1   **A simulation-based evaluation of STRUCTURE software for**

2   **exploring the introduction routes of invasive species**

3

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

5

6   [1] INRA, UMR 1355 ISA, 06903 Sophia-Antipolis, France

7   [2] Université de Nice Sophia Antipolis, UMR ISA, 06903 Sophia-Antipolis, France

8   [3] CNRS, UMR 7254 ISA, 06903 Sophia-Antipolis, France

9

10   **Keywords:** biological invasion, microsatellite, invasion routes, STRUCTURE software,

11   simulation, bottleneck.

12

13   **Corresponding author:**

14   Eric Lombaert

15   INRA, UMR 1301 IBSV (INRA / Université de Nice Sophia Antipolis / CNRS). 400 Route des

16   Chappes. BP 167 - 06903 Sophia Antipolis cedex. FRANCE

17   E-mail: lombaert@sophia.inra.fr

18   Tel: +33 4 92 38 65 06

19   Fax: +33 4 92 38 64 01

20

21   **Running title:** Clustering analysis of introduction routes

22

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

46

47 One of the most fundamental questions in invasion biology is understanding what determines

48 the success of invasive alien species. Purely ecological factors may be the most important, but

49 it has been suggested that rapid evolution may also have played a major role in many

50 successful invasions (Lawson Handley *et al.* 2011). Retracing the introduction routes of

51 invasive alien species is a prerequisite to understanding the determinants of their success.

52 However, identification of the source of an introduced population is a complex task, because

53 of the highly stochastic nature of the introduction process (Estoup & Guillemaud 2010). Many

54 classic population genetics methods and tools are now widely used to retrace the introduction

55 routes of invasive species. This approach is somewhat risky, because the methods involved

56 are often dependent on demographic and genetic equilibria, but invasions often involve

57 demographic disequilibrium, through strong bottlenecks followed by rapid population growth,

58 for example. Despite this limitation and the risks of using population genetics methods

59 inappropriately in the specific context of biological invasions, only a few of these methods

60 have been formally evaluated (e.g. Estoup & Guillemaud 2010; Guillemaud *et al.* 2010).

61 Unsupervised individual Bayesian clustering methods are widely used in population

62 genetics. The popularity of these methods is due to their ability to infer genetic structure

63 correctly in many situations and their apparent simplicity (several "click-and-play" software

64 suites are available). STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003; Hubisz *et al.*

65 2009) is the most frequently used software for clustering in this context, with more than

66 25,000 citations for the three references indicated above in Google Scholar in November

67 2016. STRUCTURE aims to sort individuals in an unsupervised way into K clusters (K being

68 defined by the user), assuming Hardy-Weinberg/linkage equilibrium within clusters (Porras-

69 Hurtado *et al.* 2013). This method is known to perform well in most cases, but it can be

70   misleading in some situations, particularly in the presence of isolation by distance (Frantz *et*

71   *al.* 2009; Schwartz & McKelvey 2009), clusters of very different sizes (Kalinowski 2011;

72   Puechmaille 2016), the sampling of family groups (Anderson & Dunham 2008), or a high

73   proportion of missing data (Smith & Wang 2014).

74   STRUCTURE and other software suites based on similar methods are frequently used

75   in the context of introduction routes inferences (Estoup & Guillemaud 2010; Lawson Handley

76   *et al.* 2011; Cristescu 2015). STRUCTURE can be used to decrease the number of genetic

77   units from a large number of population samples to a few main clusters before quantitative

78   analyses, such as approximate Bayesian computation (Zepeda-Paulo *et al.* 2010; Ascunce *et*

79   *al.* 2011; Boissin *et al.* 2012; Rius *et al.* 2012; Konecny *et al.* 2013; Lombaert *et al.* 2014;

80   Gaubert *et al.* 2015; Guillemaud *et al.* 2015). In some cases, STRUCTURE is used directly to

81   contrast models of invasion history, particularly for comparisons of scenarios involving single

82   and multiple independent introductions from a native population. This approach is used to

83   determine the source of invasive populations. Indeed, if samples from an invasive population

84   group with the samples of only one of the potential source populations, this is considered to

85   provide fairly conclusive evidence about the origin of this invasive population (Lachmuth *et*

86   *al.* 2010; Robert *et al.* 2012; Bolte *et al.* 2013; Fontaine *et al.* 2013; Sanz *et al.* 2013; Zhang *et*

87   *al.* 2014; Yu *et al.* 2014; Zhou *et al.* 2015; Karsten *et al.* 2015; Mallez *et al.* 2015; Rewicz *et*

88   *al.* 2015; Dieni *et al.* 2016). However, invasions frequently involve major demographic

89   events, such as strong bottlenecks followed by genetic drift, which may significantly impair

90   our ability to determine introduction routes correctly from a given STRUCTURE result. This

91   problem may account for the contradictory outcomes sometimes obtained with different

92   population genetics methods (e.g. Mallez *et al.* 2015).

93   In this study, we evaluated the risk of incorrect introduction route inferences based on

94   STRUCTURE analyses, for the simple case of an invasion scenario involving one native

95   population and two independently introduced populations. We simulated a large number of

96   microsatellite datasets drawn from populations of various effective sizes and bottleneck

97   severities. STRUCTURE analyses were performed on these simulated datasets and the

98   resulting clustering patterns were classified as "misleading" or "non-misleading". We then

99   explored the effect of demographic parameters on the likelihood of misleading patterns being

100   obtained, to identify and predict the situations in which the use of STRUCTURE in a context

101   of introduction routes inference may be risky.

102

Methods

103   **Methods**

104

105   *Scenario description and data simulation*

106

107   We defined a simple historical scenario in which two invasive populations (populations 2 and

108   3) were independently founded 50 generations ago from the same native population

109   (population 1). Both invasive populations were subject to a demographic bottleneck lasting 20

110   generations (Fig. 1a). The effective sizes of all three populations at equilibrium (N) and the

111   effective number of founders of the two invasive populations during the bottlenecks (NF)

112   could take different values: 10000, 1000, 100, 10 and 2 individuals, with $N \geq NF$.

113    $\text{Log}_{10}(N/NF)$ was considered to quantify bottleneck severity.

114   We used DIYABC version 2.0.4 software (Cornuet *et al.* 2014) to generate 500

115   microsatellite multilocus genotype datasets for each of the 15 different combinations of N and

116   NF values, through a coalescent process. For all datasets, a sample of 30 diploid individuals

117   per population was simulated. We evaluated the effect of the number of loci on the analyses,

118   by performing simulations with 10, 20 and 100 unlinked microsatellite markers. We used a

119   generalized stepwise mutation model, with realistic values for all three parameters (Jarne &

5

120   Lagoda 1996; Estoup *et al.* 2002): the mean mutation rate (set to 5×10^{-4}), the mean parameter

121   of the geometric distribution defining the number of microsatellite repeats gained or lost

122   during mutation events (set to 0.22) and the mean mutation rate for single-nucleotide

123   insertion/deletion (set to 10^{-8}). In total, we simulated 22,500 datasets (15 sets of parameters x

124   500 datasets per set x 3 numbers of loci). We developed a pipeline with PERL scripts,

125   available on request, to automate the processing of the datasets (simulation and subsequent

126   STRUCTURE and post-STRUCTURE analyses).

127

128   *STRUCTURE analyses and misleading clustering*

129

130   For each of the 22,500 simulated datasets, a Bayesian clustering analysis was performed in

131   parallel, on a 120-node computer cluster, with STRUCTURE software version 2.3.4

132   (Pritchard *et al.* 2000). We chose the admixture model with correlated allele frequencies. We

133   used default values for all the other parameters. Each run consisted of a burn-in period of 10^5

134   Markov chain Monte Carlo (MCMC) iterations, followed by 5×10^5 MCMC iterations. This

135   run length is considered to be long enough to obtain precise estimates of parameters

136   (Pritchard *et al.* 2010), but we also tried runs of double this length for some combinations of

137   parameters with 100 loci. The results obtained were the same (data not shown). We carried

138   out ten replicate runs for each dataset and each value of K, the number of genetic clusters,

139   with K taking values of 1, 2, 3 and 4.

140   We investigated the ability of STRUCTURE to clarify introduction routes, by focusing

141   on analyses with $K=2$. With this K value, the two samples from an introduced population may

142   or may not cluster together. With the scenario simulated here, in which the two invasive

143   populations result from two independent introductions, the two samples of the introduced

144   populations would not be expected to cluster together. Indeed, the two independent drift

6

145   pulses at work during these two introductions (i.e. the bottleneck events) should make the

146   introduced populations more genetically different from each other than from the native

147   population, from which they are separated by a single drift pulse. Consequently,

148   STRUCTURE would yield a misleading pattern if the native population sample belonged to

149   one cluster and the two invasive population samples both belonged to the other. Indeed, this

150   could be considered evidence for a lack of independence of the two populations, with one

151   invasive population being the source of the other (Fig. 1b). Such a clustering pattern, hereafter

152   referred to as "misleading clustering", would lead most STRUCTURE users to an incorrect

153   interpretation, according to which a "successive introductions" scenario would be more likely

154   than the "independent introductions" scenario. Note that STRUCTURE analyses carried out

155   on three population samples with K=3 are, theoretically, unsuitable for comparisons of

156   independent and successive introduction scenarios, because each population sample would

157   probably form its own cluster.

158   For analysis of the 225,000 STRUCTURE runs with K=2 and estimation of the

159   frequency of misleading clusterings, the STRUCTURE output was characterized as follows.

160   From the output file of each run, we extracted the proportion of membership Qi_A and Qi_B of

161   population sample i for clusters A and B, respectively (with $Qi_B = 1 - Qi_A$). The Qi_A and Qi_B

162   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],

163   ]0.6;0.8[or [0.8;1] intervals, respectively. For each STRUCTURE run, we summarized the

164   clustering pattern by a code $C_{1A}/C_{2A}/C_{3A}$, where C_{iA} is the membership code of population

165   sample i for cluster A. For example, the clustering code would be *0/0/100* for a STRUCTURE

166   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

167   has no specific meaning, and the subscripts A and B can thus be permuted. For example,

168   clustering codes *0/0/100* and *100/100/0* summarize the same pattern and are pooled together

169   as *0/0/100*. Given the simulated scenario of independent introductions of the two invasive

170   populations, *0/100/100* was the code considered to correspond to misleading clustering (Fig.

171   1b). All other clustering codes were considered non-misleading in the context of introduction

172   routes inference. Focusing on the codes instead of the proportions of membership made it

173   possible to pool together slightly different clustering patterns in the same category.

174   Given the stochastic processes involved in the MCMC analysis, the ten replicated

175   STRUCTURE runs performed on a single dataset could conceivably generate different

176   results, a phenomenon called genuine multimodality (Jakobsson & Rosenberg 2007; Porras-

177   Hurtado *et al.* 2013). Clustering results for a given dataset were considered to be

178   homogeneous if the same clustering code (as defined above) was obtained in all ten runs.

179   They were otherwise considered to be heterogeneous. We evaluated the global occurrence of

180   misleading clustering in the analyses of the simulated datasets, and focused on two critical

181   categories of misleading clusterings (Fig. 1c):

182   (i) "Misleading homogeneous clusterings": for one dataset, all ten runs homogeneously

183   provide the misleading clustering pattern *0/100/100*.

184   (ii) "Misleading heterogeneous clusterings": for one dataset, the ten runs are not

185   homogeneous (i.e. genuine multimodality is observed) and the misleading clustering pattern

186   *0/100/100* predominates.

187   For each dataset, we also inferred the best value of *K,* as follows: if the mean natural

188   logarithm of the likelihood of the data $ln(P(X|K))$ with K in [1, 2, 3, 4] is maximal for $K=1$,

189   then the inferred number of clusters is 1; otherwise, we determined the best value of K by the

190   ΔK method (Evanno *et al.* 2005).

191

192   *Effect of demographic parameter values on misleading clustering*

193

194   For each number of simulated microsatellite loci (10, 20 or 100), the variables "the

195   proportion of analyses yielding misleading homogeneous clusterings" and "the proportion of

196   analyses yielding misleading heterogeneous clusterings" were analyzed independently with a

197   generalized linear model, using a binomial probability distribution of the residual error and a

198   logit link function. The following factors were included as fixed effects: effective population

199   size N and bottleneck severity $\log_{10}(N/NF)$. We used the Akaike information criterion (AIC)

200   to select the best model from the various models of different complexity. Analyses were

201   performed with R software V3.2.2 (R Development Core Team 2015).

202

203   *Link between summary statistics of genetic diversity and STRUCTURE patterns*

204

205   We summarized each simulated dataset, by using ARLSUMSTAT version 3.5 software

206   (Excoffier & Lischer 2010) to compute the mean number of alleles and the mean expected

207   heterozygosity in each population sample, and the pairwise F_{ST} values between each pair of

208   populations. We also used in-house PERL scripts to compute (i) the mean individual

209   assignment likelihood (Rannala & Mountain 1997) ($L_{i \to j}$) of each invading population

210   (samples 2 and 3) to each possible source population (i.e. either the native population or the

211   other invasive population), and (ii) the number of alleles shared by the invasive population

212   samples.

213   For the comparison of datasets leading to "misleading homogeneous clusterings",

214   "misleading heterogeneous clusterings" and "non-misleading clusterings", we specifically

215   explored a few genetic diversity summary statistics: (i) expected heterozygosity of the native

216   population sample, (ii) mean expected heterozygosity of both invasive population samples

217   and (iii) the ratio of alleles shared by the two invasive population samples to the total number

9

218   of alleles in the two samples. For each summary statistic and each number of loci, pairwise

219   Mann-Whitney tests with Holmes correction for multiple comparisons were performed.

220   We also compared STRUCTURE results with those obtained by two other methods

221   traditionally used to identify source populations: (i) the "F_{ST}-based method" and the (ii) the

222   "assignment likelihood-based method" (Genton *et al.* 2005; Pascual *et al.* 2007; Ciosi *et al.*

223   2008; Tepolt *et al.* 2009; Thibault *et al.* 2009; Papura *et al.* 2012; Mallez *et al.* 2015; Dieni *et*

224   *al.* 2016). For an "independent introductions" scenario, we would expect the F_{ST} between the

225   two invasive population samples to be larger than the F_{ST} values between the native

226   population and each of the invasive population samples (i.e. F_{ST} 2-3 > F_{ST} 1-2 and F_{ST} 2-3 >

227   F_{ST} 1-3). We would also expect both invasive population samples to be best assigned to the

228   native population sample (i.e. $L_{2 \rightarrow 1} > L_{2 \rightarrow 3}$ and $L_{3 \rightarrow 1} > L_{3 \rightarrow 2}$).

229   For each dataset, a global exact test for population genotypic differentiation (Raymond

230   & Rousset, 1995a) was carried out with GENEPOP software version 4.3 (Raymond &

231   Rousset, 1995b). If a dataset displayed no population differentiation, we made the prudent and

232   standard decision of not trying to infer any evolutionary relationship between the population

233   samples. Consequently, such datasets were considered to generate non-misleading results for

234   all methods.

235

Results

237

Effect of demographic parameter values on simulated datasets

239

240   The 500 simulated datasets for each parameter set are summarized with some common

241   statistics in Table S1. Decreasing effective population sizes (*N)* generate lower intra-

242   population and higher inter-population genetic diversities. Increasing bottleneck severity

243   ($\log_{10}(N/NF)$) generates lower intra-population genetic diversities for both invasive samples,

244   and overall higher inter-population genetic diversity. The main impact of a larger number of

245   loci is a decrease in the variance of all summary statistics. Overall, the chosen parameter

246   values (for N and NF) yield a large number of different combinations of genetic diversity for

247   evaluation of the ability of STRUCTURE software to explore introduction routes in different

248   situations.

249

250   *Overall STRUCTURE results*

251

252   The best value of K inferred by the ΔK method (Evanno *et al.* 2005) was most frequently

253   three (Fig. 2 and Fig. S1). The proportion of datasets for which the best number of clusters

254   was $K=3$ increased strongly with increasing numbers of loci (41.9%, 50.6% and 74.9% for 10,

255   20 and 100 loci, respectively). More than 80% of the simulated datasets for which $K=3$ was

256   inferred by the ΔK method had heterogeneous clustering codes (i.e. genuine multimodality) at

257    $K=2$ (Fig. 2). By contrast, when the number of inferred clusters was one or two,

258   multimodality at $K=2$ was found in less than 10% of all datasets.

259

260   *Occurrence of misleading STRUCTURE patterns*

261

262   Three categories of clustering codes at $K=2$ accounted for more than 95% of all runs (see

263   Table S2 for details): (i) clusterings in which all populations were fully admixed and

264   undistinguishable with STRUCTURE (i.e. the *50/50/50* code), (ii) clusterings in which the

265   two invasive samples belonged to two different clusters (i.e. the *C_{1A}/100/0* and *C_{1A}/0/100*

266   codes) and (iii) the misleading clusterings defined earlier, in which the two invasive samples

267   belonged to the same cluster, whereas the native sample belong to the other cluster (i.e. the

268   *0/100/100* code, Fig. 1b).

269   Overall, the proportion of datasets with at least one misleading clustering pattern over

270   the ten STRUCTURE runs ("misleading homogeneous clusterings", "misleading

271   heterogeneous clusterings" and non-misleading clusterings with at least one run yielding a

272   misleading pattern) was 15.31%, 22.07% and 47.01% for 10, 20 and 100 simulated loci,

273   respectively (Fig. 3a and Fig. S2), and very similar proportions were obtained with more (0.1

274   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

275   the encoding of pattern results (Table S3).

276   The frequency of "misleading homogeneous clusterings" was similar for different

277   numbers of loci, and was rather low overall (between 4.24% and 5.59% of the datasets, Fig.

278   3a). "Misleading heterogeneous clusterings" were also infrequent, but their frequency

279   increased with the number of loci: 2.71%, 3.96% and 8.41% for 10, 20 and 100 loci,

280   respectively (Fig. 3a). Overall, 7.45%, 9.55% and 12.65% of datasets for 10, 20 and 100 loci,

281   respectively, yielded misleading results. For some combinations of parameters, this

282   proportion reached 36.8% of datasets (Fig. S2). *K*=2 was most often (70%) inferred for

283   datasets yielding "misleading homogeneous clusterings", and *K*=3 was most often (91%)

284   inferred for datasets leading to "misleading heterogeneous clusterings" (Fig. 3b).

285

286   *Effect of demographic parameter values on STRUCTURE results*

287

288   For the response variable "proportion of analyses yielding misleading homogeneous

289   clusterings", the best model according to the AIC always included the effective population

290   size at equilibrium *N*, which was highly significant whatever the number of simulated loci

291   (Table 1). Lower *N* values resulted in a higher proportion of misleading homogeneous

12

292   clusterings (Fig. S3a). The best model also included bottleneck severity, $\log_{10}(N/NF)$, and the

293   interaction between the two main factors for 10 and 100 loci. Bottleneck severity was

294   significant only for 10 loci, and had a positive effect: the stronger the bottleneck, the higher

295   the proportion of misleading homogeneous clusterings. The interaction between the two

296   factors was significant in both models (Table 1 and Fig. S3a).

297   For the response variable "proportion of analyses yielding misleading heterogeneous

298   clusterings", the full model was selected for all numbers of simulated loci (Table 1). The

299   effective population size at equilibrium N was significant in all cases, and had a negative

300   effect (Fig. S3b). Bottleneck severity $\log_{10}(N/NF)$ was also strongly significant for all

301   numbers of loci, but its effect was positive for 10 and 20 loci and negative for 100 loci. The

302   interaction between the two factors was significant for 20 and 100 loci, with a positive effect

303   (Table 1).

304

305   *Links between summary statistics for genetic diversity and STRUCTURE patterns*

306

307   As expected from the analysis of demographic parameters, the diversity of the native

308   population, as assessed by its expected heterozygosity in the datasets with "misleading

309   homogeneous clusterings", was lower than that for "non-misleading" datasets, whatever the

310   number of loci considered (Fig. 4). On the contrary, no clear trend could be observed for

311   datasets with "misleading heterogeneous clusterings". For these datasets, the mean expected

312   heterozygosity was relatively high with 10 loci, intermediate with 20 loci and low with 100

313   loci, but, in each case, extreme low and high values were observed. The diversity of invasive

314   populations, which was affected by both the diversity of the native population and bottleneck

315   severity, was low for both kinds of misleading clusterings (Fig. 4). In comparisons with the

316   "non-misleading" datasets, the proportion of alleles shared by the two invasive populations

13

317   was higher for the datasets with "misleading homogeneous clusterings", and lower for

318   "misleading heterogeneous clusterings", unless 100 loci were considered (Fig. 4).

319   Outcomes for comparisons of STRUCTURE clusterings with results from F_{ST}-based

320   and assignment likelihood-based methods were very mixed, depending on the type of

321   "misleading clusterings" considered. 86.5%, 93.6% and 99.1% of datasets with "misleading

322   homogeneous clusterings" in STRUCTURE provided misleading results with at least one of

323   the methods based on F_{ST} or assignment likelihood, when considering 10, 20 and 100 loci,

324   respectively (Fig. S4). By contrast, datasets with "misleading heterogeneous clusterings" in

325   STRUCTURE analysis were rarely (for 10 and 20 loci), or at least not as strongly (for 100

326   loci), associated with misleading results with the other methods: this was the case for 15.2%,

327   24.9% and 67.5% of these datasets for 10, 20 and 100 loci, respectively (Fig. S4).

328

Discussion

330

331   We used simulated microsatellite datasets for a particular invasion scenario to determine

332   whether the method implemented in the widely used STRUCTURE software (Pritchard *et al.*

333   2000) could mislead users trying to infer introduction routes. For a true scenario of two

334   independent invasions from a single source, we found that STRUCTURE runs could give

335   misleading clustering patterns (i.e. the two invasive populations clustered together at K=2). In

336   about 10% of all simulated datasets, the results led to incorrect interpretation, with all

337   ("homogeneous misleading clusterings") or most ("heterogeneous misleading clusterings") of

338   the runs for a given dataset yielding the misleading pattern. Some combinations of

339   demographic parameters resulted in higher frequencies of misleading results with

340   STRUCTURE, and, contrary to expectations, increasing the number of loci also lead to an

341   overall increase in the frequency of misleading results. Our results suggested that the two

14

342   types of misleading clustering hazard, homogeneous and heterogeneous misleading

343   clustering, were very different. We suggest that (i) "homogeneous misleading clusterings"

344   probably arise from a large probability of independently drawing the same alleles twice from

345   a native population with low genetic diversity and that (ii) "heterogeneous misleading

346   clusterings" probably arise from convergence problems in STRUCTURE.

347   For "homogeneous misleading clustering", the effective size of the native population

348   has the strongest effect: the smaller this effective population size, the higher the risk of

349   obtaining misleading clustering patterns over all STRUCTURE runs. Such "homogeneous

350   misleading clustering" occurred principally when the two invasive populations shared a large

351   proportion of alleles, and the F_{ST}-based and likelihood assignment-based methods frequently

352   yielded the same clustering pattern. Accordingly, the number of clusters inferred by the

353   Evanno method was most frequently K=2. Invasive populations result from founder effects

354   and genetic drift (Simberloff 2009; Lawson Handley *et al.* 2011), which are random

355   processes. The probability of independently drawing the same alleles twice, with similar

356   frequencies, from a given native population is usually low (when random processes are at

357   work), but can actually be quite large when the diversity of the native population is itself low.

358   This is particularly true in cases of low heterozygosity (Allendorf 1986), in which one or a

359   few alleles occur at high frequency at each locus.

360   The interpretation of "heterogeneous misleading clusterings" is less clear-cut, but

361   several lines of evidence suggest the involvement of convergence issues in STRUCTURE

362   runs at K=2, when the best K value was undoubtedly 3. For 10 and 20 simulated loci, the

363   proportion of "heterogeneous misleading clusterings" (i) was positively related to bottleneck

364   severity, which accentuates differences between populations, (ii) was associated with a lower

365   proportion of alleles shared by the two invasive populations than for non-misleading datasets,

366   (iii) was higher when the F_{ST}- and assignment likelihood-based methods provided expected

15

367 results, and (iv) was higher in datasets for which K=3 was inferred with Evanno's method

368 (Evanno *et al.* 2005). With 100 loci, we observed a relatively large proportion of

369 "heterogeneous misleading clusterings" when the bottleneck was benign or not very serious,

370 probably because it was easier for STRUCTURE to cluster individuals correctly into the three

371 actual populations when there was a large number of loci (Evanno *et al.* 2005; Waples &

372 Gaggiotti 2006; Hubisz *et al.* 2009). Indeed, the ability of Evanno's method to infer K=3 and

373 the occurrence of multimodality at K=2 increases steadily with the number of loci. Overall,

374 these results are consistent with the notion that "heterogeneous misleading clusterings" often

375 occur when the three populations are clearly differentiated and should yield an unambiguous

376 K=3 result in STRUCTURE analyses. We therefore suggest that "heterogeneous misleading

377 clustering" probably results from a convergence problem in the MCMC procedure of

378 STRUCTURE: when an inappropriate number of clusters is imposed — here K=2 whereas

379 the data are more consistent with K=3 — multimodalities are often observed (Pritchard *et al.*

380 2000; Jakobsson & Rosenberg 2007), and sometimes, by chance, a large proportion of

381 misleading clustering events occur in the various runs, resulting in "heterogeneous misleading

382 clustering".

383

384 *Conclusion and general recommendations*

385

386 This study was based on a single simple invasion scenario with only three populations. More

387 complex scenarios should be studied in the future, but this study constitutes a crucial first

388 step, providing important information about the use of clustering methods in the context of

389 biological invasions.

390 We found that STRUCTURE yielded misleading results, but at a low

391 frequency. However, our results suggest that some situations should be analyzed with care.

16

392 First, invasion biologists should be very cautious if the diversity of the native population is

393 low: independent introductions from a single source population with low genetic diversity are

394 likely to produce genetic signals similar to that expected for successive introductions. Such

395 misleading results are difficult to spot, as they are consistent with the results of other

396 methods, such as F_{ST}- or assignment likelihood-based methods. In this context, quantitative

397 methods, such as approximate Bayesian computation (ABC), may be very useful. Second,

398 multimodal STRUCTURE results should be interpreted very cautiously, particularly if large

399 numbers of loci are used. This is sobering news, because many published studies interpret

400 STRUCTURE results at different K values, including those displaying genuine multimodality.

401 This problem is not specific to the exploration of introduction routes and has much wider

402 implications (Meirmans 2015). Multimodality is often a sign of poor convergence of

403 STRUCTURE runs, and is therefore likely to lead to results of limited biological meaning. In

404 such situations, other methods (e.g. F_{ST}-based, assignment likelihood or ABC) may make it

405 possible to determine whether the STRUCTURE results are misleading or not.

406

Acknowledgments

408

412

References

414

415 Allendorf FW (1986) Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology*,

416 **5**, 181–190.

17

417   Anderson EC, Dunham KK (2008) The influence of family groups on inferences made with

418   the program Structure. *Molecular ecology resources*, **8**, 1219–1229.

419   Ascunce MS, Yang CC, Oakey J *et al.* (2011) Global Invasion History of the Fire Ant

420   *Solenopsis invicta*. *Science*, **331**, 1066–1068.

421   Boissin E, Hurley B, Wingfield MJ *et al.* (2012) Retracing the routes of introduction of

422   invasive species: the case of the *Sirex noctilio* woodwasp. *Molecular Ecology*, **21**, 5728–

423   5744.

424   Bolte S, Fuentes V, Haslob H *et al.* (2013) Population genetics of the invasive ctenophore

425   *Mnemiopsis leidyi* in Europe reveal source–sink dynamics and secondary dispersal to the

426   Mediterranean Sea. *Marine Ecology Progress Series*, **485**, 25–36.

427   Ciosi M, Miller NJ, Kim KS *et al.* (2008) Invasion of Europe by the western corn rootworm,

428   *Diabrotica virgifera virgifera*: multiple transatlantic introductions with various

429   reductions of genetic diversity. *Molecular Ecology*, **17**, 3614–3627.

430   Cornuet J-M, Pudlo P, Veyssier J *et al.* (2014) DIYABC v2.0: a software to make

431   approximate Bayesian computation inferences about population history using single

432   nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, **30**,

433   1187–1189.

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

435   2212–2225.

436   Dieni A, Brodeur J, Turgeon J (2016) Reconstructing the invasion history of the lily leaf

437   beetle, *Lilioceris lilii*, in North America. *Biological Invasions*, **18**, 31–44.

438   Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why,

439   how and so what? *Molecular Ecology*, **19**, 4113–4130.

440   Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci

441   and their consequences for population genetics analysis. *Molecular Ecology*, **11**, 1591–

442   1604.

443 Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using

444   the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

445 Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform

446   population genetics analyses under Linux and Windows. *Molecular ecology resources*,

447   **10**, 564–567.

448 Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus

449   genotype data: Linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

450 Fontaine MC, Gladieux P, Hood ME, Giraud T (2013) History of the invasion of the anther

451   smut pathogen on *Silene latifolia* in North America. *New Phytologist*, **198**, 946–956.

452 Frantz AC, Cellina S, Krier A, Schley L, Burke T (2009) Using spatial Bayesian methods to

453   determine the genetic structure of a continuously distributed population: clusters or

454   isolation by distance? *Journal of Applied Ecology*, **46**, 493–505.

455 Gaubert P, Del Cerro I, Centeno-Cuadros A *et al.* (2015) Tracing historical introductions in

456   the Mediterranean Basin: the success story of the common genet (Genetta genetta) in

457   Europe. *Biological Invasions*, **17**, 1897–1913.

458 Genton BJ, Shykoff JA, Giraud T (2005) High genetic diversity in French invasive

459   populations of common ragweed, *Ambrosia artemisiifolia*, as a result of multiple sources

460   of introduction. *Molecular Ecology*, **14**, 4275–4285.

461 Guillemaud T, Beaumont MA, Ciosi M, Cornuet JM, Estoup A (2010) Inferring introduction

462   routes of invasive species using approximate Bayesian computation on microsatellite

463   data. *Heredity*, **104**, 88–99.

464 Guillemaud T, Blin A, Le Goff I *et al.* (2015) The tomato borer, *Tuta absoluta*, invading the

465   Mediterranean Basin, originates from a single introduction from Central Chile. *Scientific

466   Reports*, **5**, 8371.

467 Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure

468   with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–

469   1332.

470 Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program

471   for dealing with label switching and multimodality in analysis of population structure. ,

472   **23**, 1801–1806.

473 Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. *Trends*

474   *in Ecology and Evolution*, **11**, 424–429.

475 Kalinowski ST (2011) The computer program STRUCTURE does not reliably identify the

476   main genetic clusters within species: simulations and implications for human population

477   structure. *Heredity*, **106**, 625–632.

478 Karsten M, Jansen van Vuuren B, Addison P, Terblanche JS (2015) Deconstructing

479   intercontinental invasion pathway hypotheses of the Mediterranean fruit fly (Ceratitis

480   capitata) using a Bayesian inference approach: are port interceptions and quarantine

481   protocols successfully preventing new invasions? *Diversity and Distributions*, **21**, 813–

482   825.

483 Konecny A, Estoup A, Duplantier JM *et al.* (2013) Invasion genetics of the introduced black

484   rat (*Rattus rattus*) in Senegal, West Africa. *Molecular Ecology*, **22**, 286–300.

485 Lachmuth S, Durka W, Schurr FM (2010) The making of a rapid plant invader: genetic

486   diversity and differentiation in the native and invaded range of *Senecio inaequidens*.

487   *Molecular Ecology*, **19**, 3952–3967.

488 Lawson Handley L-J, Estoup A, Evans DM *et al.* (2011) Ecological genetics of invasive alien

489   species. *Biocontrol*, **56**, 409–428.

490 Lombaert E, Guillemaud T, Lundgren J *et al.* (2014) Complementarity of statistical

491   treatments to reconstruct worldwide routes of invasion: the case of the Asian ladybird

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

493   Mallez S, Castagnone C, Espada M *et al.* (2015) Worldwide invasion routes of the pinewood

494   nematode: What can we infer from population genetics analyses? *Biological Invasions*,

495   **17**, 1199–1213.

496   Meirmans PG (2015) Seven common mistakes in population genetics and how to avoid them.

497   *Molecular Ecology*, **24**, 3223–3231.

498   Papura D, Burban C, van Helden M *et al.* (2012) Microsatellite and Mitochondrial Data

499   Provide Evidence for a Single Major Introduction for the Neartic Leafhopper

500   *Scaphoideus titanus* in Europe. *Plos One*, **7**, e36882.

501   Pascual M, Chapuis MP, Mestres F *et al.* (2007) Introduction history of *Drosophila*

502   *subobscura* in the New World: a microsatellite-based survey using ABC methods.

503   *Molecular Ecology*, **16**, 3069–3083.

504   Porras-Hurtado L, Ruiz Y, Santos C *et al.* (2013) An overview of STRUCTURE:

505   applications, parameter settings, and supporting software. *Frontiers in genetics*, **4**, 98.

506   Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using

507   multilocus genotype data. *Genetics*, **155**, 945–959.

508   Pritchard JK, Wen W, Falush D (2010) Documentation for structure software: Version 2. ,

509   Available from http://pritch.bsd.uchicago.edu.

510   Puechmaille SJ (2016) The program STRUCTURE does not reliably recover the correct

511   population structure when sampling is uneven: subsampling and new estimators alleviate

512   the problem. *Molecular Ecology Resources*, **16**, 608–627.

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

514   R Foundation for Statistical Computing.

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

516   *Proceedings of the National Academy of Sciences of the United States of America*, **94**,

517   9197–9201.

518   Rewicz T, Wattier R, Grabowski M, Rigaud T, Bącela-Spychalska K (2015) Out of the Black

519   Sea: Phylogeography of the Invasive Killer Shrimp *Dikerogammarus villosus* across

520   Europe. *Plos One*, **10**, e0118121.

521   Rius M, Turon X, Ordonez V, Pascual M (2012) Tracking Invasion Histories in the Sea:

522   Facing Complex Scenarios Using Multilocus Data. *Plos One*, **7**, e35815.

523   Robert S, Ravigne V, Zapater MF, Abadie C, Carlier J (2012) Contrasting introduction

524   scenarios among continents in the worldwide invasion of the banana fungal pathogen

525   *Mycosphaerella fijiensis*. *Molecular Ecology*, **21**, 1098–1114.

526   Sanz N, Araguas RM, Vidal O *et al.* (2013) Genetic characterization of the invasive

527   mosquitofish (*Gambusia* spp.) introduced to Europe: population structure and

528   colonization routes. *Biological Invasions*, **15**, 2333–2346.

529   Schwartz MK, McKelvey KS (2009) Why sampling scheme matters: the effect of sampling

530   scheme on landscape genetic results. *Conservation Genetics*, **10**, 441–452.

531   Simberloff D (2009) The Role of Propagule Pressure in Biological Invasions. *Annual Review

532   of Ecology Evolution and Systematics*, **40**, 81–102.

533   Smith O, Wang J (2014) When can noninvasive samples provide sufficient information in

534   conservation genetics studies? *Molecular Ecology Resources*, **14**, 1011–1023.

535   Tepolt CK, Darling JA, Bagley MJ *et al.* (2009) European green crabs (*Carcinus maenas*) in

536   the northeastern Pacific: genetic evidence for high population connectivity and current-

537   mediated expansion from a single introduced source population. *Diversity and

538   Distributions*, **15**, 997–1009.

539   Thibault I, Bernatchez L, Dodson JJ (2009) The contribution of newly established populations

540   to the dynamics of range expansion in a one-dimensional fluvial-estuarine system:

541   rainbow trout (*Oncorhynchus mykiss*) in Eastern Quebec. *Diversity and Distributions*,

542   **15**, 1060–1072.

543   Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some

544   genetic methods for identifying the number of gene pools and their degree of

545   connectivity. *Molecular Ecology*, **15**, 1419–1439.

546   Yu X, He T, Zhao J, Li Q (2014) Invasion genetics of *Chromolaena odorata* (Asteraceae):

547   extremely low diversity across Asia. *Biological Invasions*, **16**, 2351–2366.

548   Zepeda-Paulo FA, Simon JC, Ramirez CC *et al.* (2010) The invasion route for an insect pest

549   species: the tobacco aphid in the New World. *Molecular Ecology*, **19**, 4738–4752.

550   Zhang B, Edwards O, Kang L, Fuller S (2014) A multi-genome analysis approach enables

551   tracking of the invasion of a single Russian wheat aphid (*Diuraphis noxia*) clone

552   throughout the New World. *Molecular ecology*, **23**, 1940–51.

553   Zhou H-X, Zhang R-M, Tan X-M *et al.* (2015) Invasion Genetics of Woolly Apple Aphid

554   (Hemiptera: Aphididae) in China. *Journal of Economic Entomology*, **108**, 1040–1046.

555

Author contributions

557

558   E.L., T.G. and E.D. conceived and designed the study. E.D. wrote the scripts and ran the

559   simulations and analyses. E.L. and E.D. analyzed the data. E.L., T.G. and E.D. wrote the

560   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**
		$N \times \log_{10}(N/NF)$	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
		$N \times \log_{10}(N/NF)$	-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**
		$N \times \log_{10}(N/NF)$	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**
		$N \times \log_{10}(N/NF)$	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**
		$N \times \log_{10}(N/NF)$	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.
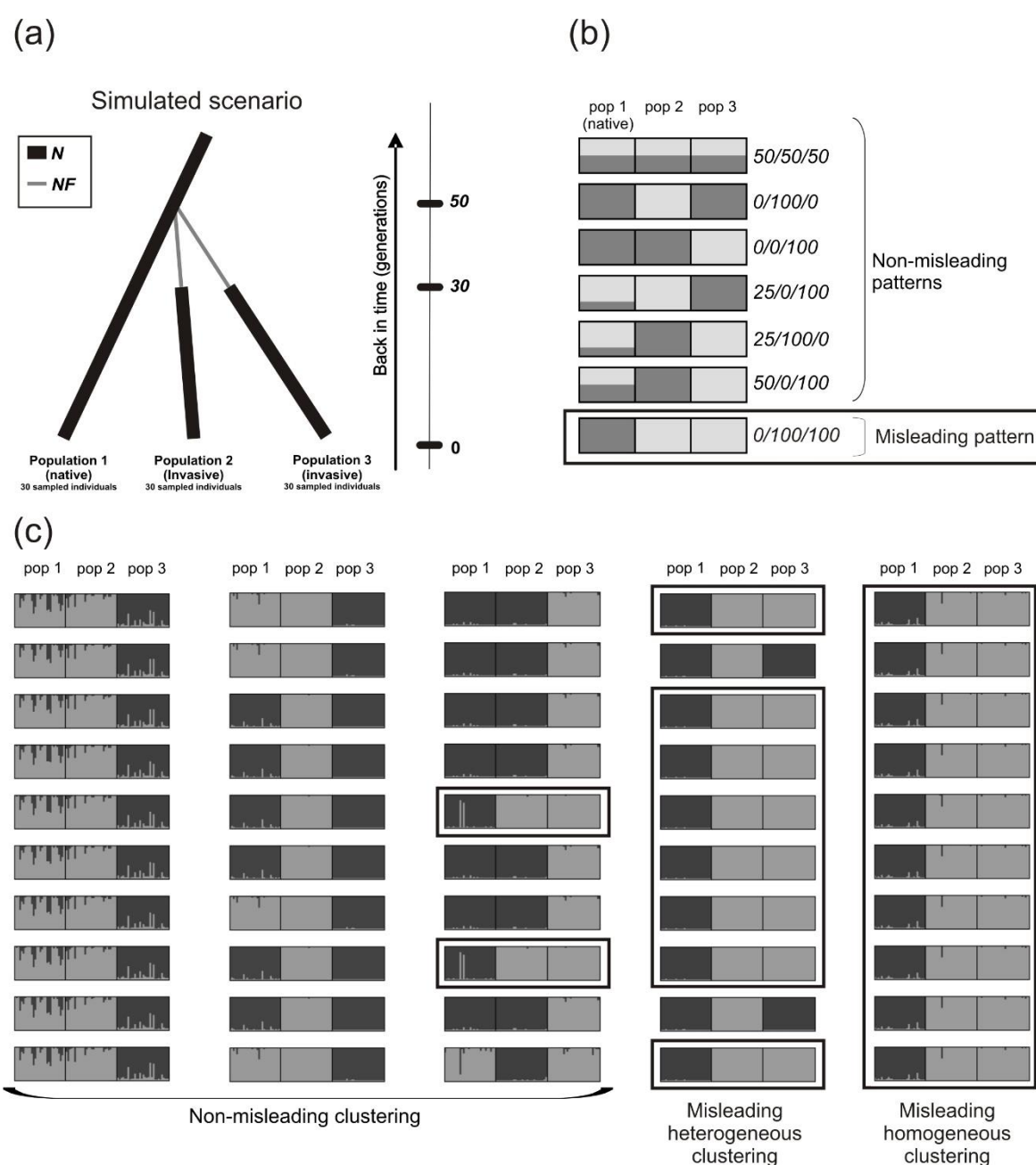
Figure 2: Distribution of the best number of clusters K inferred by Evanno's method for each number of loci, and the proportion for which there was an absence (homogeneous clustering) or presence (heterogeneous clustering) of genuine multimodality in the ten STRUCTURE runs carried out at K=2.
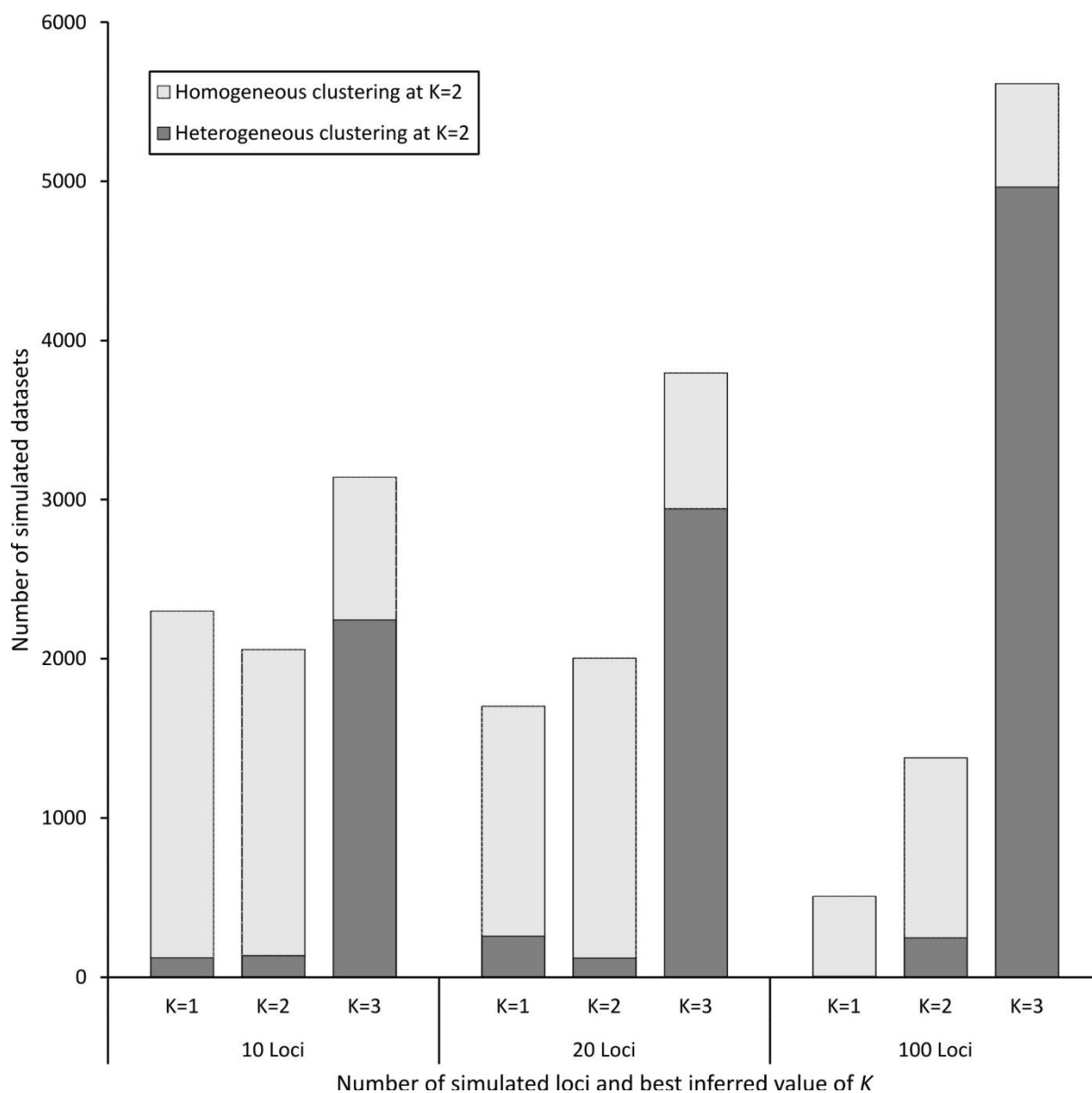
Figure 3: (**a**) Proportion of datasets with and without misleading patterns (Fig. 1b) for the ten STRUCTURE runs at *K*=2. (**b**) Best inferred number of clusters K obtained by Evanno's method for each number of loci within the datasets displaying misleading homogeneous (left) and misleading heterogeneous clusterings (right).
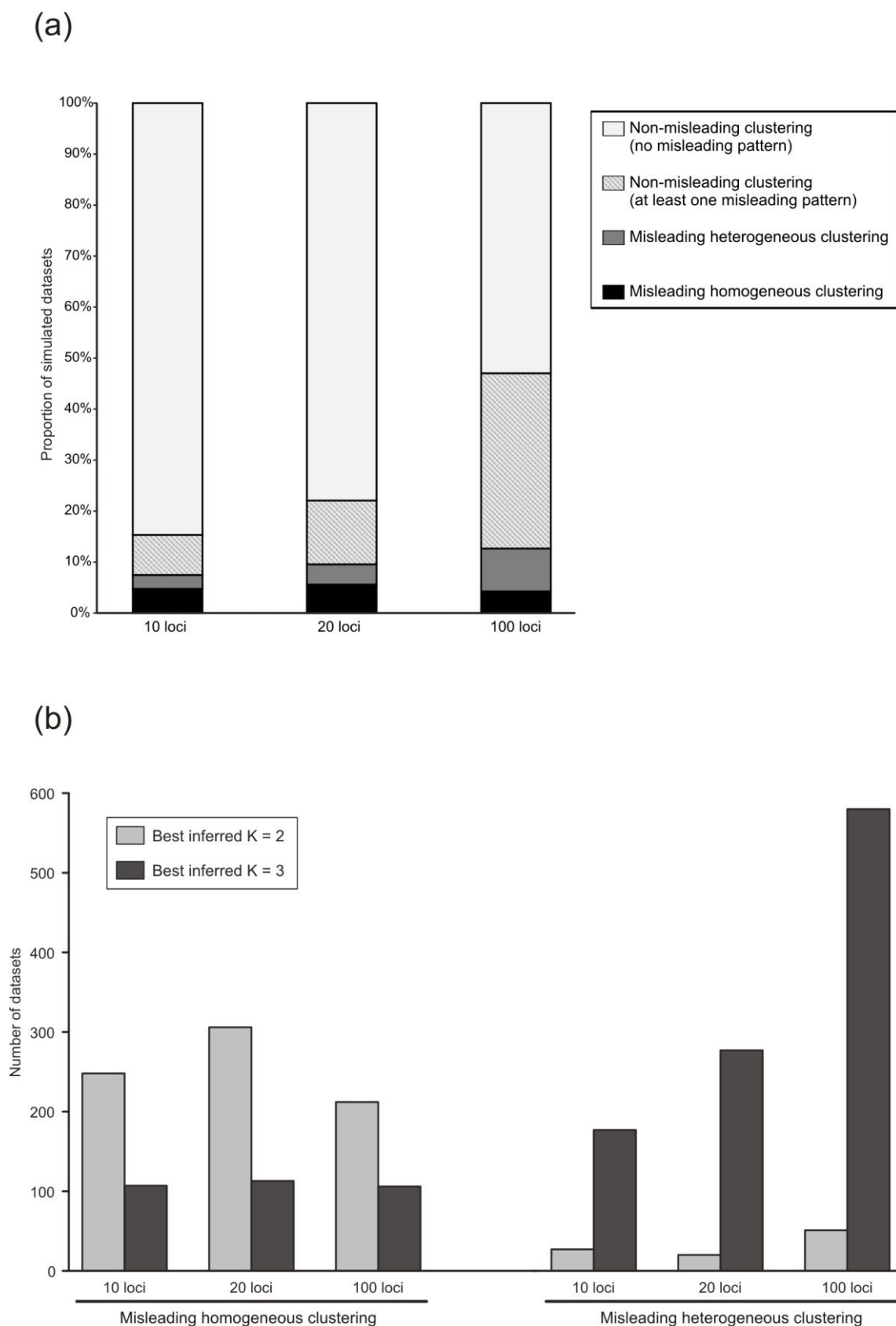
Figure 4: Tukey boxplots representing population genetics summary statistics for simulated datasets yielding non-misleading clusterings, misleading homogeneous clusterings or misleading heterogeneous clusterings. Within each frame, plots labeled with different letters are significantly different at the 5% level of significance (Mann-Whitney tests).