Sampling schemes and drift can bias admixture proportions inferred by structure

Ken S. Toyama, Pierre-André Crochet, Raphaël Leblois

Received: 10 February 2020 | Revised: 17 July 2020 | Accepted: 24 July 2020

DOI: 10.1111/1755-0998.13234

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Sampling schemes and drift can bias admixture proportions inferred by STRUCTURE

Ken S. Toyama[1] | Pierre-André Crochet[2] | Raphaël Leblois[3,4]

[1]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada

[2]CEFE, CNRS, University of Montpellier, Université Paul Valéry Montpellier 3, EPHE, IRD, Montpellier, France

[3]CBGP, INRAE, CIRAD, IRD, Montpellier SupAgro, University of Montpellier, Montpellier, France

[4]Institut de Biologie Computationnelle, University of Montpellier, Montpellier, France

Correspondence
Ken S. Toyama, Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON M5S 3B2, Canada.
Email: ken.toyama@mail.utoronto.ca

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-16-CE02-0008 ANR-19-CE02-0011

Abstract

The interbreeding of individuals coming from genetically differentiated but incompletely isolated populations can lead to the formation of admixed populations, having important implications in ecology and evolution. In this simulation study, we evaluate how individual admixture proportions estimated by the software STRUCTURE are quantitatively affected by different factors. Using various scenarios of admixture between two diverging populations, we found that unbalanced sampling from parental populations may seriously bias the inferred admixture proportions; moreover, proportionally large samples from the admixed population can also decrease the accuracy and precision of the inferences. As expected, weak differentiation between parental populations and drift after the admixture event strongly increase the biases caused by uneven sampling. We also show that admixture proportions are generally more biased when parental populations unequally contributed to the admixed population. Finally, with few exceptions, using a large number of markers reduces those biases, but using alternative priors for individual ancestry or the uncorrelated allele model only marginally affect the inference of admixture in most situations. We conclude that unbalanced sampling may cause important biases in the admixture proportions estimated by STRUCTURE, especially when a small number of markers are used, and those biases can be worsened by the effect of drift and unequal genetic contribution of parental populations. Empirical studies should thus be careful with their sampling design and consider historical characteristics when using this software to estimate the ancestry of individuals from admixed populations.

KEYWORDS

admixture, drift, simulations, STRUCTURE, unbalanced sampling

1 | INTRODUCTION

In organisms with sexual reproduction, the interbreeding of individuals from genetically differentiated but incompletely isolated lineages can result in the inclusion of new alleles in an ancestral gene pool through the backcrossing of hybrid individuals, sometimes leading to the formation of new admixed lineages made up of alleles from both ancestral gene pools (Dannemann, Andrés, & Kelso, 2016; De Carvalho et al., 2010; Ellstrand, Prentice, & Hancock, 1999; Hedrick, 2013; Kolbe et al., 2007; Mallet, 2007; Simonti et al., 2016). Introgression and admixture are so widespread that taking these phenomena into account in empirical studies is needed to reconstruct the evolutionary history of populations or understand the evolution of reproductive isolation and speciation (Barton, 1989; Coyne & Orr, 2004; Hewitt, 2001; Miraldo, Faria, Hewitt, Paulo, & Emerson, 2013; Nielsen, Hansen, Ruzzante, Meldrup, & Grønkjær, 2003; Trigo et al., 2008).

Introgression and admixture also have ecological and evolutionary consequences per se: the genetic impact of introgression from non-native lineages may be a major risk for many organisms

of conservation concern (Vilà, Weber, & Antonio, 2000; Vilaca et al., 2012; Wolf, Takebayashi, & Rieseberg, 2001), while the concept of adaptive introgression is gaining strength in the context of present and future consequences of climate change (Hamilton & Miller, 2016; Hedrick, 2013; von Holdt, Brzeski, Wilcove, & Rutledge, 2018). In these empirical contexts, an accurate determination of admixture proportions at the population or individual level is crucial (Aldrich et al., 2008; Tang, Peng, Wang, & Risch, 2005).

The individual admixture proportion, usually referred to as *Q*, represents the fraction of an individual's alleles coming from a given population. Individuals are classified as pure ($Q = 0$ or 1), hybrids ($Q = 0.5$, first = F1 or second = F2 generations), first-generation backcross ($Q = 0.25$) or, more generally, partially introgressed (in theory any *Q*-value between 0 and 1). Many empirical studies use admixture proportions to fit genetic models related to hybrid zones and introgression patterns (Cornille et al., 2015; Gagnaire et al., 2009; Schaefer, Duvernell, & Campbell, 2016; Zohren et al., 2016). In all these situations, the relevance of biological conclusions strongly depends on the quality of the estimation of individual admixture proportions.

Several methods and software have been developed to estimate admixture proportions (see Padhukasahasram, 2014 for a review). Among these, STRUCTURE (Pritchard, Stephens, & Donnelly, 2000) is the most commonly used for quantifying admixture levels or identifying hybrid individuals (e.g. Burgarella et al., 2009; Cornille et al., 2015; Hermansen et al., 2014; Nielsen et al., 2003; Porras-Hurtado et al., 2013). To this end, STRUCTURE uses an unsupervised clustering method implemented in a Bayesian framework to assign individuals to homogeneous genetic groups (often named clusters) that maximize Hardy–Weinberg and linkage equilibrium (Pritchard et al., 2000). Besides hybrid detection and admixture quantification, STRUCTURE was originally designed to infer the population structure (i.e. homogeneous genetic groups) in a genetic sample and assign individuals to populations (Pritchard et al., 2000). As expected, its performance in identifying the true population structure depends on the quantity of information present in the data set (e.g. number of loci and individuals, differentiation level) but also on relative sample sizes: when the samples from different populations are highly unbalanced in size, STRUCTURE tends to merge individuals from populations with small samples into a single cluster (Kalinowski, 2011; Puechmaille, 2016). However, Wang (2017) recently demonstrated how using an alternative ancestry prior instead of the default one for STRUCTURE solved most problems caused by uneven sampling in the estimation of population structure.

Several studies have also tested STRUCTURE's performance in detecting hybridization and introgression. STRUCTURE has been shown to perform well in terms of precision (i.e. how variable are the results) and accuracy (i.e. how far are the results from the expected values) in the detection of F1 hybrids (Bohling, Adams, & Waits, 2013; Larcombe, Vaillancourt, Jones, & Potts, 2014; Sanz, Araguas, Fernández, Vera, & García-Marín, 2009; Vähä & Primmer, 2006). However, the detection of backcrossed individuals seems more difficult (Marie, Bernatchez, & Garant, 2011; Sanz et al., 2009). For example, Vähä and Primmer (2006) found that backcrossed individuals could be erroneously classified as purebreds when using low threshold *Q*-values, suggesting that STRUCTURE did not estimate individual *Q*-values with accuracy in those situations. As expected, better results are obtained when the level of differentiation between parental populations is high and when a large number of loci are used. Studies with low differentiation levels between parental populations thus necessitate large numbers of loci to obtain precise results (Kalinowski, 2011; Sanz et al., 2009; Vähä & Primmer, 2006). Last, sampling scheme also seems to affect the ability of STRUCTURE to detect hybrid and pure individuals: Marie et al. (2011) and Vähä and Primmer (2006) found that a greater accuracy of detection is reached when the sample includes a larger number of hybrids (see also Neophytou, 2014).

Although not specifically investigated in previous studies, the age of the admixture event should also affect how well admixture is inferred. If it is recent, implying limited drift in all populations after admixture, it will be easier for STRUCTURE to identify alleles coming from parental populations in the admixed population and to assign them to one of the parental clusters. In contrast, if the admixture event is old, drift will modify allelic frequencies in the parental and admixed populations and will reduce the ability of STRUCTURE to quantify admixture proportions, as previously shown by Kalinowski (2011).

Clearly, the sampling scheme, the degree of differentiation between parental populations, the age of the admixture event and the choice of ancestry priors may all have important effects on STRUCTURE's ability to accurately estimate admixture proportions. Given the importance of their accurate estimation in empirical studies, it is surprising that, to the best of our knowledge, no previous study has evaluated how admixture proportion estimates provided by STRUCTURE are affected by these factors.

In this work, we assessed by simulations how individual admixture proportions inferred by STRUCTURE are affected by different patterns of unbalanced sampling, differentiation between parental populations and the amount of drift after the admixture event. To this end, we simulated SNP data under various scenarios where two divergent populations give rise to a third admixed population and then evolve independently until sampling time. We used various combinations of sample sizes from those three populations and analysed them using STRUCTURE to infer admixture proportions in the admixed sample. We also evaluated the effects of the number of markers used, the initial admixture rate (the proportion of individuals from each parental population during the admixture event) and the choice of the ancestry prior on the estimation of individual admixture proportions.

2 | MATERIALS AND METHODS

2.1 | Simulation settings

Data were simulated under a demographic scenario of divergence and admixture in which two populations (defined as the "parental"

populations P1 and P2) diverged from an ancestral panmictic population *TD* generations ago (time of divergence); later, *TA* (time of admixture, *TA* < *TD*) generations in the past, a third population PA, the admixed population, was created by mixing a proportion p_{adx} of individuals from P1 and a proportion $(1 - p_{adx})$ from P2 (Figure 1a). Note that, for ease of reading, all times are expressed backward in time since sampling, and in generations, and effective population sizes are expressed in number of diploid individuals, whereas all simulations were run with the program *ms* (Hudson, 2002), which considers scaled parameters (i.e. times in T/4N and relative population sizes).

Polymorphism was generated with the single mutation SNP generating option -s 1 (see Appendix S1 for a detailed *ms* command line example). Effective diploid population sizes were fixed to *Ne* = 1,000 for P1, P2 and PA and to 1,000,000 for the ancestral population (equalling 1,000 times the size of P1 or P2). This very large ancestral population size ensured that the single mutation added on each genealogy (option -s 1) most likely occurred in the ancestral population and not after the divergence event (the branch where the mutation occurred being chosen with a probability proportional to its length relative to the total gene tree length). No migration was allowed between populations. The admixture rate (p_{adx}) was set at 0.5 (admixed population made of 50% of individuals from P1 and 50% from P2) for most scenarios but we also explored the influence of asymmetric admixture rates (p_{adx} = 0.1, 0.3) on a subset of situations.

Three different divergence times (*TD* = 250, 500 and 1,000 generations; or equivalently 0.0625, 0.125 and 0.25 for *TD*/4*Ne*) and two different admixture times (*TA* = 10 and 100 generations; 0.0025 and 0.025 for *TA*/4*Ne*) were considered. Given the size of the admixed and parental populations, *TA* = 10 is sufficiently low to minimize the effects of drift and mimic empirical situations where admixture is very recent or ongoing ("weak drift" hereafter), whereas *TA* = 100 represents relatively strong drift in all three populations after the admixture event and mimics empirical situations where admixture is more ancient and allelic frequencies for any given locus have diverged from the frequency observed just after admixture ("strong drift" hereafter).

Two hundred and fifty (250) independent SNP loci were simulated for all scenarios, assuming complete linkage equilibrium between markers. We chose to simulate only 250 SNP loci, a relatively low number given the current sequencing capacities, as a baseline in our study for at least three reasons: (a) to allow reasonable computation times, (b) to focus on situations where STRUCTURE might have reduced performances and (c) to approximate the level of genetic information carried by a few tens of microsatellite loci (Brumfield, Beerli, Nickerson, & Edwards, 2003; Morin et al., 2012; Morin, Luikart, Wayne, & the SNP workshop group, 2004; Narum et al., 2008) as many studies using STRUCTURE were based on microsatellite markers. This number of SNPs is also typical of many empirical situations: in a recent review, Janes et al. (2017) found that the majority of SNP studies used less than 100 loci. Many current-day RADseq studies are based on much larger numbers of loci, however, so we also generated data sets with 2,500 loci to check the effect of using a larger number of loci on a subset of situations. Pairwise F_{ST}'s between the two parental populations at sampling time were calculated using the approach of Weir and Cockerham (1984) to quantify their level of differentiation.

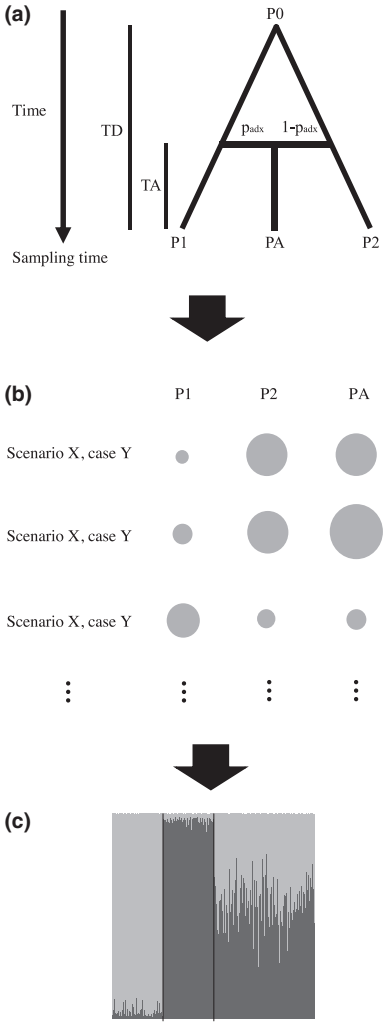


FIGURE 1 Stages of the study. (a) Simulations of populations. This first section shows a schematic representation of the simulated demographic scenario. *TD* = time of divergence; *TA* = time of admixture; P0 = ancestral population; P1 and P2 = parental populations; PA = admixed population; p_{adx} = gene flow from P1, that is initial proportion of individuals of PA coming from P1 at t = *TA*. At this stage, TA, TD, p_{adx} and the number of SNPs used are selected. (b) Sampling. At this stage, the sampling of the three populations is performed according to different scenarios and cases (see Table 1) which will define the sample sizes for each population (here represented as circles of different size). (c) STRUCTURE analyses. Each sampling case, represented by 100 data sets, is analysed using STRUCTURE. At this stage, the prior of individual ancestry is selected. See main text for details

2.2 | Sampling schemes

The different sampling schemes we explored were organized in "scenarios" and "cases" (Table 1 and Figure 1b). Each scenario represents

TABLE 1 Sampling scheme. Combinations of sample sizes organized in five scenarios, each containing six cases. P1 and P2 represent the sample sizes of parental populations, and PA represents that of the admixed population

		Case					
Scenario	Population	1	2	3	4	5	6
1	P1	100	100	100	100	100	100
	P2	100	100	100	100	100	100
	PA	10	33	50	100	200	300
2	P1	10	33	50	100	200	300
	P2	10	33	50	100	200	300
	PA	100	100	100	100	100	100
3	P1	10	33	50	100	200	300
	P2	100	100	100	100	100	100
	PA	100	100	100	100	100	100
4	P1	10	33	50	100	200	300
	P2	20	66	100	200	400	600
	PA	100	100	100	100	100	100
5	P1	10	33	50	100	200	300
	P2	20	66	100	200	400	600
	PA	10	10	10	10	10	10

a general situation regarding the way in which sample sizes can vary and each case corresponds to a particular combination of sample sizes for P1, P2 and PA. Each scenario thus contains six cases investigating the effect of gradual changes in the amplitude of sample size imbalance (see Appendix S1). These different cases allowed us to assess the effect of systematic changes in the sampling scheme within a given general scenario. We explored five scenarios giving a total of 27 different cases (some scenarios share identical cases; see Table 1). For each case, 100 data sets were simulated with *ms* and analysed with STRUCTURE. From here on to refer to particular sampling cases, we will use an expression of the form "SxCy," followed by the sample sizes of P1, P2 and Pa (e.g. scenario 4–case 3 will be denoted as "S4C3 (50-100-100)").

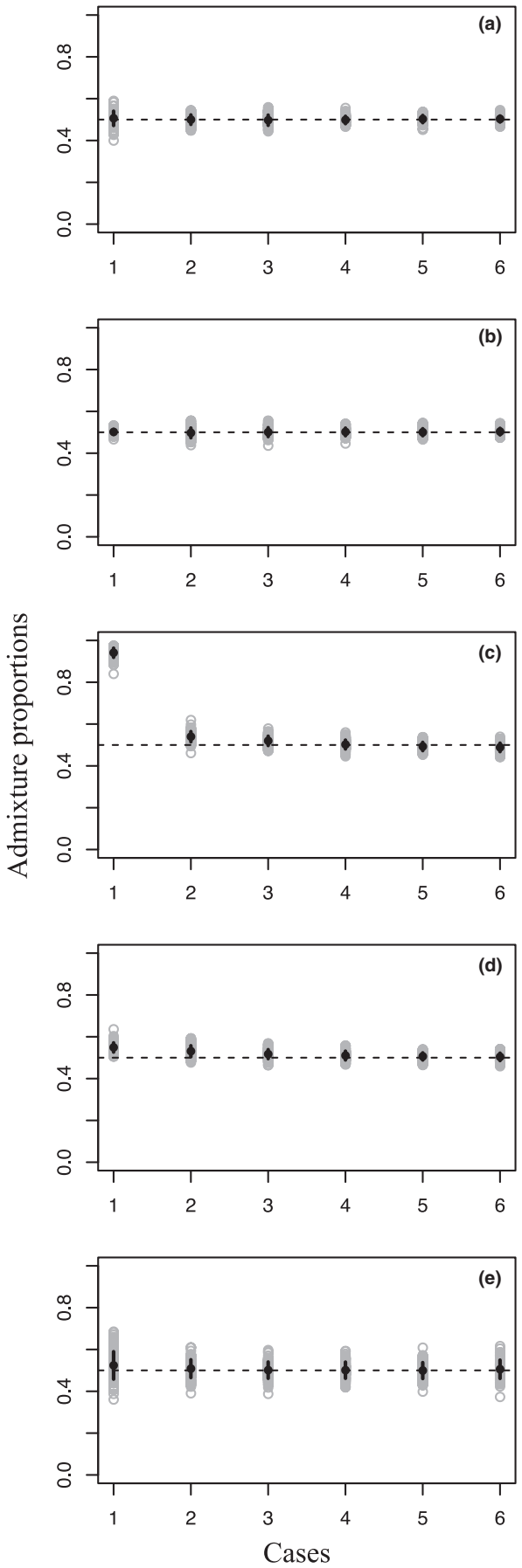
2.3 | STRUCTURE analyses

For each case, four independent analyses of STRUCTURE were run on each of the 100 simulated data sets for a number of clusters $K = 2$ using models with admixture, correlated allele frequencies and using the default value of the ancestry prior (single common parameter, ALPHA, in the Dirichlet distribution for priors on the contribution of each cluster to the total allelic pool, and initial alpha value ALPHA = 1), as done in most empirical studies and as recommended in the STRUCTURE manual, except for asymmetric admixture (see below for the use of alternative ancestry priors and other *K* values). Each chain of a STRUCTURE run consisted in a fixed burn-in of 10,000 iterations followed by 1,000 iterations. This value is quite small compared to the number of iterations used on empirical studies, but we chose it in order to keep the duration of the study more manageable. However, we made sure this methodological choice did not affect our results by (a) performing preliminary tests under different scenarios that showed that, once the chain had reached equilibrium after a sufficiently long burn-in, 1,000 iterations were sufficient to obtain consistent estimates of *Q*-values among runs for a given data set, (b) carefully checking that runs had converged (see Appendix S1) and increasing the burn-in from 10,000 to 100,000 iterations when it was not the case and (c) performing several runs for each data set and making sure results were consistent between runs. The run with the highest probability from the four performed ones was retained for further analyses.

In this study, the estimated individual admixture proportions (*Q*-value) always refer to the proportion of an individual genome coming from P1. The mean *Q*-value of a data set refers to the average of all individual *Q*-values from PA within the data set and the mean *Q*-value of a case to the average of all individual *Q*-values for that case (or equivalently the average of the mean *Q*-value across all data sets for that case). For each case, the standard deviation of the mean *Q*-values among data sets is reported (*SD* of the 100 mean *Q*-values), as well as the average of the variances of each data set's *Q*-values (average of the 100 between-individuals within-data sets variance values, see also legend of Figure 2).

Although it is true that longer drift times could make $K = 3$ the most adequate option for some of our cases (see below), from

FIGURE 2 Influence of sampling schemes in the mean admixture proportions estimated by STRUCTURE for $TD = 250$ and $TA = 10$. The results for 100 data sets are shown per case represented by grey circles. Black dots and solid lines represent the mean and standard deviation for each case, respectively, and the horizontal dashed line represents the expected $Q = 0.5$. (a–e) correspond to scenarios 1–5, respectively. Next to the figure of each scenario is a table showing sampling scheme and average values of admixture proportions, their standard deviations (*SD*) and the average of the within-data set variances of *Q*-values of individuals from PA



Scenario 1

	Cases					
	1	2	3	4	5	6
P1	100	100	100	100	100	100
P2	100	100	100	100	100	100
PA	10	33	50	100	200	300
Mean	0.506	0.500	0.497	0.498	0.502	0.503
SD	0.034	0.023	0.024	0.018	0.017	0.015
Variance	0.013	0.013	0.012	0.011	0.011	0.010

Scenario 2

	Cases					
	1	2	3	4	5	6
P1	10	33	50	100	200	300
P2	10	33	50	100	200	300
PA	100	100	100	100	100	100
Mean	0.502	0.497	0.501	0.502	0.500	0.503
SD	0.012	0.023	0.021	0.018	0.017	0.016
Variance	0.010	0.011	0.011	0.012	0.012	0.012

Scenario 3

	Cases					
	1	2	3	4	5	6
P1	10	33	50	100	200	300
P2	100	100	100	100	100	100
PA	100	100	100	100	100	100
Mean	0.941	0.540	0.520	0.502	0.493	0.488
SD	0.022	0.024	0.022	0.021	0.020	0.020
Variance	0.009	0.012	0.012	0.012	0.012	0.012

Scenario 4

	Cases					
	1	2	3	4	5	6
P1	10	33	50	100	200	300
P2	20	66	100	200	400	600
PA	100	100	100	100	100	100
Mean	0.549	0.531	0.517	0.510	0.506	0.504
SD	0.022	0.026	0.022	0.021	0.017	0.018
Variance	0.010	0.011	0.012	0.012	0.012	0.012

Scenario 5

	Cases					
	1	2	3	4	5	6
P1	10	33	50	100	200	300
P2	20	66	100	200	400	600
PA	10	10	10	10	10	10
Mean	0.524	0.508	0.501	0.501	0.499	0.505
SD	0.066	0.043	0.039	0.039	0.038	0.044
Variance	0.017	0.014	0.013	0.013	0.012	0.012

now on we will refer to any deviation from the expected value of 0.5 in the estimated average admixture proportions as a bias, as our study is focused on the estimation of admixture proportions when considering $K = 2$, as it is usually done in empirical uses of STRUCTURE.

2.4 | "Best-K" estimation

In empirical situations where the history and structure of the sampled populations are unknown, STRUCTURE is routinely used to assess how many populations have been sampled and if any originates from admixture events. To do so, the program infers the "best-*K*," that is the most likely number of homogeneous genetic clusters present in the data, using the differences in probabilities among runs with different *K* values, or with more sophisticated technics like the Δ*K* method of Evanno, Regnaut, and Goudet (2005). In our simulations, the "best-*K*" value recovered by STRUCTURE may vary depending on the history of the populations. The inferred "best-*K*" value may be 2 if the admixed population is detected as admixed, and both parental populations are identified as different clusters, or 3 if the admixed population is identified as a third independent cluster. For example, contrary to situations with weak drift, strong drift implies that allelic frequencies at sampling in the admixed and parental populations do not reflect allelic frequencies from the admixed and parental populations at admixture time. It is thus expected that, in such situations, STRUCTURE may consider the parental and admixed populations as three independent genetic clusters. If so, "forcing" STRUCTURE to use a model with two genetic clusters ($K = 2$), as is often done when admixture is suspected, may lead to inaccurate results. In such cases, STRUCTURE may cluster individuals more or less arbitrarily for $K = 2$, for example (a) grouping the two parental samples together and the admixed sample aside, or (b) grouping one parental population with PA, and the other parental population aside, both cases resulting in estimates close to $Q = 1$ or $Q = 0$ for all admixed individuals. We will refer to such behaviour as the "forcing $K = 2$ with three genetic clusters" case.

We thus analysed all data sets with $K = 2$ as described previously to infer admixture proportions but also looked for the "best-*K*" value inferred by STRUCTURE with the Δ*K* method, exploring possible values of *K* from 1 to 5. This allowed us to examine how reliable is the inference of the "best-*K*" in the situations we explored, and how they relate to the estimation of admixture proportions under $K = 2$. We used the Δ*K* method because preliminary simulations showed that (a) it was easier to implement than considering the probability of the data $\ln\Pr(X|K)$ for different *K* values because of higher variances and/or the presence of a plateau for the probabilities with *K* values larger or equal to 2; and (b) in our case, the results were concordant with those obtained by visual inspection of the probabilities of the data. We did not use the new method proposed by Verity and Nichols (2016) to find the "best-*K*," because this was not our main goal and, according to the authors, this method is computationally demanding.

2.5 | Effect of the number of markers, admixture rates and alternative ancestry priors

To assess the effect of the number of markers on the estimation of admixture proportions, we simulated data sets of 2,500 SNPs for three cases that showed different types of biases in our results: a case where parental populations had equal sample sizes and PA was represented by a small sample (S1C1 [100-100-10]), a case where the sample sizes of the parental populations were equal and the sample size of PA was substantially larger (S2C1 [10-10-100]), and a case where the sample sizes of the parental populations were unequal and the sample size of PA was relatively large (S3C1 [10-100-100]). These cases were tested using $TD = 250$ and $TA = 10$ and 100 to assess the influence of weak and strong drift, respectively.

We also ran simulations with different initial admixture rates (p_{adx}) to assess how STRUCTURE estimates admixture proportions different from 0.5 in various situations. Values of 0.1 and 0.3 were both tested on scenarios 1 and 3 with $TA = 10$ and $TD = 500$. Since important biases were observed under these conditions, we selected the cases with the strongest biases (for $p_{adx} = 0.3$, S3C1 [10-100-100]; for $p_{adx} = 0.1$, S1C6 [100-100-300], S3C1 [10-100-100] and S3C6 [300-100-100]) and repeated the simulations increasing the number of markers to 2,500. Fifty data sets were analysed for each of the repeated cases. This allowed us to explore whether increasing the number of markers would correct the combined effects of sampling schemes and asymmetric rates of admixture.

Wang (2017) recently highlighted that most studies assessing the performance of STRUCTURE, or using STRUCTURE in empirical situations, used the default ancestry prior option. With this default option, STRUCTURE assumes that each allele has its ancestry originating from each of the *K* populations with an equal probability of 1/*K*. He showed how using an alternative prior, which allows STRUCTURE to model *K* different probabilities for each cluster, increased the accuracy of the "best-*K*" inference and individual assignments in cases of unbalanced sampling. We thus compared the accuracy of STRUCTURE's results when using the default and alternative ancestry priors in the three scenarios with unbalanced sampling (scenarios 3, 4 and 5). These tests were performed for $TD = 500$ and $TA = 10$ and 100 (weak and strong drift). We also tested the effect of the alternative prior for asymmetric admixture rates (i.e. $p_{adx} = 0.1, 0.3$) as it may be especially adapted for such situations (Pritchard, Wen, & Falush, 2010). Additionally, Wang (2017) also showed that when sampling is unbalanced and many ancestral populations are included, the use of the uncorrelated allele frequency model (default = correlated allele frequency model) and a lower value for ALPHA (the initial value of the Dirichlet prior parameters used to calculate an individual's ancestry) of approximately 1/*K* (default = 1) improved the performance of STRUCTURE. Although our study only considered two parental populations, we tested the influence of using the alternative prior together with the uncorrelated allele frequency model and ALPHA = 0.5 on scenarios 3, 4 and 5 with $TD = 500$ and $TA = 10$ and 100.

3 | RESULTS

3.1 | Effect of unbalanced sampling

In this section, we will comment on results from simulations with a short drift time (*TA* = 10) to assess the effect of different sampling schemes alone. Average F_{ST} estimates between P1 and P2 among scenarios in our simulations were of 0.09 (range = 0.066–0.117) for *TD* = 250, 0.16 (range = 0.126–0.199) for *TD* = 500 and 0.28 (range = 0.225–0.336) for *TD* = 1,000.

Cases with large and equal sample sizes of P1 and P2 and varying sample sizes of PA (scenario 1) showed very accurate results, but the precision slightly decreased with smaller sample sizes of PA (Figure 2a). Cases with large sample sizes of PA and variable but even sample sizes of P1 and P2 (scenario 2) also showed accurate results. No clear pattern for changes in precision was observed (Figure 2b).

Cases where P1 and P2 samples had unequal sizes but the PA had the same large sample size as P2 (scenario 3) showed the most biased results. S3C1 (10-100-100) was particularly bad; however, estimates of admixture proportions gained accuracy as the sample size of P1 became closer to the one of P2 and PA. Then, the accuracy decreased again but to a lesser extent when the sample size of P1 got larger than P2's and PA's. Hence, average admixture proportions were never fully accurate except in case 4, where the three populations had equal sample sizes. The bias always overestimated the contribution of the population with the smaller sample (Figure 2c).

In cases with unequal sample sizes of pure and admixed populations with a constant large PA sample (scenario 4, Figure 2d), the accuracy depended on the proportional differences between the sample sizes of P1 and P2 and the sample size of PA. As the sample size of P2 was never more than twice the size of P1, the bias was not as strong as in S3C1. Again, the contribution of the population with the smaller sample size was overestimated.

Finally, cases with unequal sample sizes of pure and admixed populations with a constant small PA sample (scenario 5) showed slightly more accurate results compared to scenario 4 (which also had cases with unequal sample sizes of pure and admixed populations but a constant large PA sample), but the precision was lower than in all other scenarios (Figure 2e). Reducing the sample size of PA globally increased accuracy when parental samples were unequal but, as expected, also reduced precision.

In summary, we found that (a) the accuracy of admixture proportion estimates can be seriously affected by unbalanced sampling of the parental populations, especially when the sample size is small for one of them, and (b) estimated admixture proportions always overestimate the contribution of the parental population with the smaller sample size. Sample sizes of PA that differed from the parental populations did not affect the accuracy much in cases where parental populations had equal sample sizes (i.e. scenarios 1 and 2), but a small sample size of PA relative to the parental populations increased the accuracy and lowered the precision when sample sizes of parental populations were unequal (i.e. scenarios 4 and 5).

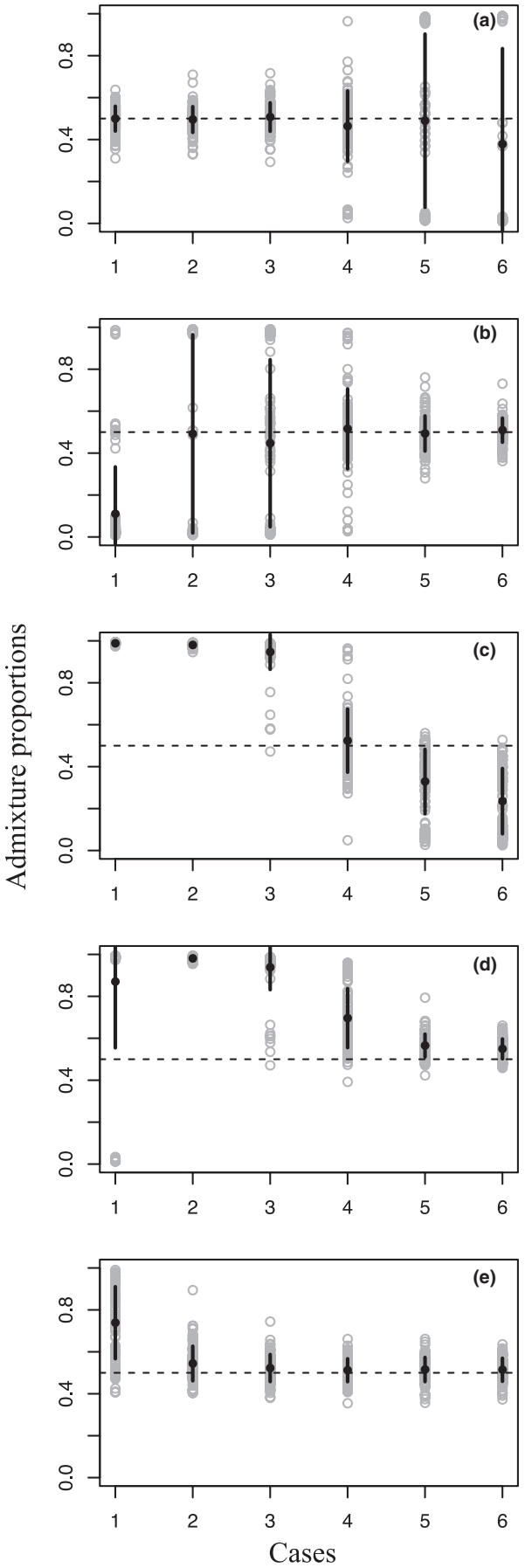
The results showed the same patterns for all *TD*s but as expected, a longer *TD* clearly increased the accuracy and precision of the results, reducing the biasing effect of unbalanced sampling (Figures S1 and S2).

3.2 | Effect of strong drift

Admixture proportions estimated by STRUCTURE under strong drift (*TA* = 100, Figure 3) were more variable, and the effects of uneven sampling were often stronger. When differentiation between the parental populations was the lowest (*TD* = 250, Figure 3) and they showed equal sample sizes (scenarios 1 and 2), the average admixture proportions were highly variable around the value of 0.5. Moreover, in the cases where the sample size of PA was equal to or larger than that of P1 and P2, we also obtained admixture proportions close to 0 and 1 (e.g. Figure 3, S1C4–6, S2C1–4), illustrating the behaviour of "forcing *K* = 2 with three genetic clusters." To better understand the origin of these results, we looked at particular data sets in some cases showing admixture proportions close to 0 and 1 (S1C5 [100-100-200] and S2C1 [10-10-100], Figure 4) and confirmed three different patterns: (a) P1 and P2 assigned to two separate clusters and PA showing admixture proportions close to 0.5 (Figure 4a, d); (b) PA and one of the parental populations assigned to a single cluster and the other parental population assigned to another cluster, giving admixture proportions close to 0 or 1 (Figure 4b, e); and (c) PA assigned to its own cluster, while the parental populations were clustered together (Figure 4c, f). This last configuration always gave admixture proportions close to 0, biasing the average admixture proportions of the whole case towards 0. This case is more common when the sample size of PA is substantially larger than for P1 and P2 (e.g. S1C6 [100-100-300] and S2C1 [10-10-100]).

When the sample sizes of pure populations were unequal (scenarios 3–5), the patterns of admixture proportions were generally similar to those observed under weak drift (*TA* = 10), but a higher variability of the results made the trends less clear and the "forcing *K* = 2 with three genetic clusters" behaviour led to more estimates of *Q* close to 1 or 0. As with even sampling, proportionally larger sample sizes of PA tended to cause PA individuals to be assigned to their own cluster, or PA to be clustered with one of the pure populations because of the "forcing *K* = 2 with three genetic clusters" behaviour (e.g. S4C1 [10-20-100], Figure 3d). Finally, as observed with *TA* = 10 (Figure 2, Figures S1 and S2), increasing *TD* from 250 to 500 and 1,000 increased the accuracy and precision of the results, although biases were still present (Figure 3, Figures S3 and S4).

To sum up, the long periods of drift greatly affected the results, but did not globally modify the way different sampling schemes biased the results. Scenarios 1 and 2, for which the sample sizes of pure populations were equal, remained more accurate, while scenarios 3, 4 and 5, with unequal sample sizes of parental populations, always showed stronger biases. By strengthening the effect of the "forcing *K* = 2 with three genetic clusters" behaviour, proportionally larger samples of PA always increased the biases caused by uneven sampling of the pure



Scenario 1	Cases					
	1	2	3	4	5	6
P1	100	100	100	100	100	100
P2	100	100	100	100	100	100
PA	10	33	50	100	200	300
Mean	0.500	0.495	0.507	0.464	0.490	0.378
SD	0.059	0.061	0.068	0.168	0.413	0.455
Variance	0.014	0.012	0.013	0.012	0.005	0.002

Scenario 2	Cases					
	1	2	3	4	5	6
P1	10	33	50	100	200	300
P2	10	33	50	100	200	300
PA	100	100	100	100	100	100
Mean	0.110	0.492	0.447	0.516	0.493	0.509
SD	0.224	0.472	0.398	0.190	0.084	0.058
Variance	0.006	0.002	0.005	0.012	0.012	0.012

Scenario 3	Cases					
	1	2	3	4	5	6
P1	10	33	50	100	200	300
P2	100	100	100	100	100	100
PA	100	100	100	100	100	100
Mean	0.988	0.980	0.947	0.524	0.329	0.235
SD	0.004	0.008	0.083	0.151	0.153	0.156
Variance	<0.001	0.002	0.005	0.013	0.013	0.014

Scenario 4	Cases					
	1	2	3	4	5	6
P1	10	33	50	100	200	300
P2	20	66	100	200	400	600
PA	100	100	100	100	100	100
Mean	0.870	0.981	0.938	0.696	0.565	0.549
SD	0.316	0.008	0.106	0.141	0.055	0.047
Variance	0.001	0.002	0.004	0.013	0.013	0.012

Scenario 5	Cases					
	1	2	3	4	5	6
P1	10	33	50	100	200	300
P2	20	66	100	200	400	600
PA	10	10	10	10	10	10
Mean	0.739	0.544	0.522	0.512	0.516	0.514
SD	0.172	0.083	0.065	0.055	0.058	0.055
Variance	0.015	0.014	0.015	0.013	0.012	0.012

FIGURE 3 Influence of sampling schemes in the admixture proportions estimated by STRUCTURE for *TD* = 250 and *TA* = 100. See legend in Figure 2 for details

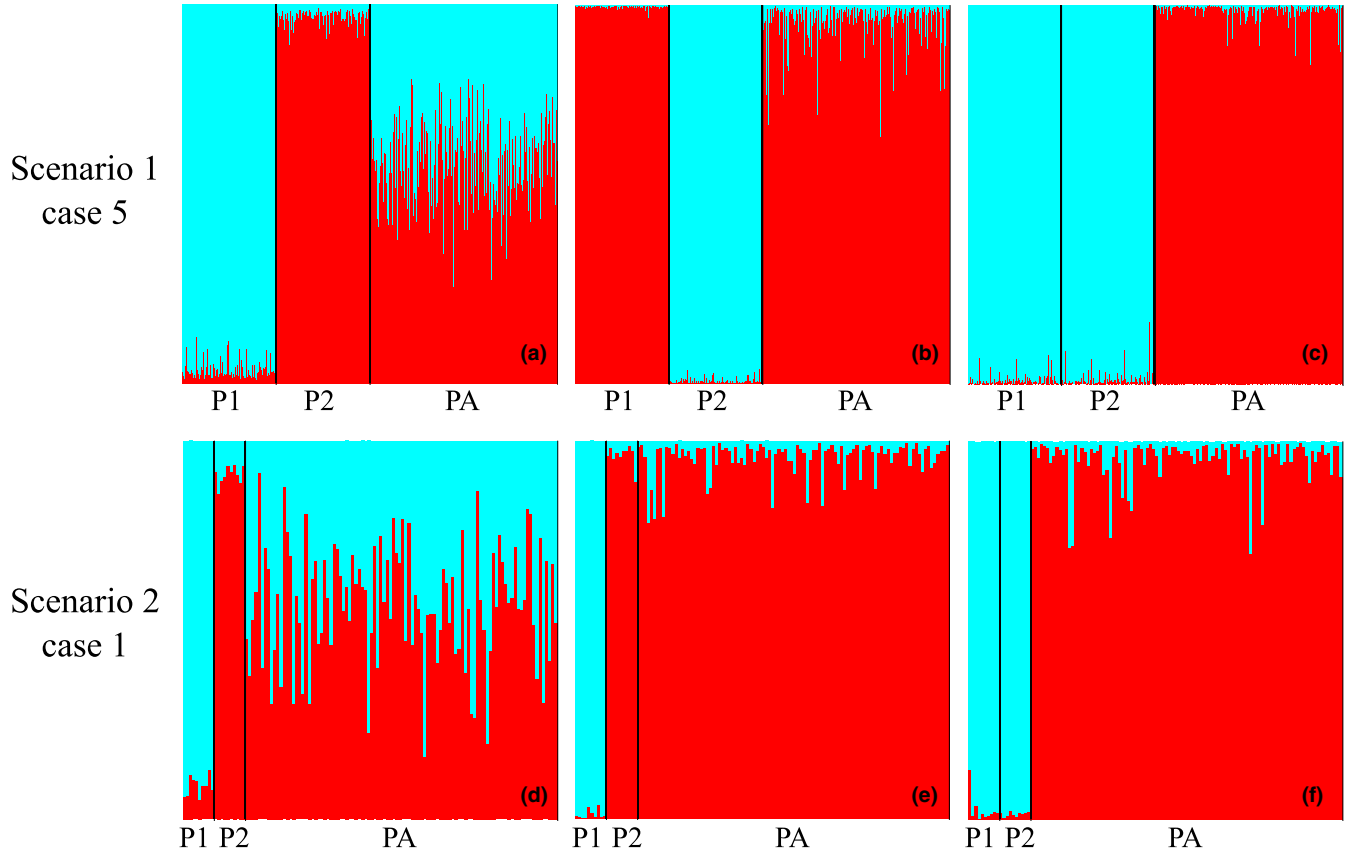


FIGURE 4 Clustering and individual admixture proportions in particular data sets. Three different patterns of admixture and clustering were found for S1C5 (a–c) and S2C1 (d–f). Each individual bar represents one individual. Different colours represent affinity with a given cluster ($K = 2$). P1 and P2 = parental populations; PA = admixed population. (a and d) show the expected pattern of admixture for PA (~0.5), (b and e) show cases where PA is clustered with one of the parental populations, (c and f) show cases where individuals from PA are assigned to their own cluster, while individuals from the parental populations are assigned to a different cluster together

populations and, contrary to situations with weak drift, even generated important biases in cases of even sampling of the pure populations.

3.3 | "Best-*K*" inference

Our analysed cases only showed "best-*K*" values of either 2 or 3. When drift was low (*TA* = 10), "best-*K*" values of 2 were obtained for the great majority of simulations in all scenarios and cases, most of them reaching occurrences of 100% (Table 2). The different levels of differentiation between parental populations had the expected effect: shorter *TD*s decreased the proportion of simulations for which the inferred "best-*K*" value was 2, indicating that P1, P2 and PA were identified as three different clusters in more data sets than for higher differentiation levels.

The effect of the sampling scheme on the "best-*K*" inferences was much stronger when drift was strong (*TA* = 100) and a clear pattern was apparent: whenever the sample size of PA was proportionally large compared to that of the parental populations, a lower percentage of "best-*K*"=2 was obtained (Table 2).

However, higher percentages of "best-*K*"=2 did not guarantee that the two parental populations were correctly identified. For example, for S2C1 (10-10-100) with *TD* = 250 65% of the inferences resulted in "best-*K*"=2 (Table 2); however, our calculated admixture proportions in that particular case show an average admixture proportion close to 0 when forcing $K = 2$ (Figure 3b), suggesting that PA was often being recognized as a cluster on its own, or being merged with P2. Similar patterns are observed when sampling is uneven; for example, S3C1 (10-100-100) with *TD* = 250 shows 100% of "best-*K*" = 2 inferences (Table 2), but our previous results evidence how that particular case shows average admixture proportions close to 1 (Figure 3c).

3.4 | Effect of the number of SNPs

We selected three cases (S1C1 [100-100-10], S2C1 [10-10-100] and S3C1 [10-100-100]) for both admixture times (*TA* = 10, 100) and *TD* = 250 to explore the effect of the number of markers (250 vs. 2,500) on the accuracy and precision of the inferred admixture proportions.

Under a short drift time (*TA* = 10), S1C1 already showed accurate results when using 250 SNPs (Figures 2a and 5a), but using 2,500 SNPs improved the precision (Figure 5b), as expected. S2C1 also

TABLE 2 Percentages of data sets with inferred "best-*K*" = 2

		Cases											
		TA = 10						*TA* = 100					
TD	Scenarios	1	2	3	4	5	6	1	2	3	4	5	6
250	1	100	99	100	99	99	96	100	85	79	58	13	5
	2	81	98	99	100	100	100	65	15	18	49	76	89
	3	99	100	100	99	99	99	100	68	38	50	58	60
	4	92	98	100	100	100	100	83	64	34	56	88	94
	5	95	100	100	100	100	100	87	100	98	100	100	100
500	1	100	100	100	100	99	98	98	99	95	91	69	63
	2	89	100	100	100	100	100	42	61	73	90	96	99
	3	88	100	99	100	100	100	100	62	80	89	95	97
	4	98	99	100	100	100	100	72	67	81	90	94	97
	5	100	100	100	100	100	100	99	100	100	100	100	100
1,000	1	100	100	100	100	100	100	100	100	97	93	91	88
	2	99	100	100	100	100	100	78	89	91	95	100	98
	3	100	100	100	100	100	100	86	96	94	94	98	99
	4	100	99	100	100	100	100	84	90	92	97	98	99
	5	100	100	100	100	100	100	99	100	100	100	100	100

Note: Values represent percentages of data sets for which "best-*K*" = 2 is inferred across all different scenarios and cases, for all *TD*s and *TA*s. Values of "best-*K*" were inferred with the method of Evanno et al. (2005).

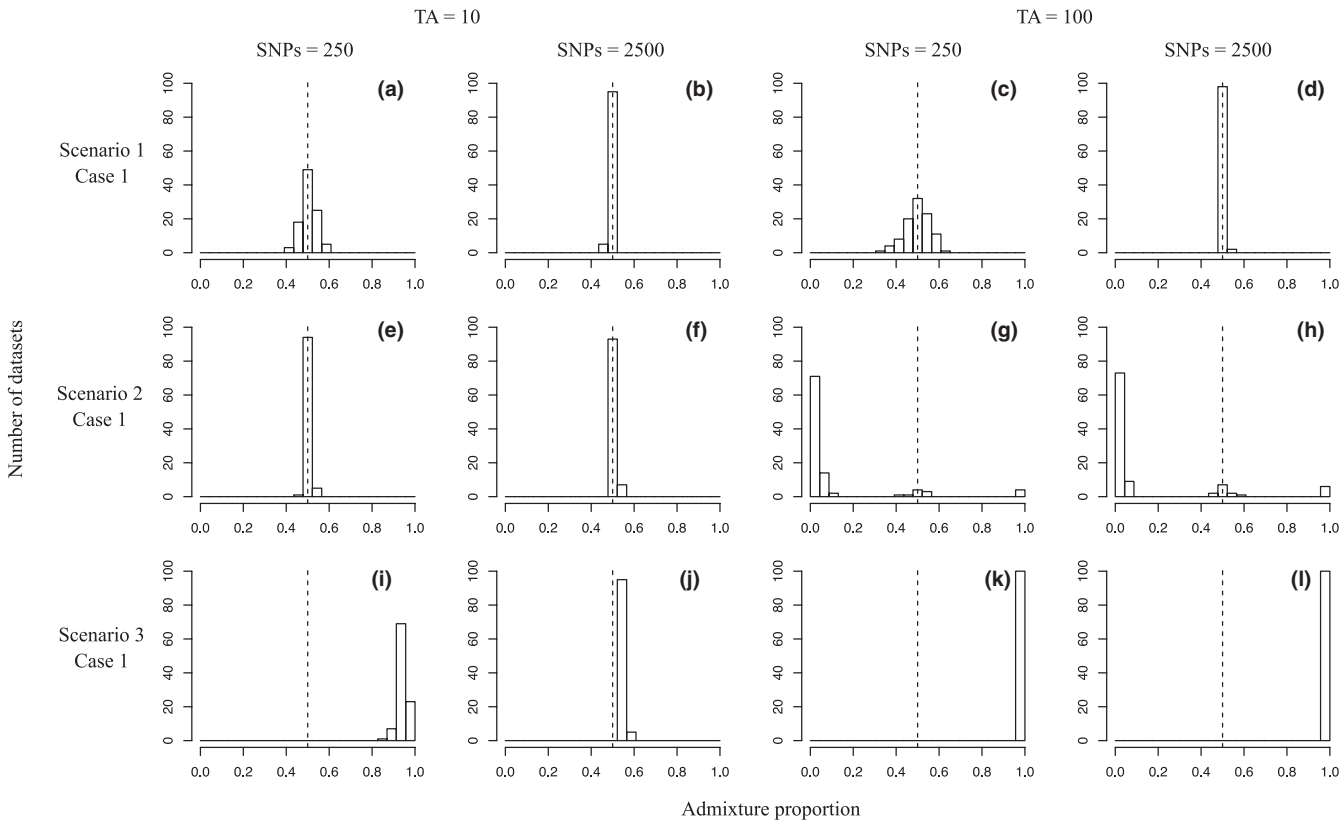


FIGURE 5 The effect of the number of SNPs on estimated admixture proportions. Each row corresponds to the indicated case. Number of SNPs used and *TA* values are shown on top of each column. *TD* = 250. The vertical dashed line indicates the expected average value of *Q* = 0.5

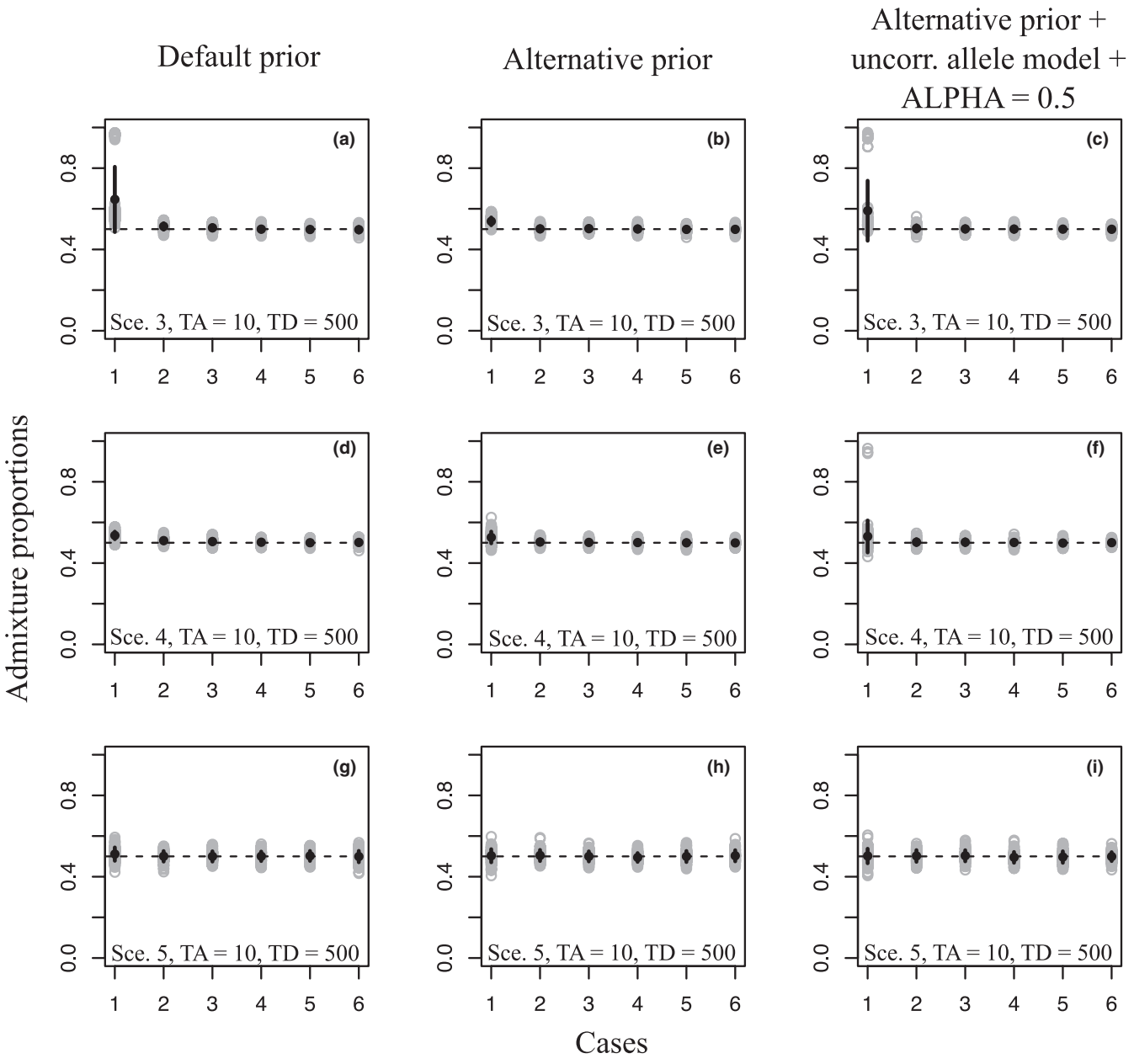


FIGURE 6 Admixture proportions obtained with different priors of individual ancestry. Grey circles indicate average admixture proportions found for each data set. Black dots and solid lines indicate means and standard deviations for the admixture proportions in each case, respectively, and the horizontal dashed line represents the expected $Q = 0.5$. Each panel represents different combinations of scenarios and parameters as indicated in the figure. *TA* = 10 and *TD* = 500 were used

showed accurate and precise results using 250 SNPs (Figures 2b and 5e) and remained so when using 2,500 SNPs (Figure 5f). S3C1 was strongly biased when using 250 SNPs (Figures 2c and 5i) but became more accurate and precise with 2,500 markers, although a small bias was still present (Figure 5j).

Under a long drift time (*TA* = 100), the effect of increasing the number of markers was weak for cases with an initial bias. S1C1 showed accurate results with 250 SNPs (Figures 3a and 5c) but increasing the number to 2,500 increased the precision (Figure 5d). S2C1 showed strongly biased results when using 250 SNPs (Figures 3b and 5g), and increasing the number of markers did not correct this bias (Figure 5h). S3C1 showed strongly biased results with 250 SNPs (Figures 3c and 5k), and using 2,500 SNPs did not make the results accurate (Figure 5l).

3.5 | Comparisons between different priors for individual ancestry

In the situation of weak drift and moderate population divergence (*TA* = 10 and *TD* = 500), the alternative ancestry prior improved the accuracy of admixture estimation for scenario 3, especially in case 1 where the sample size of P1 was very small compared to that of P2 and PA (Figure 6a,b). However, this case still showed a bias towards

P1 in the inferred admixture proportion, as with the default ancestry prior option. The use of the uncorrelated allele frequency model and an initial ALPHA = 0.5 did not improve the results even compared to the default prior (Figure 6c). For scenario 4, using the alternative prior did not improve accuracy (Figure 6d,e) but actually decreased the precision of the results for case 1, where the sample sizes of pure populations were the smallest. No improvement in accuracy was observed when using the uncorrelated allele frequency model and an initial ALPHA = 0.5, but the first case became particularly imprecise (Figure 6f). For scenario 5, no differences in accuracy were observed between default and alternative priors, or even when changing the allele frequency model and the initial ALPHA value (Figure 6g–i).

When the effect of drift was strong (*TA* = 100), no improvement in accuracy was observed in any of the three scenarios as the same biases obtained using the default prior were still present (Figure S5).

To sum up, the use of the uncorrelated allele frequency model and an initial ALPHA = 0.5 did not improve the inferences in any of the scenarios and cases. The effect of the alternative prior of individual ancestry was limited, as it only improved the accuracy of the case with the most unbalanced sampling (S3C1 [10-100-100]).

3.6 | Effect of asymmetric admixture

STRUCTURE did not perform well when admixture was asymmetric even in scenarios where it was found to give good estimates under symmetric admixture. In general, the performance of the inferences decreased as the asymmetry of admixture increased, admixture proportions being almost always biased towards the population that contributed the most.

For cases with large and equal sample sizes of P1 and P2 and varying sample sizes of PA (scenario 1), when $p_{adx} = 0.5$, STRUCTURE gave good estimates (Figure 7a). An intermediate degree of asymmetry ($p_{adx} = 0.3$) still produced accurate results (Figure 7c). When the asymmetry was strong ($p_{adx} = 0.1$), the smaller admixture proportions were underestimated and this bias increased as the proportional sample size of PA increased (Figure 7e). In cases where P1 and P2 samples have unequal sizes but the PA has the same large sample size as P2 (scenario 3), when results where poor for $p_{adx} = 0.5$ due to the sampling scheme (case 1, Figure 7b), an intermediate degree of asymmetry ($p_{adx} = 0.3$) actually improved the results as only a slight bias caused by the sampling scheme is noticed (Figure 7d). A stronger asymmetry ($p_{adx=}0.1$) resulted in stronger underestimation of the smaller contribution, the bias increasing with the sample size of P1 (Figure 7f).

We repeated the analysis of the cases with the most biased results, increasing the number of SNPs to 2,500. All results became more accurate, with inferred admixture proportions closer to the expected values (S3C1 [10-100-100], $p_{adx} = 0.3$, mean = 0.312; S3C1 [10-100-100], $p_{adx} = 0.1$, mean = 0.102; S3C6 [300-100-100], $p_{adx} = 0.1$, mean = 0.094; S1C6 [100-100-300], $p_{adx} = 0.1$, mean = 0.086).

We used the alternative prior of individual ancestry with 250 SNPs to test whether it could solve the problems associated with asymmetric p_{adx} on scenarios with unbalanced sampling of parental populations and a small number of markers, as suggested in the STRUCTURE manual. Using the alternative prior did not eliminate the

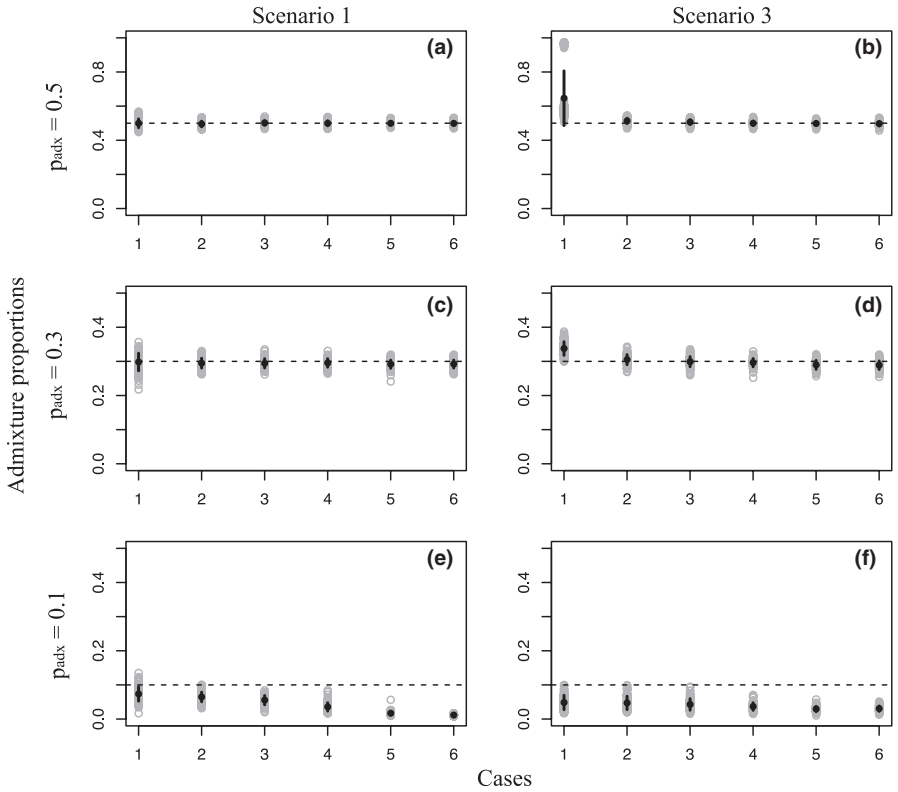


FIGURE 7 The effect of asymmetric admixture rates p_{adx}. Black dots and solid lines represent the mean and standard deviation for each case, respectively, and the horizontal dashed line represents the expected Q-value. Each panel represents different combinations of scenarios and p_{adx} values as indicated in the figure. Results from individual data sets are represented by grey circles. *TA* = 10 and *TD* = 500 were used

underestimation found for $p_{adx} = 0.1$ and 0.3 (Figure S6). Similarly, as with the default prior, the underestimation was stronger for $p_{adx} = 0.1$ (Figure S6a–c). More accurate results were obtained using the alternative prior for $p_{adx} = 0.3$, but small biases could still be observed (e.g. S4C1 [10-20-100], Figure S6e).

4 | DISCUSSION

As shown by numerous theoretical and empirical studies, STRUCTURE accurately estimates admixture proportions in a large range of historical scenarios and parameter values. However, STRUCTURE's performance is sensitive to the time elapsed since admixture (as expected given prior works on the subject) and, perhaps less intuitively, to the sampling scheme and the level of asymmetry in the admixture rates. The effects of drift following admixture are not alleviated by increasing the number of markers, but the effects of sampling scheme and asymmetry of admixture are mainly restricted to data sets employing a moderate number of markers. These limits are apparently not well known to empirical users of the software. Although some effects of sampling scheme have been previously documented (e.g. Kalinowski, 2011; Meirmans, 2019; Puechmaille, 2016), our study is, to the best of our knowledge, the first to document how the proportion of pure and admixed individuals in a sample affects STRUCTURE's abilities to estimate admixture proportions.

4.1 | Effect of drift and sampling scheme on the "Best-*K*" inference

"Ancient" admixture is a recurrent subject in hybridization and speciation studies (Good et al., 2008; Miller et al., 2012). Our results show that ancient admixture can be difficult to identify when the empirical test to distinguish it from a scenario of pure population divergence is based on finding the "best-*K*" value. If a sampled population originates from past admixture between two sampled populations, a correct interpretation of STRUCTURE results in terms of admixture will only be straightforward if STRUCTURE infers that $K = 2$ and that the admixed population is indeed admixed. However, in cases of long drift relative to divergence times, we have shown that STRUCTURE may often identify the admixed population as an independent cluster ($K = 3$) when using a low number of markers, indicating that drift had enough time to alter the information used by STRUCTURE to infer the past admixture history of the population. While $K = 3$ can be seen as the correct population structure when admixture is ancient, this result shows that using the best-*K* to distinguish between a scenario of pure vicariance and a scenario of past admixture is not valid if admixture is not very recent (see also Janes et al., 2017, Cullingham et al., 2020 for other issues related to the "best-*K*" estimation).

Besides, even under a situation with a short drift time, the sampling scheme can affect STRUCTURE's ability to infer the correct population structure. When drift was weak ($TA = 10$), STRUCTURE inferred a $K = 2$ for most data sets (>95%) but in some cases of unbalanced sampling where the sample size of PA is much larger than for P1 and P2, the "best-*K*" can be estimated to be 3 (Table 2). In empirical situations, even recent admixture events can thus be "missed" in cases of unbalanced sampling if STRUCTURE is the only test of admixture. Moreover, admixture time and sampling scheme interact strongly to affect the probability of finding $K = 2$ (Table 2). Finally, STRUCTURE can infer a flawed population structure even when the correct *K* is inferred, failing to delimit correctly the parental and admixed populations (see below). The best way to maximize the chances to identify past admixture may thus be (a) to oversample the candidate parental populations or to subsample the candidate admixed population when hypotheses on population history exist, as situations where the sample of PA is small compared to the parental samples always gave better results in our simulation settings and (b) always check different *K* values in STRUCTURE analyses since several *K* values, and not a single "best-*K*" value, can be biologically informative and reveal the various effects of sampling and drift on STRUCTURE results.

4.2 | Effect of sampling schemes, divergence and drift on admixture proportions

Our results showed that, as expected, the amount of differentiation between parental populations and the drift experienced by the admixed population had positive and negative effects on STRUCTURE's accuracy, respectively (see also Kalinowski, 2011). Strong differentiation of parental populations and recent admixture events can decrease the effect of unbalanced sampling. On the contrary, low differentiation before admixture and/or large amount of drift after admixture often caused extreme biases when assuming $K = 2$, even in cases of balanced sampling (e.g. Figure 3) independently of the number of markers (Figure 5g,h).

Besides the influence of divergence and drift, sampling schemes had two main effects on the inference of admixture proportions, especially when using a moderate number of markers. First, when the sampling of pure populations was uneven, average admixture proportions were always biased towards the pure population with the smaller sample size, overestimating its contribution to the admixed population. This pattern was observed for both recent and ancient admixture events. Second, the performance of STRUCTURE was sometimes poor when the sample size of the admixed population was proportionally large compared to the parental ones, even when the sample sizes of parental populations were equal and especially for ancient admixture events. In such situations, STRUCTURE often considered the admixed population, or a combination of the admixed and one parental population, as a single cluster (Figure 4), resulting in estimated admixture proportions close to 0 or 1 and greatly affecting the inferences.

Intuitively, researchers often aim to have large samples in order to obtain better resolution in statistical analyses, without necessarily considering proportionality. Previous studies suggested that STRUCTURE should not be affected by the proportion of hybrids in the sample (Marie et al., 2011, but see Vähä & Primmer, 2006). Based

on our results, this is not always the case, since a relatively large proportion of admixed individuals relative to pure individuals biased the estimated admixture proportions even in cases where pure populations were balanced (e.g. S2C1 [10-10-100]).

4.3 | Effect of asymmetric rates of admixture

STRUCTURE's accuracy and precision in estimating admixture proportions is greatly affected by the contribution levels of parental populations to the admixed one (asymmetric admixture rates). When using a moderate number of markers, the proportion of the genome inferred as coming from the parental population with the smaller contribution will be underestimated, this underestimation increasing with the asymmetry in the admixture rates. Using a high number of markers alleviates this underestimation when drift is low, but for ancient admixture events this bias could remain strong independently of the number of markers used. Most simulation studies have assumed equal admixture rates between parental populations in their experimental designs, making this pattern impossible to detect. This pattern poses a problem for empirical studies as highly asymmetric admixture events are common in natural and domestic populations (Elmer, Recknagel, Thompson, & Meyer, 2012; Muranishi, Tamaki, Setsuko, & Tomaru, 2013; Papa & Gepts, 2003). STRUCTURE's admixture proportions are often used as a hybrid index to build clines in hybrid zone analyses (Arntzen et al., 2016; Dufresnes et al., 2018; Wielstra et al., 2017). Clearly, underestimating low admixture proportions will result in modelling steeper clines and narrower hybrid zones than in reality so studies of hybrid zones should use a large number of markers.

4.4 | Effect of the number of SNPs and different priors for individual ancestry

Our results showed that increasing the number of markers increased the accuracy and precision of the estimation of admixture proportions by STRUCTURE, as expected. Although the benefits of increasing the number of markers have been previously reported (e.g. Vähä & Primmer, 2006), we showed that biases caused by uneven sampling can still persist for extremely unbalanced sampling schemes and when drift has been strong enough after the admixture event (Figure 5k,l).

Since Wang (2017) showed how the use of different prior parameters over default ones in STRUCTURE improved the inference of the number of clusters as well as individual assignment to each cluster under an island model of population structure (see "Section 2"), we also tested the effect of different prior parameters for individual ancestry on the estimation of admixture proportions using a small number of markers. Our results showed only a moderate improvement in a single case of recent admixture with highly unbalanced sampling between parental populations (i.e. S3C1 [10-100-100]). This suggests that the use of this alternative prior may sometimes alleviate biases in cases of high unbalance in sampling schemes, but that the effect of this unbalance cannot be entirely corrected. Interestingly, no improvement was observed for cases of highly asymmetric rates of admixture (i.e. $p_{adx} = 0.1$) in spite of this being the type of situations for which the use of this alternative prior is suggested (Pritchard et al., 2010).

4.5 | Recommendations for empirical studies

Our simulation study differs in its complexity from situations found in nature. For example, we considered a single, isolated event of admixture before allowing drift to act on the three simulated populations instead of having continuous admixture through time. These settings, however, allowed us to observe more clearly the effects of other parameters found in empirical studies (e.g. low levels of differentiation between populations; Sarno et al., 2017).

This being said, our simulations clearly suggest that STRUCTURE's estimates of individual admixture proportions must be interpreted cautiously, especially when the number of markers is moderate. Regarding sampling schemes for simple demographic scenarios of admixture similar to ours, we suggest avoiding strongly unequal sample sizes of parental populations and avoiding oversampling populations believed to be admixed. However, it is obviously difficult to ensure balanced sampling of populations belonging to different genetic groups as the boundaries of such groups are, many times, not known a priori (e.g. Baker, Daugherty, Colbourne, & McLennan, 1995; Carriconde et al., 2008; Rueness et al., 2003). This means that if perfectly balanced sampling cannot be guaranteed, the possibility of some bias must always be kept in mind. This can be especially relevant for detection of admixture, since thresholds based on admixture proportions are often used as boundaries between pure and admixed individuals (Marie et al., 2011; Randi, 2008), or estimation of variable admixture rates as in most studies of hybrid zones. In general, we recommend keeping balanced sample sizes of pure parental populations and relatively smaller sample sizes of admixed populations, although this proportion should not be too low (see Vähä & Primmer, 2006) and test the robustness of STRUCTURE analyses towards unbalanced sampling schemes by running multiple analyses with different subsampling schemes.

If possible, the evolutionary history of the populations should also be considered. First, populations that experienced long drift periods after admixture and/or coming from weakly differentiated parental populations are expected to show stronger biases when analysed with STRUCTURE. We thus recommend using STRUCTURE cautiously when divergence between parental populations is low and/or admixture is not recent. Note, however, that this is not specific to STRUCTURE, and in these empirical situations, correctly identifying the pattern of admixture and estimating individual admixture proportions will always be more challenging than for recent admixture events between well-differentiated populations. Second, populations coming from asymmetric admixture events are also likely to show strong biases in their admixture proportion estimates. Although the use of

the alternative prior of individual ancestry is recommended to deal with cases of asymmetric admixture (Pritchard et al., 2010), no real improvements were observed in our results on this type of situations (Figure S6). However, we recommend its use over the default one, as it was shown to be useful for the most biased cases of unbalanced sampling. Notice, however, that its power seems to be limited to cases with low drift, as it could not alleviate the biases observed with long drift times. We do not recommend the use of the alternative prior together with the uncorrelated allele model and a lower value of ALPHA suggested by Wang (2017) as it did not improve the results obtained by using the alternative prior alone and it showed worse results in some cases. However, we acknowledge that this may be due to the fact that under our divergence model, allelic frequencies are clearly correlated, and that this divergence with admixture model fits relatively well with the *F*-model of correlated allele frequencies implemented in STRUCTURE (Falush, Stephens, & Pritchard, 2003).

Finally, using more markers usually improves the estimation of individual admixture proportions (and we suggest doing so as far as possible) although we noticed that in some of our simulations increasing the number of markers did not entirely solve the problems produced by strong drift and unbalanced sampling. In conclusion, we suggest not to use STRUCTURE alone to infer admixture, but to compare/contrast its results with those from other approaches (e.g. PCA). When parental populations are known, the hybrid index inferred by genomic cline approaches (Gompert & Buerkle, 2011, 2012) could also be used. When they can be applied, population-level tests of admixture (such as proposed by Patterson et al., 2012) provide another way of assessing the reliability of individual-level estimates. Last, Lawson, van Dorp, and Falush (2018) developed a novel STRUCTURE-like clustering method that works well with genomic haplotypic data (i.e. using linkage disequilibrium information), but is less precise with unlinked markers and remains to be tested on small numbers of unlinked markers, which constitute the most frequently used type of data in many empirical studies.

ACKNOWLEDGEMENTS

We thank Mathieu Gautier for initial discussions, sharing code and reading the final version of the manuscript, as well as two anonymous reviewers for constructive comments. Part of this work was carried out by using the resources of the national INRA MIGALE (http://migale.jouy.inra.fr) and GENOTOUL (Toulouse Midi-Pyrénées) 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. All authors were supported by the Agence Nationale de la Recherche (RL: project GENOSPACE ANR-16-CE02-0008; PAC & RL: project INTROSPEC ANR-19-CE02-0011; RL & KST: project "Investissement d'Avenir" IBC Institut de Biologie Computationnelle)

AUTHOR CONTRIBUTIONS

The study was designed by P.-A.C. and R.L. The research was performed by K.S.T. and R.L. The data were analysed and the manuscript was written by all authors.

DATA AVAILABILITY STATEMENT

The data generated in this study, as well as the scripts and files necessary to run the analyses, are available via Data Dryad: https://doi.org/10.5061/dryad.gf1vhhmkw (Toyama et al., 2020).

ORCID


Ken S. Toyama https://orcid.org/0000-0002-8331-4894

REFERENCES

Aldrich, M. C., Selvin, S., Hansen, H. M., Barcellos, L. F., Wrensch, M. R., Sison, J. D., … Wiencke, J. K. (2008). Comparison of statistical methods for estimating genetic admixture in a lung cancer study of African Americans and Latinos. *American Journal of Epidemiology*, *168*(9), 1035–1046. https://doi.org/10.1093/aje/kwn224

Arntzen, J. W., Trujillo, T., Butôt, R., Vrieling, K., Schaap, O., Gutiérrez-Rodríguez, J., & Martínez-Solano, I. (2016). Concordant morphological and molecular clines in a contact zone of the Common and Spined toad (*Bufo bufo* and *B. spinosus*) in the northwest of France. *Frontiers in Zoology*, *13*(1), 52. https://doi.org/10.1186/s12983-016-0184-7

Baker, A. J., Daugherty, C. H., Colbourne, R., & McLennan, J. L. (1995). Flightless brown kiwis of New Zealand possess extremely subdivided population structure and cryptic species like small mammals. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(18), 8254–8258. https://doi.org/10.1073/pnas.92.18.8254

Barton, N. H., & Hewitt, G. M. (1989). Adaptation, speciation and hybrid zones. *Nature*, *341*, 497–503. https://doi.org/10.1038/341497a0

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

Brumfield, R. T., Beerli, P., Nickerson, D. A., & Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, *18*(5), 249–256. https://doi.org/10.1016/S0169-5347(03)00018-1

Burgarella, C., Lorenzo, Z., Jabbour-Zahab, R., Lumaret, R., Guichoux, E., Petit, R. J., … Gil, L. (2009). Detection of hybrids in nature: Application to oaks (*Quercus suber* and *Q. ilex*). *Heredity*, *102*(5), 442–452. https://doi.org/10.1038/hdy.2009.8

Carriconde, F., Gardes, M., Jargeat, P., Heilmann-Clausen, J., Mouhamadou, B., & Gryta, H. (2008). Population evidence of cryptic species and geographical structure in the cosmopolitan ectomycorrhizal fungus, *Tricholoma scalpturatum*. *Microbial Ecology*, *56*(3), 513–524. https://doi.org/10.1007/s00248-008-9370-2

Cornille, A., Feurtey, A., Gélin, U., Ropars, J., Misvanderbrugge, K., Gladieux, P., & Giraud, T. (2015). Anthropogenic and natural drivers of gene flow in a temperate wild fruit tree: A basis for conservation and breeding programs in apples. *Evolutionary Applications*, *8*(4), 373–384. https://doi.org/10.1111/eva.12250

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sunderland, MA: Sinauer Associates.

Cullingham, C. I., Miller, J. M., Peery, R. M., Dupuis, J. R., Malenfant, R. M., Gorrell, J. C., & Janes, J. K. (2020). Confidently identifying the correct K value using the ΔK method: When does K= 2? *Molecular Ecology*, *29*(5), 862–869. https://doi.org/10.1111/mec.15374

Dannemann, M., Andrés, A. M., & Kelso, J. (2016). Introgression of neandertal-and denisovan-like haplotypes contributes to adaptive variation in human Toll-like receptors. *American Journal of Human Genetics*, *98*(1), 22–33. https://doi.org/10.1016/j.ajhg.2015.11.015

De Carvalho, D., Ingvarsson, P. K., Joseph, J., Suter, L., Sedivy, C., Macaya-Sanz, D., … Lexer, C. (2010). Admixture facilitates adaptation from standing variation in the European aspen (*Populus tremula* L.), a

widespread forest tree. *Molecular Ecology*, *19*(8), 1638–1650. https://doi.org/10.1111/j.1365-294X.2010.04595.x

Dufresnes, C., Mazepa, G., Rodrigues, N., Brelsford, A., Litvinchuk, S. N., Sermier, R., … Jeffries, D. L. (2018). Genomic evidence for cryptic speciation in tree frogs from the Apennine Peninsula, with description of *Hyla perrini* sp. nov. *Frontiers in Ecology and Evolution*, *6*, 144. https://doi.org/10.3389/fevo.2018.00144

Ellstrand, N. C., Prentice, H. C., & Hancock, J. F. (1999). Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics*, *30*(1), 539–563. https://doi.org/10.1146/annurev.ecolsys.30.1.539

Elmer, K. R., Recknagel, H., Thompson, A., & Meyer, A. (2012). Asymmetric admixture and morphological variability at a suture zone: Parapatric burbot subspecies (*Pisces*) in the Mackenzie River basin. *Canada*. *Hydrobiologia*, *683*(1), 217–229. https://doi.org/10.1007/s10750-011-0959-y

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

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.

Gagnaire, P. A., Minegishi, Y., Aoyama, J., Reveillac, E., Robinet, T., Bosc, P., … Berrebi, P. (2009). Ocean currents drive secondary contact between *Anguilla marmorata* populations in the Indian Ocean. *Marine Ecology Progress Series*, *379*, 267–278. https://doi.org/10.3354/meps07895

Gompert, Z., & Buerkle, C. A. (2011). Bayesian estimation of genomic clines. *Molecular Ecology*, *20*(10), 2111–2127. https://doi.org/10.1111/j.1365-294X.2011.05074.x

Gompert, Z., & Buerkle, C. A. (2012). bgc: Software for Bayesian estimation of genomic clines. *Molecular Ecology Resources*, *12*(6), 1168–1176. https://doi.org/10.1111/1755-0998.12009.x

Good, J. M., Hird, S., Reid, N., Demboski, J. R., Steppan, S. J., Martin-Nims, T. R., & Sullivan, J. (2008). Ancient hybridization and mitochondrial capture between two species of chipmunks. *Molecular Ecology*, *17*(5), 1313–1327. https://doi.org/10.1111/j.1365-294X.2007.03640.x

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

Hedrick, P. W. (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, *22*(18), 4606–4618. https://doi.org/10.1111/mec.12415

Hermansen, J. S., Haas, F., Trier, C. N., Bailey, R. I., Nederbragt, A. J., Marzal, A., & Sætre, G. P. (2014). Hybrid speciation through sorting of parental incompatibilities in Italian sparrows. *Molecular Ecology*, *23*(23), 5831–5842. https://doi.org/10.1111/mec.12910

Hewitt, G. M. (2001). Speciation, hybrid zones and phylogeography—or seeing genes in space and time. *Molecular Ecology*, *10*(3), 537–549. https://doi.org/10.1046/j.1365-294x.2001.01202.x

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Janes, J. K., Miller, J. M., Dupuis, J. R., Malenfant, R. M., Gorrell, J. C., Cullingham, C. I., & Andrew, R. L. (2017). The K= 2 conundrum. *Molecular Ecology*, *26*(14), 3594–3602. https://doi.org/10.1111/mec.14187

Kalinowski, S. T. (2011). The computer program STRUCTURE does not reliably identify the main genetic clusters within species: Simulations and implications for human population structure. *Heredity*, *106*(4), 625–632. https://doi.org/10.1038/hdy.2010.95

Kolbe, J. J., Glor, R. E., Schettino, L. R., Lara, A. C., Larson, A., & Losos, J. B. (2007). Multiple sources, admixture, and genetic variation in introduced *Anolis* lizard populations. *Conservation Biology*, *21*(6), 1612–1625. https://doi.org/10.1111/j.1523-1739.2007.00826.x

Larcombe, M. J., Vaillancourt, R. E., Jones, R. C., & Potts, B. M. (2014). Assessing a Bayesian approach for detecting exotic hybrids between plantation and native Eucalypts. *International Journal of Forestry Research*, *2014*, 1–13. https://doi.org/10.1155/2014/650202

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

Mallet, J. (2007). Hybrid speciation. *Nature*, *446*(7133), 279–283. https://doi.org/10.1038/nature05706

Marie, A. D., Bernatchez, L., & Garant, D. (2011). Empirical assessment of software efficiency and accuracy to detect introgression under variable stocking scenarios in brook charr (*Salvelinus fontinalis*). *Conservation Genetics*, *12*(5), 1215–1227. https://doi.org/10.1007/s10592-011-0224-y

Meirmans, P. G. (2019). Subsampling reveals that unbalanced sampling affects STRUCTURE results in a multi-species dataset. *Heredity*, *122*(3), 276–287. https://doi.org/10.1038/s41437-018-0124-8

Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., … Lindqvist, C. (2012). Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(36), E2382–E2390. https://doi.org/10.1073/pnas.1210506109

Miraldo, A., Faria, C., Hewitt, G. M., Paulo, O. S., & Emerson, B. C. (2013). Genetic analysis of a contact zone between two lineages of the ocellated lizard (*Lacerta lepida* Daudin 1802) in south-eastern Iberia reveal a steep and narrow hybrid zone. *Journal of Zoological Systematics and Evolutionary Research*, *51*(1), 45–54. https://doi.org/10.1111/jzs.12005

Morin, P. A., Archer, F. I., Pease, V. L., Hancock-Hanser, B. L., Robertson, K. M., Huebinger, R. M., … Taylor, B. L. (2012). Empirical comparison of single nucleotide polymorphisms and microsatellites for population and demographic analyses of bowhead whales. *Endangered Species Research*, *19*(2), 129–147. https://doi.org/10.3354/esr00459

Morin, P. A., Luikart, G., Wayne, R. K., & the SNP Workshop Group. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, *19*(4), 208–216. https://doi.org/10.1016/j.tree.2004.01.009

Muranishi, S., Tamaki, I., Setsuko, S., & Tomaru, N. (2013). Asymmetric introgression between *Magnolia stellata* and *M. salicifolia* at a site where the two species grow sympatrically. *Tree Genetics & Genomes*, *9*(4), 1005–1015. https://doi.org/10.1007/s11295-013-0612-1

Narum, S. R., Banks, M., Beacham, T. D., Bellinger, M. R., Campbell, M. R., Dekoning, J., … Garza, J. C. (2008). Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology*, *17*(15), 3464–3477. https://doi.org/10.1111/j.1365-294X.2008.03851.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 & Genomes*, *10*(2), 273–285. https://doi.org/10.1007/s11295-013-0680-2

Nielsen, E. E., Hansen, M. M., Ruzzante, D. E., Meldrup, D., & Grønkjær, P. (2003). Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. *Molecular Ecology*, *12*(6), 1497–1508. https://doi.org/10.1046/j.1365-294X.2003.01819.x

Padhukasahasram, B. (2014). Inferring ancestry from population genomic data and its applications. *Frontiers in Genetics*, *5*(204), 1–5. https://doi.org/10.3389/fgene.2014.00204

Papa, R., & Gepts, P. (2003). Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. *Theoretical and Applied Genetics*, *106*(2), 239–250. https://doi.org/10.1007/s00122-002-1085-z

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., … Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., & Lareu, M. V. (2013). An overview of STRUCTURE: Applications, parameter settings, and supporting software. *Frontiers in Genetics*, *4*(98), 1–13. https://doi.org/10.3389/fgene.2013.00098

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

Pritchard, J. K., Wen, X., & Falush, D. (2010). *Documentation for structure software: Version 2.3*. Chicago, IL: University of Chicago.

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

Randi, E. (2008). Detecting hybridization between wild species and their domesticated relatives. *Molecular Ecology*, *17*(1), 285–293. https://doi.org/10.1111/j.1365-294X.2007.03417.x

Rueness, E. K., Jorde, P. E., Hellborg, L., Stenseth, N. C., Ellegren, H., & Jakobsen, K. S. (2003). Cryptic population structure in a large, mobile mammalian predator: The Scandinavian lynx. *Molecular Ecology*, *12*(10), 2623–2633. https://doi.org/10.1046/j.1365-294X.2003.01952.x

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

Sarno, S., Boattini, A., Pagani, L., Sazzini, M., De Fanti, S., Quagliariello, A., … Pettener, D. (2017). Ancient and recent admixture layers in Sicily and Southern Italy trace multiple migration routes along the Mediterranean. *Scientific Reports*, *7*(1), 1–12. https://doi.org/10.1038/s41598-017-01802-4

Schaefer, J., Duvernell, D., & Campbell, D. C. (2016). Hybridization and introgression in two ecologically dissimilar *Fundulus* hybrid zones. *Evolution*, *70*(5), 1051–1063. https://doi.org/10.1111/evo.12920

Simonti, C. N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D. S., Chisholm, R. L., … Capra, J. A. (2016). The phenotypic legacy of admixture between modern humans and Neandertals. *Science*, *351*(6274), 737–741. https://doi.org/10.1126/science.aad2149

Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *28*(4), 289–301. https://doi.org/10.1002/gepi.20064

Toyama, K. S., Crochet, P. A., & Leblois, R. (2020). Data and analysis scripts from: Sampling schemes and drift can bias admixture proportions inferred by STRUCTURE, v2. Dryad Dataset, https://doi.org/10.5061/dryad.gf1vhhmkw

Trigo, T. C., Freitas, T. R. O., Kunzler, G., Cardoso, L., Silva, J. C. R., Johnson, W. E., … Eizirik, E. (2008). Inter-species hybridization among Neotropical cats of the genus *Leopardus*, and evidence for an introgressive hybrid zone between *L. geoffroyi* and *L. tigrinus* in southern Brazil. *Molecular Ecology*, *17*(19), 4317–4333. https://doi.org/10.1111/j.1365-294X.2008.03919.x

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

Verity, R., & Nichols, R. A. (2016). Estimating the number of subpopulations (K) in structured populations. *Genetics*, *203*(4), 1827–1839. https://doi.org/10.1534/genetics.115.180992

Vilà, M., Weber, E., & Antonio, C. M. (2000). Conservation implications of invasion by plant hybridization. *Biological Invasions*, *2*(3), 207–217. https://doi.org/10.1023/A:1010003603310

Vilaça, S. T., Vargas, S. M., Lara-ruiz, P., Molfetti, É., Reis, E. C., Lôbo-hajdu, G., … Santos, F. R. (2012). Nuclear markers reveal a complex introgression pattern among marine turtle species on the Brazilian coast. *Molecular Ecology*, *21*(17), 4300–4312. https://doi.org/10.1111/j.1365-294X.2012.05685.x

von Holdt, B. M., Brzeski, K. E., Wilcove, D. S., & Rutledge, L. Y. (2018). Redefining the role of admixture and genomics in species conservation. *Conservation Letters*, *11*(2), e12371. https://doi.org/10.1111/conl.12371

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

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*, 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Wielstra, B., Burke, T., Butlin, R. K., Avcı, A., Üzüm, N., Bozkurt, E., … Arntzen, J. W. (2017). A genomic footprint of hybrid zone movement in crested newts. *Evolution Letters*, *1*(2), 93–101. https://doi.org/10.1002/evl3.9

Wolf, D. E., Takebayashi, N., & Rieseberg, L. H. (2001). Predicting the risk of extinction through hybridization. *Conservation Biology*, *15*(4), 1039–1053. https://doi.org/10.1046/j.1523-1739.2001.01500041039.x

Zohren, J., Wang, N., Kardailsky, I., Borrell, J. S., Joecker, A., Nichols, R. A., & Buggs, R. J. (2016). Unidirectional diploid–tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. *Molecular Ecology*, *25*(11), 2413–2426. https://doi.org/10.1111/mec.13644

SUPPORTING INFORMATION

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

How to cite this article: Toyama KS, Crochet P-A, Leblois R. Sampling schemes and drift can bias admixture proportions inferred by STRUCTURE. *Mol Ecol Resour*. 2020;00:1–17. https://doi.org/10.1111/1755-0998.13234