j.biocon.2007.01.014

Beaumont, M., Barratt, E. M., Gottelli, D., Kitchener, A. C., Daniels, M. J., Pritchard, J. K., & Bruford, M. W. (2001). Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology*, *10*(2), 319–336. https://doi.org/10.1046/j.1365-294X.2001.01196.x

Bohling, J. H., Adams, J. R., & Waits, L. P. (2013). Evaluating the ability of Bayesian clustering methods to detect hybridization and

introgression using an empirical red wolf data set. *Molecular Ecology*, *22*(1), 74–86. https://doi.org/10.1111/mec.12109

Burgarella, C., Barnaud, A., Kane, N. A., 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

Burgarella, C., Cubry, P., Kane, N. A., Varshney, R. K., Mariac, C., Liu, X., Shi, C., Thudi, M., Couderc, M., Xu, X., Chitikineni, A., Scarcelli, N., Barnaud, A., Rhoné, B., Dupuy, C., François, O., Berthouly-Salazar, C., & Vigouroux, Y. (2018). A western Sahara centre of domestication inferred from pearl millet genomes. *Nature Ecology and Evolution*, *2*(9), 1377–1380. https://doi.org/10.1038/s41559-018-0643-y

Cheng, J. Y., Mailund, T., & Nielsen, R. (2017). Fast admixture analysis and population tree estimation for SNP and NGS data. *Bioinformatics*, *33*(14), 2148–2155. https://doi.org/10.1093/bioinformatics/btx098

De Barba, M., Miquel, C., Lobréaux, S., Quenette, P. Y., Swenson, J. E., & Taberlet, P. (2017). High-throughput microsatellite genotyping in ecology: Improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. *Molecular Ecology Resources*, *17*(3), 492–507. https://doi.org/10.1111/1755-0998.12594

Drake, J. W., Charlesworth, B., Charlesworth, D., & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, *148*(4), 1667–1686.

Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics*, *5*, 435–445. https://doi.org/10.1038/nrg1348

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567–1587. https://doi.org/10.1111/j.1471-8286.2007.01758.x

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, *196*(4), 973–983. https://doi.org/10.1534/genetics.113.160572

Glover, K. A., Solberg, M. F., McGinnity, P., Hindar, K., Verspoor, E., Coulson, M. W., Hansen, M. M., Araki, H., Skaala, Ø., & Svåsand, T. (2017). Half a century of genetic interaction between farmed and wild Atlantic salmon: Status of knowledge and unanswered questions. *Fish and Fisheries*, *18*(5), 890–927. https://doi.org/10.1111/faf.12214

Godinho, R., Llaneza, L., Blanco, J. C., Lopes, S., Álvares, F., García, E. J., Palacios, V., Cortés, Y., Talegón, J., & Ferrand, N. (2011). Genetic evidence for multiple events of hybridization between wolves and domestic dogs in the Iberian Peninsula. *Molecular Ecology*, *20*(24), 5154–5166. https://doi.org/10.1111/j.1365-294X.2011.05345.x

Gómez-Sánchez, D., Olalde, I., Sastre, N., Enseñat, C., Carrasco, R., Marques-Bonet, T., Lalueza-Fox, C., Leonard, J. A., Vilà, C., & Ramírez, O. (2018). On the path to extinction: Inbreeding and admixture in a declining grey wolf population. *Molecular Ecology*, *27*(18), 3599–3612. https://doi.org/10.1111/mec.14824

Hamilton, J. A., & Miller, J. M. (2016). Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conservation Biology*, *30*(1), 33–41. https://doi.org/10.1111/cobi.12574

Hothorn, T., Van De Wiel, M. A., Hornik, K., & Zeileis, A. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, *28*(8), 1–23. https://doi.org/10.18637/jss.v028.i08

Johnson, B. B., White, T. A., Phillips, C. A., & Zamudio, K. R. (2015). Asymmetric introgression in a spotted salamander hybrid zone. *Journal of Heredity*, *106*(5), 608–617. https://doi.org/10.1093/jhered/esv042

Lavretsky, P., Janzen, T., & McCracken, K. G. (2019). Identifying hybrids & the genomics of hybridization: Mallards & American black ducks of Eastern North America. *Ecology and Evolution*, *9*(6), 3470–3490. https://doi.org/10.1002/ece3.4981

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*(1), 1–11. https://doi.org/10.1038/s41467-018-05257-7

Lecis, R., Pierpaoli, M., Birò, Z. S., Szemethy, L., Ragni, B., Vercillo, F., & Randi, E. (2006). Bayesian analyses of admixture in wild and domestic cats (*Felis silvestris*) using linked microsatellite loci. *Molecular Ecology*, *15*(1), 119–131. https://doi.org/10.1111/j.1365-294X.2005.02812.x

Lorenzini, R., Fanelli, R., Grifoni, G., Scholl, F., & Fico, R. (2014). Wolf-dog crossbreeding: "Smelling" a hybrid may not be easy. *Mammalian Biology*, *79*(2), 149–156. https://doi.org/10.1016/j.mambio.2013.07.080

McFarlane, S. E., & Pemberton, J. M. (2019). Detecting the true extent of introgression during anthropogenic hybridization. *Trends in Ecology and Evolution*, *34*, 315–326. https://doi.org/10.1016/j.tree.2018.12.013

Mckelvey, K. S., Young, M. K., Wilcox, T. M., Bingham, D. M., Pilgrim, K. L., & Schwartz, M. K. (2016). Patterns of hybridization among cutthroat trout and rainbow trout in northern Rocky Mountain streams. *Ecology and Evolution*, *6*(3), 688–706. https://doi.org/10.1002/ece3.1887

Muñoz-Fuentes, V., Vilà, C., Green, A. J., Negro, J. J., & Sorenson, M. D. (2007). Hybridization between white-headed ducks and introduced ruddy ducks in Spain. *Molecular Ecology*, *16*(3), 629–638. https://doi.org/10.1111/j.1365-294X.2006.03170.x

Neophytou, C. (2014). Bayesian clustering analyses for genetic assignment and study of hybridization in oaks: Effects of asymmetric phylogenies and asymmetric sampling schemes. *Tree Genetics and Genomes*, *10*(2), 273–285. https://doi.org/10.1007/s11295-013-0680-2

Nussberger, B., Wandeler, P., Weber, D., & Keller, L. F. (2014). Monitoring introgression in European wildcats in the Swiss Jura. *Conservation Genetics*, *15*(5), 1219–1230. https://doi.org/10.1007/s10592-014-0613-0

Oliveira, R., Godinho, R., Randi, E., Ferrand, N., & Alves, P. C. (2008). Molecular analysis of hybridisation between wild and domestic cats (*Felis silvestris*) in Portugal: Implications for conservation. *Conservation Genetics*, *9*(1), 1–11. https://doi.org/10.1007/s10592-007-9297-z

Oliveira, R., Randi, E., Mattucci, F., Kurushima, J. D., Lyons, L. A., & Alves, P. C. (2015). Toward a genome-wide approach for detecting hybrids: Informative SNPs to detect introgression between domestic cats and European wildcats (*Felis silvestris*). *Heredity*, *115*(3), 195–205. https://doi.org/10.1038/hdy.2015.25

Ortego, J., Gugger, P. F., & Sork, V. L. (2017). Impacts of human-induced environmental disturbances on hybridization between two ecologically differentiated Californian oak species. *New Phytologist*, *213*(2), 942–955. https://doi.org/10.1111/nph.14182

Ortego, J., Gugger, P. F., & Sork, V. L. (2018). Genomic data reveal cryptic lineage diversification and introgression in Californian golden cup oaks (section Protobalanus). *New Phytologist*, *218*(2), 804–818. https://doi.org/10.1111/nph.14951

Oziolor, E. M., Reid, N. M., Yair, S., Lee, K. M., Guberman VerPloeg, S., Bruns, P. C., Shaw, J. R., Whitehead, A., & Matson, C. W. (2019). Adaptive introgression enables evolutionary rescue from extreme environmental pollution. *Science*, *364*(6439), 455–457. https://doi.org/10.1126/science.aav4155

Pacheco, C., López-Bao, J. V., García, E. J., Lema, F. J., Llaneza, L., Palacios, V., & Godinho, R. (2017). Spatial assessment of Wolf-dog hybridization in a single breeding period. *Scientific Reports*, *7*(1), 1–10. https://doi.org/10.1038/srep42475

Payseur, B. A., & Nachman, M. W. (2000). Microsatellite variation and recombination rate in the human genome. *Genetics*, *156*(3), 1285–1298.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Puechmaille, S. J. (2016). The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, *16*(3), 608–627. https://doi.org/10.1111/1755-0998.12512

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Randi, E. (2008). Detecting hybridization between wild species and their domesticated relatives. *Molecular Ecology*, *17*, 285–293. https://doi.org/10.1111/j.1365-294X.2007.03417.x

Randi, E., Hulva, P., Fabbri, E., Galaverni, M., Galov, A., Kusak, J., Bigi, D., Bolfíková, B. Č., Smetanová, M., & Caniglia, R. (2014). Multilocus detection of wolf x dog hybridization in Italy, and guidelines for marker selection. *PLoS One*, *9*(1), e86409. https://doi.org/10.1371/journal.pone.0086409

Rhymer, J. M., & Simberloff, D. (1996). Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, *27*(1), 83–109. https://doi.org/10.1146/annurev.ecolsys.27.1.83

RStudio Team (2020). *RStudio: Integrated development for R*. RStudio, PBC. Retrieved from http://www.rstudio.com/

Sanchez-Donoso, I., Huisman, J., Echegaray, J., Puigcerver, M., Rodríguez-Teijeiro, J. D., Hailer, F., & Vilá, C. (2014). Detecting slow introgression of invasive alleles in an extensively restocked game bird. *Frontiers in Ecology and Evolution*, *2*(15), 1–17. https://doi.org/10.3389/fevo.2014.00015

Sanchez-Donoso, I., Vilà, C., Puigcerver, M., Butkauskas, D., Caballero de la Calle, J. R., Morales-Rodríguez, P. A., & Rodríguez-Teijeiro, J. D. (2012). Are farm-reared quails for game restocking really common quails (*Coturnix coturnix*)?: A genetic approach. *PLoS One*, *7*(6), e39031. https://doi.org/10.1371/journal.pone.0039031

Sanz, N., Araguas, R. M., Fernández, R., Vera, M., & García-Marín, J. L. (2009). Efficiency of markers and methods for detecting hybrids and introgression in stocked populations. *Conservation Genetics*, *10*(1), 225–236. https://doi.org/10.1007/s10592-008-9550-0

Scarcelli, N., Chaïr, H., Causse, S., Vesta, R., Couvreur, T. L. P., & Vigouroux, Y. (2017). Crop wild relative conservation: Wild yams are not that wild. *Biological Conservation*, *210*(2016), 325–333. https://doi.org/10.1016/j.biocon.2017.05.001

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Suarez-Gonzalez, A., Lexer, C., & Cronk, Q. C. B. (2018). Adaptive introgression: A plant perspective. *Biology Letters*, *14*(3), 20170688. https://doi.org/10.1098/rsbl.2017.0688

Sujii, P. S., Cozzolino, S., & Pinheiro, F. (2019). Hybridization and geographic distribution shapes the spatial genetic structure of two co-occurring orchid species. *Heredity*, *123*, 458–469. https://doi.org/10.1038/s41437-019-0254-7

Sullivan, A. R., Owusu, S. A., Weber, J. A., Hipp, A. L., & Gailing, O. (2016). Hybridization and divergence in multi-species oak (*Quercus*) communities. *Botanical Journal of the Linnean Society*, *181*(1), 99–114. https://doi.org/10.1111/boj.12393

Toyama, K. S., Crochet, P. A., & Leblois, R. (2020). Sampling schemes and drift can bias admixture proportions inferred by structure. *Molecular Ecology Resources*, *20*(6), 1769–1785. https://doi.org/10.1111/1755-0998.13234

Trigo, T. C., Schneider, A., De Oliveira, T. G., Lehugeur, L. M., Silveira, L., Freitas, T. R. O., & Eizirik, E. (2013). Molecular data reveal complex hybridization and a cryptic species of Neotropical wild cat. *Current Biology*, *23*(24), 2528–2533. https://doi.org/10.1016/j.cub.2013.10.046

Vähä, J.-P., & Primmer, C. R. (2006). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, *15*(1), 63–72. https://doi.org/10.1111/j.1365-294X.2005.02773.x

Verdu, P., & Rosenberg, N. A. (2011). A general mechanistic model for admixture histories of hybrid populations. *Genetics*, *189*(4), 1413–1426. https://doi.org/10.1534/genetics.111.132787

Wang, J. (2017). The computer program STRUCTURE for assigning individuals to populations: Easy to use but easier to misuse. *Molecular Ecology Resources*, *17*(5), 981–990. https://doi.org/10.1111/1755-0998.12650

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. Retrieved from https://cran.r-project.org/web/packages/ggplot2/citation.html

Zeileis, A., Cribari-Neto, F., Gruen, B., & Kosmidis, I.(2012). Package "betareg". Retrieved from https://cran.r-project.org/web/packages/betareg/betareg.pdf

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Ravagni S, Sanchez-Donoso I, Vilà C. Biased assessment of ongoing admixture using STRUCTURE in the absence of reference samples. *Mol Ecol Resour.* 2021;21:677–689. https://doi.org/10.1111/1755-0998.13286