A simulation-based evaluation of STRUCTURE software for exploring the introduction routes of invasive species

Eric Lombaert[1,2,3], Thomas Guillemaud[1,2,3] & Emeline Deleury[1,2,3]

[1] INRA, UMR 1355 ISA, 06903 Sophia-Antipolis, France

[2] Université de Nice Sophia Antipolis, UMR ISA, 06903 Sophia-Antipolis, France

[3] CNRS, UMR 7254 ISA, 06903 Sophia-Antipolis, France

Keywords: biological invasion, microsatellite, invasion routes, STRUCTURE software, simulation, bottleneck.

Corresponding author:

Eric Lombaert

INRA, UMR 1301 IBSV (INRA / Université de Nice Sophia Antipolis / CNRS). 400 Route des Chappes. BP 167 - 06903 Sophia Antipolis cedex. FRANCE

E-mail: lombaert@sophia.inra.fr

Tel: +33 4 92 38 65 06

Fax: +33 4 92 38 64 01

Running title: Clustering analysis of introduction routes

Abstract

Population genetic methods are widely used to retrace the introduction routes of invasive species. The unsupervised Bayesian clustering algorithm implemented in STRUCTURE is amongst the most frequently use of these methods, but its ability to provide reliable information about introduction routes has never before been assessed. We used computer simulations of microsatellite datasets to evaluate the extent to which the clustering results provided by STRUCTURE were misleading for the inference of introduction routes for two invasive populations independently introduced from a native population. For this simple scenario of independent introductions, the assumption of two clusters (i.e. K=2) would not be expected to result in STRUCTURE assigning the native population to one cluster, with both invasive populations assigned to the second cluster. We investigated the influence of two demographic parameters (effective size and bottleneck severity) and different numbers of loci on the type and frequency of misleading results. We showed that misleading STRUCTURE results were obtained for 10% of our simulated datasets and at a frequency of up to 37% for some combinations of parameters. Our results highlighted two different categories of misleading output. The first occurs in situations in which the native population has a low level of diversity. In this case, the two introduced populations may be very similar, despite their independent introduction histories. The second category results from convergence issues in STRUCTURE for K=2, with strong bottleneck severity and/or large numbers of loci resulting in high levels of differentiation between the three populations.

Introduction

One of the most fundamental questions in invasion biology is understanding what determines the success of invasive alien species. Purely ecological factors may be the most important, but it has been suggested that rapid evolution may also have played a major role in many successful invasions (Lawson Handley *et al.* 2011). Retracing the introduction routes of invasive alien species is a prerequisite to understanding the determinants of their success. However, identification of the source of an introduced population is a complex task, because of the highly stochastic nature of the introduction process (Estoup & Guillemaud 2010). Many classic population genetics methods and tools are now widely used to retrace the introduction routes of invasive species. This approach is somewhat risky, because the methods involved are often dependent on demographic and genetic equilibria, but invasions often involve demographic disequilibrium, through strong bottlenecks followed by rapid population growth, for example. Despite this limitation and the risks of using population genetics methods inappropriately in the specific context of biological invasions, only a few of these methods have been formally evaluated (e.g. Estoup & Guillemaud 2010; Guillemaud *et al.* 2010).

Unsupervised individual Bayesian clustering methods are widely used in population genetics. The popularity of these methods is due to their ability to infer genetic structure correctly in many situations and their apparent simplicity (several “click-and-play” software suites are available). STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003; Hubisz *et al.* 2009) is the most frequently used software for clustering in this context, with more than 25,000 citations for the three references indicated above in Google Scholar in November 2016. STRUCTURE aims to sort individuals in an unsupervised way into *K* clusters (*K* being defined by the user), assuming Hardy-Weinberg/linkage equilibrium within clusters (Porras-Hurtado *et al.* 2013). This method is known to perform well in most cases, but it can be

misleading in some situations, particularly in the presence of isolation by distance (Frantz *et al.* 2009; Schwartz & McKelvey 2009), clusters of very different sizes (Kalinowski 2011; Puechmaille 2016), the sampling of family groups (Anderson & Dunham 2008), or a high proportion of missing data (Smith & Wang 2014).

STRUCTURE and other software suites based on similar methods are frequently used in the context of introduction routes inferences (Estoup & Guillemaud 2010; Lawson Handley *et al.* 2011; Cristescu 2015). STRUCTURE can be used to decrease the number of genetic units from a large number of population samples to a few main clusters before quantitative analyses, such as approximate Bayesian computation (Zepeda-Paulo *et al.* 2010; Ascunce *et al.* 2011; Boissin *et al.* 2012; Rius *et al.* 2012; Konecny *et al.* 2013; Lombaert *et al.* 2014; Gaubert *et al.* 2015; Guillemaud *et al.* 2015). In some cases, STRUCTURE is used directly to contrast models of invasion history, particularly for comparisons of scenarios involving single and multiple independent introductions from a native population. This approach is used to determine the source of invasive populations. Indeed, if samples from an invasive population group with the samples of only one of the potential source populations, this is considered to provide fairly conclusive evidence about the origin of this invasive population (Lachmuth *et al.* 2010; Robert *et al.* 2012; Bolte *et al.* 2013; Fontaine *et al.* 2013; Sanz *et al.* 2013; Zhang *et al.* 2014; Yu *et al.* 2014; Zhou *et al.* 2015; Karsten *et al.* 2015; Mallez *et al.* 2015; Rewicz *et al.* 2015; Dieni *et al.* 2016). However, invasions frequently involve major demographic events, such as strong bottlenecks followed by genetic drift, which may significantly impair our ability to determine introduction routes correctly from a given STRUCTURE result. This problem may account for the contradictory outcomes sometimes obtained with different population genetics methods (e.g. Mallez *et al.* 2015).

In this study, we evaluated the risk of incorrect introduction route inferences based on STRUCTURE analyses, for the simple case of an invasion scenario involving one native

population and two independently introduced populations. We simulated a large number of microsatellite datasets drawn from populations of various effective sizes and bottleneck severities. STRUCTURE analyses were performed on these simulated datasets and the resulting clustering patterns were classified as “misleading” or “non-misleading”. We then explored the effect of demographic parameters on the likelihood of misleading patterns being obtained, to identify and predict the situations in which the use of STRUCTURE in a context of introduction routes inference may be risky.

Methods

Scenario description and data simulation

We defined a simple historical scenario in which two invasive populations (populations 2 and 3) were independently founded 50 generations ago from the same native population (population 1). Both invasive populations were subject to a demographic bottleneck lasting 20 generations (Fig. 1a). The effective sizes of all three populations at equilibrium (N) and the effective number of founders of the two invasive populations during the bottlenecks (NF) could take different values: 10000, 1000, 100, 10 and 2 individuals, with $N \geq NF$. $\text{Log}_{10}(N/NF)$ was considered to quantify bottleneck severity.

We used DIYABC version 2.0.4 software (Cornuet *et al.* 2014) to generate 500 microsatellite multilocus genotype datasets for each of the 15 different combinations of N and NF values, through a coalescent process. For all datasets, a sample of 30 diploid individuals per population was simulated. We evaluated the effect of the number of loci on the analyses, by performing simulations with 10, 20 and 100 unlinked microsatellite markers. We used a generalized stepwise mutation model, with realistic values for all three parameters (Jarne &

Lagoda 1996; Estoup *et al.* 2002): the mean mutation rate (set to $5x10^{-4}$), the mean parameter of the geometric distribution defining the number of microsatellite repeats gained or lost during mutation events (set to 0.22) and the mean mutation rate for single-nucleotide insertion/deletion (set to 10^{-8}). In total, we simulated 22,500 datasets (15 sets of parameters x 500 datasets per set x 3 numbers of loci). We developed a pipeline with PERL scripts, available on request, to automate the processing of the datasets (simulation and subsequent STRUCTURE and post-STRUCTURE analyses).

STRUCTURE analyses and misleading clustering

For each of the 22,500 simulated datasets, a Bayesian clustering analysis was performed in parallel, on a 120-node computer cluster, with STRUCTURE software version 2.3.4 (Pritchard *et al.* 2000). We chose the admixture model with correlated allele frequencies. We used default values for all the other parameters. Each run consisted of a burn-in period of 10^5 Markov chain Monte Carlo (MCMC) iterations, followed by $5x10^5$ MCMC iterations. This run length is considered to be long enough to obtain precise estimates of parameters (Pritchard *et al.* 2010), but we also tried runs of double this length for some combinations of parameters with 100 loci. The results obtained were the same (data not shown). We carried out ten replicate runs for each dataset and each value of *K*, the number of genetic clusters, with *K* taking values of 1, 2, 3 and 4.

We investigated the ability of STRUCTURE to clarify introduction routes, by focusing on analyses with *K*=2. With this *K* value, the two samples from an introduced population may or may not cluster together. With the scenario simulated here, in which the two invasive populations result from two independent introductions, the two samples of the introduced populations would not be expected to cluster together. Indeed, the two independent drift

pulses at work during these two introductions (i.e. the bottleneck events) should make the introduced populations more genetically different from each other than from the native population, from which they are separated by a single drift pulse. Consequently, STRUCTURE would yield a misleading pattern if the native population sample belonged to one cluster and the two invasive population samples both belonged to the other. Indeed, this could be considered evidence for a lack of independence of the two populations, with one invasive population being the source of the other (Fig. 1b). Such a clustering pattern, hereafter referred to as "misleading clustering", would lead most STRUCTURE users to an incorrect interpretation, according to which a "successive introductions" scenario would be more likely than the "independent introductions" scenario. Note that STRUCTURE analyses carried out on three population samples with K=3 are, theoretically, unsuitable for comparisons of independent and successive introduction scenarios, because each population sample would probably form its own cluster.

For analysis of the 225,000 STRUCTURE runs with K=2 and estimation of the frequency of misleading clusterings, the STRUCTURE output was characterized as follows. From the output file of each run, we extracted the proportion of membership Qi_A and Qi_B of population sample i for clusters A and B, respectively (with $Qi_B = 1 - Qi_A$). The Qi_A and Qi_B values were coded as *0*, *25*, *50*, *75* or *100* when belonging to the [0;0.2],]0.2;0.4[, [0.4;0.6],]0.6;0.8[or [0.8;1] intervals, respectively. For each STRUCTURE run, we summarized the clustering pattern by a code $C_{1A}/C_{2A}/C_{3A}$, where C_{iA} is the membership code of population sample i for cluster A. For example, the clustering code would be *0/0/100* for a STRUCTURE run output in which $Q1_A$=0.12, $Q2_A$=0.05 and $Q3_A$=0.96. Note that belonging to cluster A or B has no specific meaning, and the subscripts A and B can thus be permuted. For example, clustering codes *0/0/100* and *100/100/0* summarize the same pattern and are pooled together as *0/0/100*. Given the simulated scenario of independent introductions of the two invasive

populations, *0/100/100* was the code considered to correspond to misleading clustering (Fig. 1b). All other clustering codes were considered non-misleading in the context of introduction routes inference. Focusing on the codes instead of the proportions of membership made it possible to pool together slightly different clustering patterns in the same category.

Given the stochastic processes involved in the MCMC analysis, the ten replicated STRUCTURE runs performed on a single dataset could conceivably generate different results, a phenomenon called genuine multimodality (Jakobsson & Rosenberg 2007; Porras-Hurtado *et al.* 2013). Clustering results for a given dataset were considered to be homogeneous if the same clustering code (as defined above) was obtained in all ten runs. They were otherwise considered to be heterogeneous. We evaluated the global occurrence of misleading clustering in the analyses of the simulated datasets, and focused on two critical categories of misleading clusterings (Fig. 1c):

(i) "Misleading homogeneous clusterings": for one dataset, all ten runs homogeneously provide the misleading clustering pattern *0/100/100*.

(ii) "Misleading heterogeneous clusterings": for one dataset, the ten runs are not homogeneous (i.e. genuine multimodality is observed) and the misleading clustering pattern *0/100/100* predominates.

For each dataset, we also inferred the best value of K, as follows: if the mean natural logarithm of the likelihood of the data $ln(P(X|K))$ with K in [1, 2, 3, 4] is maximal for K=1, then the inferred number of clusters is 1; otherwise, we determined the best value of K by the ΔK method (Evanno *et al.* 2005).

Effect of demographic parameter values on misleading clustering

For each number of simulated microsatellite loci (10, 20 or 100), the variables "the proportion of analyses yielding misleading homogeneous clusterings" and "the proportion of analyses yielding misleading heterogeneous clusterings" were analyzed independently with a generalized linear model, using a binomial probability distribution of the residual error and a logit link function. The following factors were included as fixed effects: effective population size N and bottleneck severity $\log_{10}(N/NF)$. We used the Akaike information criterion (AIC) to select the best model from the various models of different complexity. Analyses were performed with R software V3.2.2 (R Development Core Team 2015).

Link between summary statistics of genetic diversity and STRUCTURE patterns

We summarized each simulated dataset, by using ARLSUMSTAT version 3.5 software (Excoffier & Lischer 2010) to compute the mean number of alleles and the mean expected heterozygosity in each population sample, and the pairwise F_{ST} values between each pair of populations. We also used in-house PERL scripts to compute (i) the mean individual assignment likelihood (Rannala & Mountain 1997) ($L_{i\rightarrow j}$) of each invading population (samples 2 and 3) to each possible source population (i.e. either the native population or the other invasive population), and (ii) the number of alleles shared by the invasive population samples.

For the comparison of datasets leading to "misleading homogeneous clusterings", "misleading heterogeneous clusterings" and "non-misleading clusterings", we specifically explored a few genetic diversity summary statistics: (i) expected heterozygosity of the native population sample, (ii) mean expected heterozygosity of both invasive population samples and (iii) the ratio of alleles shared by the two invasive population samples to the total number

of alleles in the two samples. For each summary statistic and each number of loci, pairwise Mann-Whitney tests with Holmes correction for multiple comparisons were performed.

We also compared STRUCTURE results with those obtained by two other methods traditionally used to identify source populations: (i) the "F_{ST}-based method" and the (ii) the "assignment likelihood-based method" (Genton *et al.* 2005; Pascual *et al.* 2007; Ciosi *et al.* 2008; Tepolt *et al.* 2009; Thibault *et al.* 2009; Papura *et al.* 2012; Mallez *et al.* 2015; Dieni *et al.* 2016). For an "independent introductions" scenario, we would expect the F_{ST} between the two invasive population samples to be larger than the F_{ST} values between the native population and each of the invasive population samples (i.e. F_{ST} 2-3 > F_{ST} 1-2 and F_{ST} 2-3 > F_{ST} 1-3). We would also expect both invasive population samples to be best assigned to the native population sample (i.e. $L_{2\rightarrow1} > L_{2\rightarrow3}$ and $L_{3\rightarrow1} > L_{3\rightarrow2}$).

For each dataset, a global exact test for population genotypic differentiation (Raymond & Rousset, 1995a) was carried out with GENEPOP software version 4.3 (Raymond & Rousset, 1995b). If a dataset displayed no population differentiation, we made the prudent and standard decision of not trying to infer any evolutionary relationship between the population samples. Consequently, such datasets were considered to generate non-misleading results for all methods.

Results

Effect of demographic parameter values on simulated datasets

The 500 simulated datasets for each parameter set are summarized with some common statistics in Table S1. Decreasing effective population sizes (*N)* generate lower intra-population and higher inter-population genetic diversities. Increasing bottleneck severity

($\log_{10}(N/NF)$) generates lower intra-population genetic diversities for both invasive samples, and overall higher inter-population genetic diversity. The main impact of a larger number of loci is a decrease in the variance of all summary statistics. Overall, the chosen parameter values (for *N* and *NF*) yield a large number of different combinations of genetic diversity for evaluation of the ability of STRUCTURE software to explore introduction routes in different situations.

Overall STRUCTURE results

The best value of *K* inferred by the Δ*K* method (Evanno *et al.* 2005) was most frequently three (Fig. 2 and Fig. S1). The proportion of datasets for which the best number of clusters was *K*=3 increased strongly with increasing numbers of loci (41.9%, 50.6% and 74.9% for 10, 20 and 100 loci, respectively). More than 80% of the simulated datasets for which *K*=3 was inferred by the Δ*K* method had heterogeneous clustering codes (i.e. genuine multimodality) at *K*=2 (Fig. 2). By contrast, when the number of inferred clusters was one or two, multimodality at *K*=2 was found in less than 10% of all datasets.

Occurrence of misleading STRUCTURE patterns

Three categories of clustering codes at *K*=2 accounted for more than 95% of all runs (see Table S2 for details): (i) clusterings in which all populations were fully admixed and undistinguishable with STRUCTURE (i.e. the *50/50/50* code), (ii) clusterings in which the two invasive samples belonged to two different clusters (i.e. the C_{1A}*/100/0* and C_{1A}*/0/100* codes) and (iii) the misleading clusterings defined earlier, in which the two invasive samples

belonged to the same cluster, whereas the native sample belong to the other cluster (i.e. the *0/100/100* code, Fig. 1b).

Overall, the proportion of datasets with at least one misleading clustering pattern over the ten STRUCTURE runs ("misleading homogeneous clusterings", "misleading heterogeneous clusterings" and non-misleading clusterings with at least one run yielding a misleading pattern) was 15.31%, 22.07% and 47.01% for 10, 20 and 100 simulated loci, respectively (Fig. 3a and Fig. S2), and very similar proportions were obtained with more (0.1 and 0.9) and less (0.3 and 0.7) stringent Q_{iA} cutoff values (instead of 0.2 and 0.8 for Q_{iA}) for the encoding of pattern results (Table S3).

The frequency of "misleading homogeneous clusterings" was similar for different numbers of loci, and was rather low overall (between 4.24% and 5.59% of the datasets, Fig. 3a). "Misleading heterogeneous clusterings" were also infrequent, but their frequency increased with the number of loci: 2.71%, 3.96% and 8.41% for 10, 20 and 100 loci, respectively (Fig. 3a). Overall, 7.45%, 9.55% and 12.65% of datasets for 10, 20 and 100 loci, respectively, yielded misleading results. For some combinations of parameters, this proportion reached 36.8% of datasets (Fig. S2). K=2 was most often (70%) inferred for datasets yielding "misleading homogeneous clusterings", and K=3 was most often (91%) inferred for datasets leading to "misleading heterogeneous clusterings" (Fig. 3b).

Effect of demographic parameter values on STRUCTURE results

For the response variable "proportion of analyses yielding misleading homogeneous clusterings", the best model according to the AIC always included the effective population size at equilibrium N, which was highly significant whatever the number of simulated loci (Table 1). Lower N values resulted in a higher proportion of misleading homogeneous

clusterings (Fig. S3a). The best model also included bottleneck severity, $\log_{10}(N/NF)$, and the interaction between the two main factors for 10 and 100 loci. Bottleneck severity was significant only for 10 loci, and had a positive effect: the stronger the bottleneck, the higher the proportion of misleading homogeneous clusterings. The interaction between the two factors was significant in both models (Table 1 and Fig. S3a).

For the response variable "proportion of analyses yielding misleading heterogeneous clusterings", the full model was selected for all numbers of simulated loci (Table 1). The effective population size at equilibrium *N* was significant in all cases, and had a negative effect (Fig. S3b). Bottleneck severity $\log_{10}(N/NF)$ was also strongly significant for all numbers of loci, but its effect was positive for 10 and 20 loci and negative for 100 loci. The interaction between the two factors was significant for 20 and 100 loci, with a positive effect (Table 1).

Links between summary statistics for genetic diversity and STRUCTURE patterns

As expected from the analysis of demographic parameters, the diversity of the native population, as assessed by its expected heterozygosity in the datasets with "misleading homogeneous clusterings", was lower than that for "non-misleading" datasets, whatever the number of loci considered (Fig. 4). On the contrary, no clear trend could be observed for datasets with "misleading heterogeneous clusterings". For these datasets, the mean expected heterozygosity was relatively high with 10 loci, intermediate with 20 loci and low with 100 loci, but, in each case, extreme low and high values were observed. The diversity of invasive populations, which was affected by both the diversity of the native population and bottleneck severity, was low for both kinds of misleading clusterings (Fig. 4). In comparisons with the "non-misleading" datasets, the proportion of alleles shared by the two invasive populations

was higher for the datasets with “misleading homogeneous clusterings”, and lower for “misleading heterogeneous clusterings”, unless 100 loci were considered (Fig. 4).

Outcomes for comparisons of STRUCTURE clusterings with results from F_{ST}-based and assignment likelihood-based methods were very mixed, depending on the type of “misleading clusterings” considered. 86.5%, 93.6% and 99.1% of datasets with “misleading homogeneous clusterings” in STRUCTURE provided misleading results with at least one of the methods based on F_{ST} or assignment likelihood, when considering 10, 20 and 100 loci, respectively (Fig. S4). By contrast, datasets with “misleading heterogeneous clusterings” in STRUCTURE analysis were rarely (for 10 and 20 loci), or at least not as strongly (for 100 loci), associated with misleading results with the other methods: this was the case for 15.2%, 24.9% and 67.5% of these datasets for 10, 20 and 100 loci, respectively (Fig. S4).

Discussion

We used simulated microsatellite datasets for a particular invasion scenario to determine whether the method implemented in the widely used STRUCTURE software (Pritchard *et al.* 2000) could mislead users trying to infer introduction routes. For a true scenario of two independent invasions from a single source, we found that STRUCTURE runs could give misleading clustering patterns (i.e. the two invasive populations clustered together at K=2). In about 10% of all simulated datasets, the results led to incorrect interpretation, with all (“homogeneous misleading clusterings”) or most (“heterogeneous misleading clusterings”) of the runs for a given dataset yielding the misleading pattern. Some combinations of demographic parameters resulted in higher frequencies of misleading results with STRUCTURE, and, contrary to expectations, increasing the number of loci also lead to an overall increase in the frequency of misleading results. Our results suggested that the two

types of misleading clustering hazard, homogeneous and heterogeneous misleading clustering, were very different. We suggest that (i) "homogeneous misleading clusterings" probably arise from a large probability of independently drawing the same alleles twice from a native population with low genetic diversity and that (ii) "heterogeneous misleading clusterings" probably arise from convergence problems in STRUCTURE.

For "homogeneous misleading clustering", the effective size of the native population has the strongest effect: the smaller this effective population size, the higher the risk of obtaining misleading clustering patterns over all STRUCTURE runs. Such "homogeneous misleading clustering" occurred principally when the two invasive populations shared a large proportion of alleles, and the F_{ST}-based and likelihood assignment-based methods frequently yielded the same clustering pattern. Accordingly, the number of clusters inferred by the Evanno method was most frequently K=2. Invasive populations result from founder effects and genetic drift (Simberloff 2009; Lawson Handley *et al.* 2011), which are random processes. The probability of independently drawing the same alleles twice, with similar frequencies, from a given native population is usually low (when random processes are at work), but can actually be quite large when the diversity of the native population is itself low. This is particularly true in cases of low heterozygosity (Allendorf 1986), in which one or a few alleles occur at high frequency at each locus.

The interpretation of "heterogeneous misleading clusterings" is less clear-cut, but several lines of evidence suggest the involvement of convergence issues in STRUCTURE runs at K=2, when the best K value was undoubtedly 3. For 10 and 20 simulated loci, the proportion of "heterogeneous misleading clusterings" (i) was positively related to bottleneck severity, which accentuates differences between populations, (ii) was associated with a lower proportion of alleles shared by the two invasive populations than for non-misleading datasets, (iii) was higher when the F_{ST}- and assignment likelihood-based methods provided expected

results, and (iv) was higher in datasets for which *K*=3 was inferred with Evanno's method (Evanno *et al.* 2005). With 100 loci, we observed a relatively large proportion of "heterogeneous misleading clusterings" when the bottleneck was benign or not very serious, probably because it was easier for STRUCTURE to cluster individuals correctly into the three actual populations when there was a large number of loci (Evanno *et al.* 2005; Waples & Gaggiotti 2006; Hubisz *et al.* 2009). Indeed, the ability of Evanno's method to infer *K*=3 and the occurrence of multimodality at *K*=2 increases steadily with the number of loci. Overall, these results are consistent with the notion that "heterogeneous misleading clusterings" often occur when the three populations are clearly differentiated and should yield an unambiguous *K*=3 result in STRUCTURE analyses. We therefore suggest that "heterogeneous misleading clustering" probably results from a convergence problem in the MCMC procedure of STRUCTURE: when an inappropriate number of clusters is imposed — here *K*=2 whereas the data are more consistent with *K*=3 — multimodalities are often observed (Pritchard *et al.* 2000; Jakobsson & Rosenberg 2007), and sometimes, by chance, a large proportion of misleading clustering events occur in the various runs, resulting in "heterogeneous misleading clustering".

Conclusion and general recommendations

This study was based on a single simple invasion scenario with only three populations. More complex scenarios should be studied in the future, but this study constitutes a crucial first step, providing important information about the use of clustering methods in the context of biological invasions.

We found that STRUCTURE yielded misleading results, but at a low frequency. However, our results suggest that some situations should be analyzed with care.

First, invasion biologists should be very cautious if the diversity of the native population is low: independent introductions from a single source population with low genetic diversity are likely to produce genetic signals similar to that expected for successive introductions. Such misleading results are difficult to spot, as they are consistent with the results of other methods, such as F_{ST}- or assignment likelihood-based methods. In this context, quantitative methods, such as approximate Bayesian computation (ABC), may be very useful. Second, multimodal STRUCTURE results should be interpreted very cautiously, particularly if large numbers of loci are used. This is sobering news, because many published studies interpret STRUCTURE results at different *K* values, including those displaying genuine multimodality. This problem is not specific to the exploration of introduction routes and has much wider implications (Meirmans 2015). Multimodality is often a sign of poor convergence of STRUCTURE runs, and is therefore likely to lead to results of limited biological meaning. In such situations, other methods (e.g. F_{ST}-based, assignment likelihood or ABC) may make it possible to determine whether the STRUCTURE results are misleading or not.

Acknowledgments

We thank Margarita Correa, Arnaud Estoup, Thibaut Malausa and Ferran Palero for fruitful comments and discussions. We also thank Alexandre Dehne-Garcia for assistance with the computer cluster.

References

Allendorf FW (1986) Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology*, **5**, 181–190.

Anderson EC, Dunham KK (2008) The influence of family groups on inferences made with the program Structure. *Molecular ecology resources*, **8**, 1219–1229.

Ascunce MS, Yang CC, Oakey J *et al.* (2011) Global Invasion History of the Fire Ant *Solenopsis invicta*. *Science*, **331**, 1066–1068.

Boissin E, Hurley B, Wingfield MJ *et al.* (2012) Retracing the routes of introduction of invasive species: the case of the *Sirex noctilio* woodwasp. *Molecular Ecology*, **21**, 5728–5744.

Bolte S, Fuentes V, Haslob H *et al.* (2013) Population genetics of the invasive ctenophore *Mnemiopsis leidyi* in Europe reveal source–sink dynamics and secondary dispersal to the Mediterranean Sea. *Marine Ecology Progress Series*, **485**, 25–36.

Ciosi M, Miller NJ, Kim KS *et al.* (2008) Invasion of Europe by the western corn rootworm, *Diabrotica virgifera virgifera*: multiple transatlantic introductions with various reductions of genetic diversity. *Molecular Ecology*, **17**, 3614–3627.

Cornuet J-M, Pudlo P, Veyssier J *et al.* (2014) DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, **30**, 1187–1189.

Cristescu ME (2015) Genetic reconstructions of invasion history. *Molecular ecology*, **24**, 2212–2225.

Dieni A, Brodeur J, Turgeon J (2016) Reconstructing the invasion history of the lily leaf beetle, *Lilioceris lilii*, in North America. *Biological Invasions*, **18**, 31–44.

Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why, how and so what? *Molecular Ecology*, **19**, 4113–4130.

Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*, **11**, 1591–

1604.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, **10**, 564–567.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fontaine MC, Gladieux P, Hood ME, Giraud T (2013) History of the invasion of the anther smut pathogen on *Silene latifolia* in North America. *New Phytologist*, **198**, 946–956.

Frantz AC, Cellina S, Krier A, Schley L, Burke T (2009) Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? *Journal of Applied Ecology*, **46**, 493–505.

Gaubert P, Del Cerro I, Centeno-Cuadros A *et al.* (2015) Tracing historical introductions in the Mediterranean Basin: the success story of the common genet (Genetta genetta) in Europe. *Biological Invasions*, **17**, 1897–1913.

Genton BJ, Shykoff JA, Giraud T (2005) High genetic diversity in French invasive populations of common ragweed, *Ambrosia artemisiifolia*, as a result of multiple sources of introduction. *Molecular Ecology*, **14**, 4275–4285.

Guillemaud T, Beaumont MA, Ciosi M, Cornuet JM, Estoup A (2010) Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, **104**, 88–99.

Guillemaud T, Blin A, Le Goff I *et al.* (2015) The tomato borer, *Tuta absoluta*, invading the Mediterranean Basin, originates from a single introduction from Central Chile. *Scientific Reports*, **5**, 8371.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. , **23**, 1801–1806.

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution*, **11**, 424–429.

Kalinowski ST (2011) The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity*, **106**, 625–632.

Karsten M, Jansen van Vuuren B, Addison P, Terblanche JS (2015) Deconstructing intercontinental invasion pathway hypotheses of the Mediterranean fruit fly (Ceratitis capitata) using a Bayesian inference approach: are port interceptions and quarantine protocols successfully preventing new invasions? *Diversity and Distributions*, **21**, 813–825.

Konecny A, Estoup A, Duplantier JM *et al.* (2013) Invasion genetics of the introduced black rat (*Rattus rattus*) in Senegal, West Africa. *Molecular Ecology*, **22**, 286–300.

Lachmuth S, Durka W, Schurr FM (2010) The making of a rapid plant invader: genetic diversity and differentiation in the native and invaded range of *Senecio inaequidens*. *Molecular Ecology*, **19**, 3952–3967.

Lawson Handley L-J, Estoup A, Evans DM *et al.* (2011) Ecological genetics of invasive alien species. *Biocontrol*, **56**, 409–428.

Lombaert E, Guillemaud T, Lundgren J *et al.* (2014) Complementarity of statistical treatments to reconstruct worldwide routes of invasion: the case of the Asian ladybird

Harmonia axyridis. *Molecular ecology*, **23**, 5979–5997.

Mallez S, Castagnone C, Espada M *et al.* (2015) Worldwide invasion routes of the pinewood nematode: What can we infer from population genetics analyses? *Biological Invasions*, **17**, 1199–1213.

Meirmans PG (2015) Seven common mistakes in population genetics and how to avoid them over confidence need diff analyses diff scales needed may not be able to do patterns errors. *Molecular Ecology*, **24**, 3223–3231.

Papura D, Burban C, van Helden M *et al.* (2012) Microsatellite and Mitochondrial Data Provide Evidence for a Single Major Introduction for the Neartic Leafhopper *Scaphoideus titanus* in Europe. *Plos One*, **7**, e36882.

Pascual M, Chapuis MP, Mestres F *et al.* (2007) Introduction history of *Drosophila subobscura* in the New World: a microsatellite-based survey using ABC methods. *Molecular Ecology*, **16**, 3069–3083.

Porras-Hurtado L, Ruiz Y, Santos C *et al.* (2013) An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in genetics*, **4**, 98.

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

Pritchard JK, Wen W, Falush D (2010) Documentation for structure software: Version 2. , Available from http://pritch.bsd.uchicago.edu.

Puechmaille SJ (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**, 608–627.

R Development Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes.

Proceedings of the National Academy of Sciences of the United States of America, **94**, 9197–9201.

Rewicz T, Wattier R, Grabowski M, Rigaud T, Bącela-Spychalska K (2015) Out of the Black Sea: Phylogeography of the Invasive Killer Shrimp *Dikerogammarus villosus* across Europe. *Plos One*, **10**, e0118121.

Rius M, Turon X, Ordonez V, Pascual M (2012) Tracking Invasion Histories in the Sea: Facing Complex Scenarios Using Multilocus Data. *Plos One*, **7**, e35815.

Robert S, Ravigne V, Zapater MF, Abadie C, Carlier J (2012) Contrasting introduction scenarios among continents in the worldwide invasion of the banana fungal pathogen *Mycosphaerella fijiensis*. *Molecular Ecology*, **21**, 1098–1114.

Sanz N, Araguas RM, Vidal O *et al.* (2013) Genetic characterization of the invasive mosquitofish (*Gambusia* spp.) introduced to Europe: population structure and colonization routes. *Biological Invasions*, **15**, 2333–2346.

Schwartz MK, McKelvey KS (2009) Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. *Conservation Genetics*, **10**, 441–452.

Simberloff D (2009) The Role of Propagule Pressure in Biological Invasions. *Annual Review of Ecology Evolution and Systematics*, **40**, 81–102.

Smith O, Wang J (2014) When can noninvasive samples provide sufficient information in conservation genetics studies? *Molecular Ecology Resources*, **14**, 1011–1023.

Tepolt CK, Darling JA, Bagley MJ *et al.* (2009) European green crabs (*Carcinus maenas*) in the northeastern Pacific: genetic evidence for high population connectivity and current-mediated expansion from a single introduced source population. *Diversity and Distributions*, **15**, 997–1009.

Thibault I, Bernatchez L, Dodson JJ (2009) The contribution of newly established populations to the dynamics of range expansion in a one-dimensional fluvial-estuarine system:

rainbow trout (*Oncorhynchus mykiss*) in Eastern Quebec. *Diversity and Distributions*, **15**, 1060–1072.

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Yu X, He T, Zhao J, Li Q (2014) Invasion genetics of *Chromolaena odorata* (Asteraceae): extremely low diversity across Asia. *Biological Invasions*, **16**, 2351–2366.

Zepeda-Paulo FA, Simon JC, Ramirez CC *et al.* (2010) The invasion route for an insect pest species: the tobacco aphid in the New World. *Molecular Ecology*, **19**, 4738–4752.

Zhang B, Edwards O, Kang L, Fuller S (2014) A multi-genome analysis approach enables tracking of the invasion of a single Russian wheat aphid (*Diuraphis noxia*) clone throughout the New World. *Molecular ecology*, **23**, 1940–51.

Zhou H-X, Zhang R-M, Tan X-M *et al.* (2015) Invasion Genetics of Woolly Apple Aphid (Hemiptera: Aphididae) in China. *Journal of Economic Entomology*, **108**, 1040–1046.

Author contributions

E.L., T.G. and E.D. conceived and designed the study. E.D. wrote the scripts and ran the simulations and analyses. E.L. and E.D. analyzed the data. E.L., T.G. and E.D. wrote the paper.

Tables

Table 1: Results obtained with the best model selected from the various statistical models run for the response variables "proportion of analyses yielding homogeneous misleading clusterings" and "proportion of analyses yielding heterogeneous misleading clusterings". Note: significant *P*-values, for a 5% threshold of significance, are shown in bold

Response variable	*Number of loci*	*Factors of selected model*	*Estimate*	*Std error*	*z-value (df=7499)*	*P*
Proportion of analyses yielding homogeneous misleading clusterings						
	10	*N*	-0.004	0.0005	-7.931	**<0.0001**
		$\log_{10}(N/NF)$	0.399	0.084	4.746	**<0.0001**
		N x $\log_{10}(N/NF)$	0.001	0.0002	5.185	**<0.0001**
	20	*N*	-0.005	0.0005	-9.658	**<0.0001**
	100	*N*	-0.0003	0.00006	-5.757	**<0.0001**
		$\log_{10}(N/NF)$	-0.194	0.259	-0.749	0.4540
		N x $\log_{10}(N/NF)$	-0.028	0.005	-5.190	**<0.0001**
Proportion of analyses yielding heterogeneous misleading clusterings						
	10	*N*	-0.0002	0.00007	-3.496	**0.0005**
		$\log_{10}(N/NF)$	1.107	0.114	9.672	**<0.0001**
		N x $\log_{10}(N/NF)$	0.00004	0.00002	1.827	0.0677
	20	*N*	-0.0004	0.00006	-5.977	**<0.0001**
		$\log_{10}(N/NF)$	0.574	0.079	7.178	**<0.0001**
		N x $\log_{10}(N/NF)$	0.00009	0.00002	4.854	**<0.0001**
	100	*N*	-0.0009	0.0001	-9.666	**<0.0001**
		$\log_{10}(N/NF)$	-0.801	0.068	-11.715	**<0.0001**
		N x $\log_{10}(N/NF)$	0.0003	0.00003	11.245	**<0.0001**

Figures

Figure 1: Simulated scenario and main observed STRUCTURE patterns at *K*=2. (**a**) Graphical representation of the simulated scenario in which two invasive populations (populations 2 and 3) are independently derived from the native population (population 1). *N* is the effective size at equilibrium and *NF* is the effective number of founders during the bottlenecks. (**b**) Schematic representations of the main patterns obtained in the STRUCTURE runs for *K*=2 and their associated summarized codes. The misleading pattern, inconsistent with the simulated scenario, is boxed. (**c**) Five examples of clusterings obtained over ten STRUCTURE runs for *K*=2, and their associated classification. In this study, we focused on "misleading homogeneous clusterings" and "misleading heterogeneous clusterings", in which "misleading patterns" were found in all ten runs or predominated, respectively, for a given dataset. Runs displaying the misleading pattern are boxed.

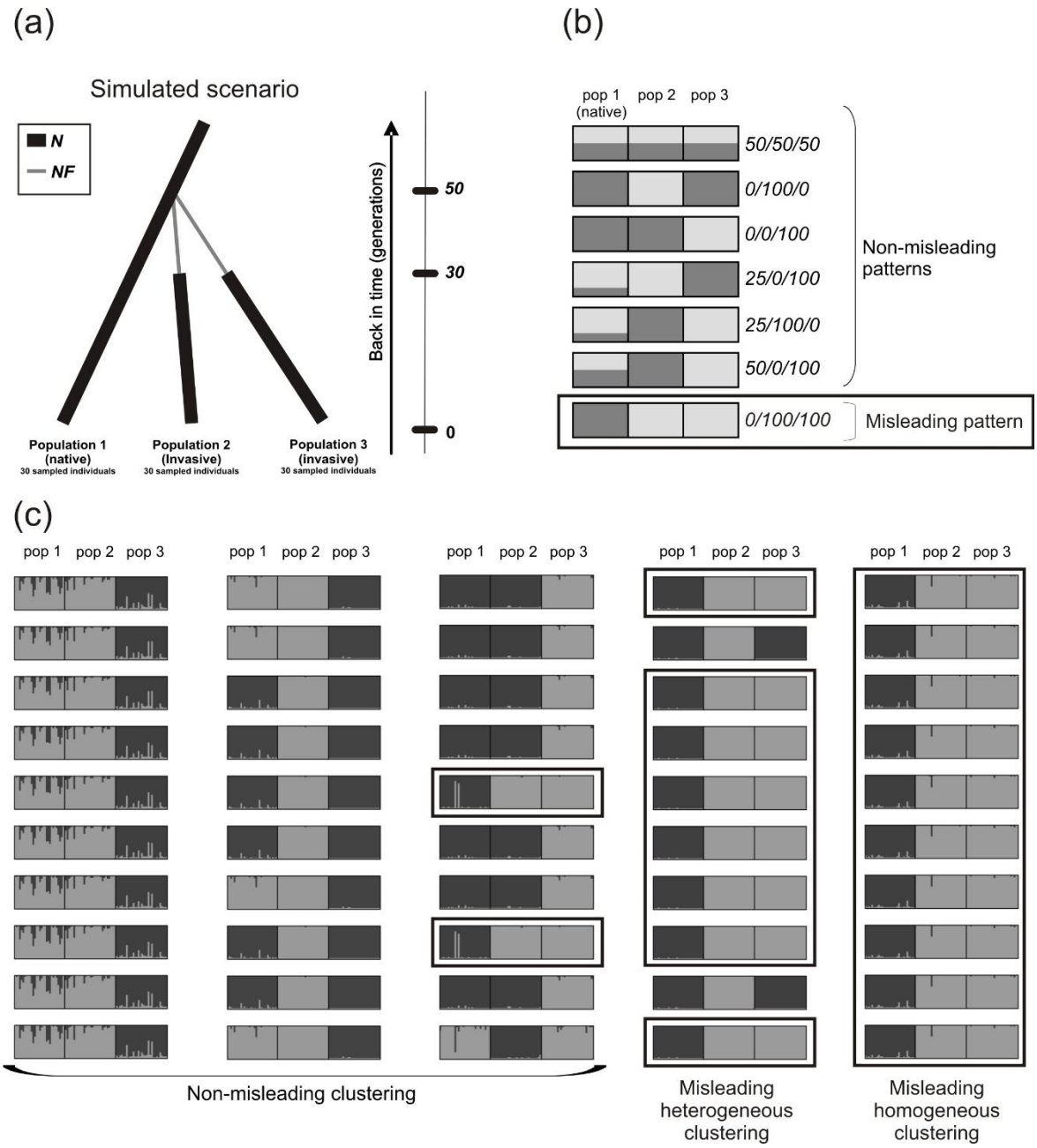

Figure 2: Distribution of the best number of clusters *K* inferred by Evanno's method for each number of loci, and the proportion for which there was an absence (homogeneous clustering) or presence (heterogeneous clustering) of genuine multimodality in the ten STRUCTURE runs carried out at *K*=2.

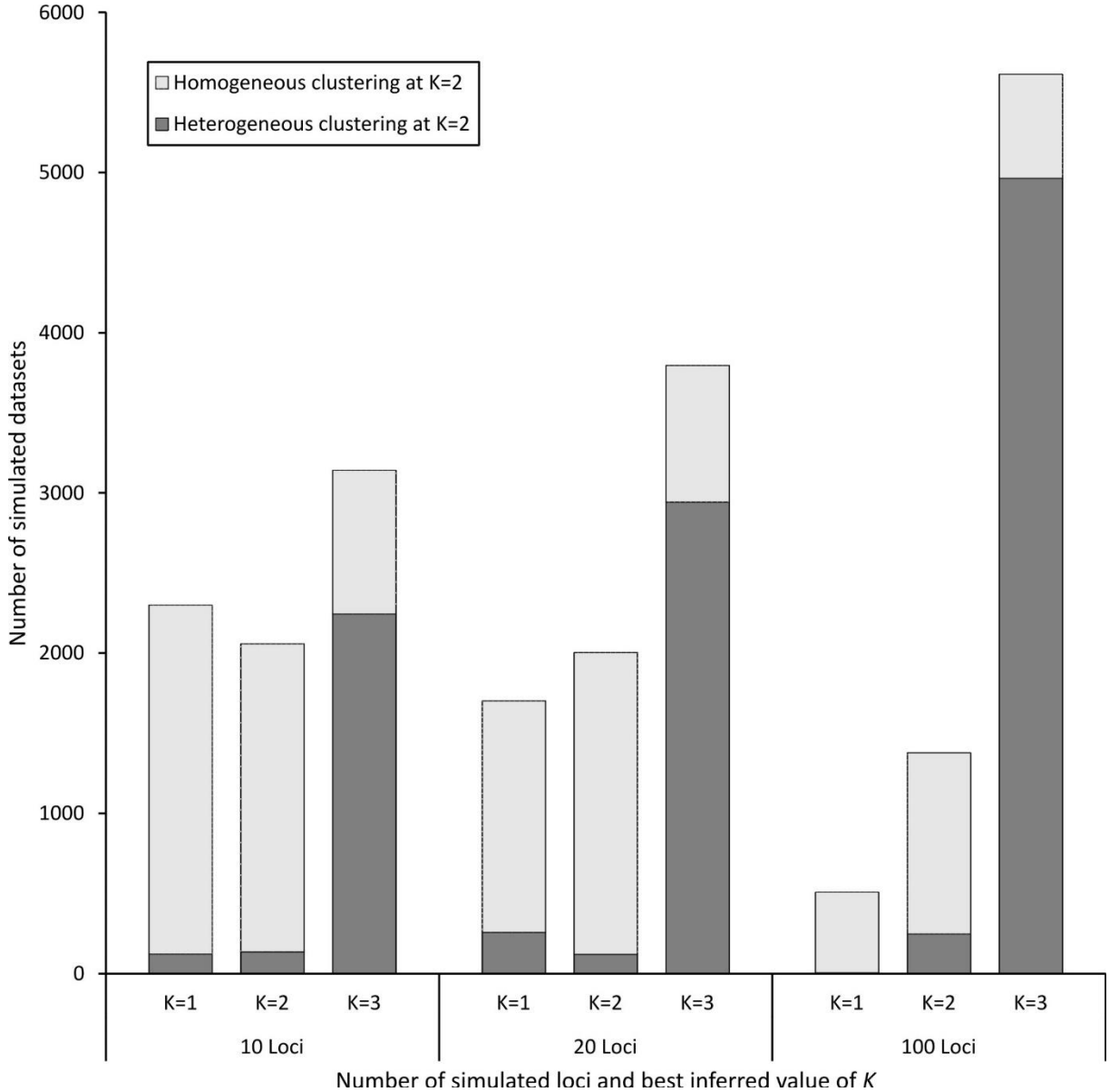

Figure 3: (**a**) Proportion of datasets with and without misleading patterns (Fig. 1b) for the ten STRUCTURE runs at *K*=2. (**b**) Best inferred number of clusters K obtained by Evanno's method for each number of loci within the datasets displaying misleading homogeneous (left) and misleading heterogeneous clusterings (right).

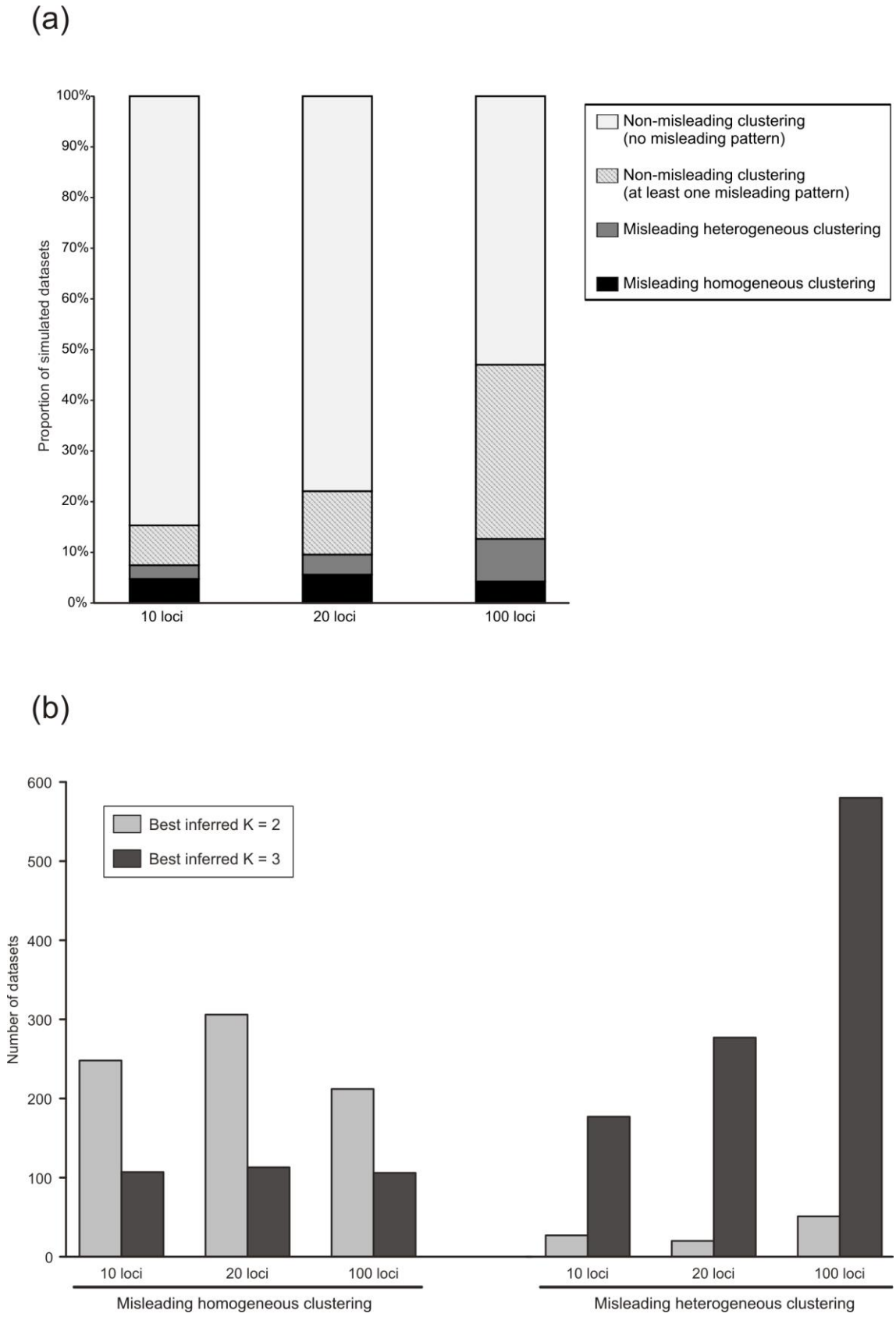
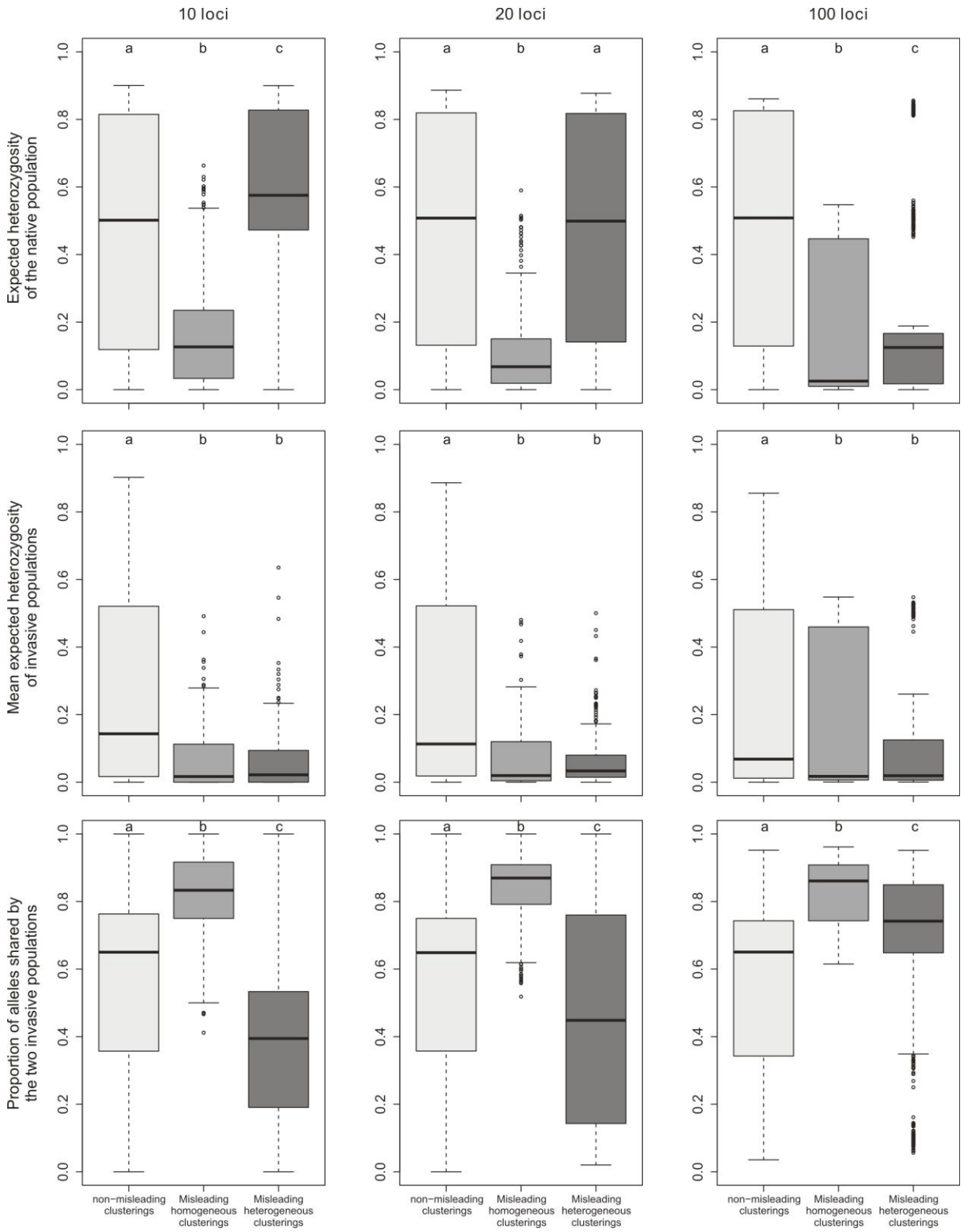
Figure 4: Tukey boxplots representing population genetics summary statistics for simulated datasets yielding non-misleading clusterings, misleading homogeneous clusterings or misleading heterogeneous clusterings. Within each frame, plots labeled with different letters are significantly different at the 5% level of significance (Mann-Whitney tests).