Heredity (2019) 122:276–287
https://doi.org/10.1038/s41437-018-0124-8

the geneticssociety

ARTICLE

Subsampling reveals that unbalanced sampling affects STRUCTURE results in a multi-species dataset

Patrick G. Meirmans[1]

Received: 7 February 2018 / Revised: 4 June 2018 / Accepted: 6 June 2018 / Published online: 19 July 2018

Abstract

Studying the genetic population structure of species can reveal important insights into several key evolutionary, historical, demographic, and anthropogenic processes. One of the most important statistical tools for inferring genetic clusters is the program STRUCTURE. Recently, several papers have pointed out that STRUCTURE may show a bias when the sampling design is unbalanced, resulting in spurious joining of underrepresented populations and spurious separation of overrepresented populations. Suggestions to overcome this bias include subsampling and changing the ancestry model, but the performance of these two methods has not yet been tested on actual data. Here, I use a data set of 12 high-alpine plant species to test whether unbalanced sampling affects the STRUCTURE inference of population differentiation between the European Alps and the Carpathians. For four of the 12 species, subsampling of the Alpine populations—to match the sample size between the Alps and the Carpathians—resulted in a drastically different clustering than the full data set. On the other hand, STRUCTURE results with the alternative ancestry model were indistinguishable from the results with the default model. Based on these results, the subsampling strategy seems a more viable approach to overcome the bias than the alternative ancestry model. However, subsampling is only possible when there is an a priori expectation of what constitute the main clusters. Though these results do not mean that the use of STRUCTURE should be discarded, it does indicate that users of the software should be cautious about the interpretation of the results when sampling is unbalanced.

Introduction

Almost all species show some form of genetic structure in the distribution of genetic variation. Be it a herb with an extremely limited distribution (Freville et al. 2001), a widely distributed tree species (Meirmans et al. 2017), or a planktonic species from the open ocean (Peijnenburg and Goetze 2013), there may be surprising genetic discontinuities across a species' range. Patterns of population structure can take many forms, from simple gradients resulting from limited dispersal to complex hierarchical patterns resulting from ecological adaptation to local conditions. Studying population structure can therefore be used to make inferences about the underlying evolutionary, historical, demographic, or anthropogenic processes, or a mixture of these (Lee and Mitchell-Olds 2011; Orsini et al. 2012; Nadeau et al. 2016).

One of the most widely used statistical tools for assessing population structure based on individual genotypes is the program STRUCTURE (Pritchard et al. 2000; Falush et al. 2003). STRUCTURE applies assignment of individuals to populations in a Bayesian framework, assuming Hardy–Weinberg equilibrium within clusters. While doing so, it performs a dual role: (1) assigning individuals to clusters, with the possibility of admixture between clusters, (2) finding the most suitable number of clusters (K) given the data. STRUCTURE generally performs both tasks very well and often gets results with a very intuitive biological explanation. Nevertheless, the method is not without flaws: Pritchard et al. (2000) themselves already acknowledged that STRUCTURE may give rise to spurious clustering in the presence of isolation by distance (see also Frantz et al. 2009; Meirmans 2012). Furthermore, the inference of the number of clusters is difficult as there may not be a single

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41437-018-0124-8) contains supplementary material, which is available to authorized users.

✉ Patrick G. Meirmans
p.g.meirmans@uva.nl

[1] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands

"optimal" value (Meirmans 2015), and different methods (Pritchard et al. 2000; Evanno et al. 2005) may yield different estimates (Janes et al. 2017).

Recently, several papers have pointed out that STRUCTURE is particularly sensitive to unbalanced sampling of populations (Kalinowski 2011; Neophytou 2013; Puechmaille 2016). With simulated data, even at the correct value of *K*, an unbalanced sampling design resulted in incorrect assignment of individuals to clusters; underrepresented populations tended to be clustered together even when they were not genetically more closely related. Conversely, the most sampled populations were often split into two or more spurious clusters with many individuals showing some degree of admixture. Patterns that are remarkably similar to those in Puechmaille's (2016) simulations can be observed when STRUCTURE is run on actual data sets. Puechmaille himself showed such a bias to be present in the analysis of Monarch butterflies. A similar pattern was noted earlier in a study of hybridisation between domesticated and wild species of cabbage, where the most sampled species (*Brassica rapa*) was split into two almost fully admixed clusters (Luijten et al. 2015). In order to reduce the bias that comes from unbalanced sampling, Puechmaille (2016) suggested subsampling the largest sample to match the size of the smaller ones. This subsampling strategy indeed removed the spurious clustering that was present in the largest sample both in Monarch butterflies (Puechmaille 2016) and in *Brassica* (P. Meirmans unpublished data).

In response to these simulation papers, Wang (2017) posited that a simple change in the settings of STRUCTURE may suffice to resolve the bias resulting from unbalanced sampling. The ancestry model used by STRUCTURE has a default setting where it is assumed that all source populations contribute equally to the total sample of individuals. Obviously, this is not the case when sampling is unbalanced. However, an alternative ancestry model can be used where a separate admixture parameter (alpha) is inferred for each cluster. Using simulated data and artificially unbalanced subsets of real data, Wang (2017) showed that changing the ancestry model setting improves STRUCTURE's performance with unbalanced data sets. Furthermore, Wang found that reducing the initial value of alpha, to a value of about 1/*K*, further improves the results.

In general, it is difficult to assess to what extent a bias that is present in simulated data is also present in real data sets. This is because the true situation cannot be known for real data sets. However, the subsampling strategy suggested by Puechmaille (2016) and the ancestry model settings suggested by Wang (2017) do present an opportunity to assess the incurred extent of the bias since the results of different methods can directly be compared. This is especially the case for species where unbalanced sampling coincides with an a priori hypothesis of divergence: e.g., for species where a large population is geographically widely separated from a much smaller population. Such species allow assessing whether any observed admixture between large and small populations is the result of bias or rather the result of unknown biological processes such as long-distance dispersal, recent fragmentation, or a shared evolutionary history.

Here, I use a data set of 12 high-alpine plant species to test whether unbalanced sampling affects the STRUCTURE inference of population differentiation between the European Alps and the Carpathians. The data set is remarkable in its scope (Gugerli et al. 2008) as all species have been uniformly sampled on the same regular grid in both mountain ranges. Even though the Carpathians are stretched out over a longer arc than the Alps, they have fewer high peaks and therefore fewer habitats for high-alpine species. Therefore, the regular sampling design resulted in an unbalanced data set, with more samples taken from the Alps than from the Carpathians (Fig. 1). The two mountain ranges are clearly geographically separated and also their floristic differences are well described (Tutin et al. 1964–1980). Therefore, it can be hypothesised that the population samples from these 12 species will generally fall into two clusters corresponding to the two mountain ranges. However, Gugerli et al. (2008) found for one of the included species—*Carex sempervirens*—partial overlap between the clusters between the Alps and the Carpathians. Such mismatch between the Structure results and the expectation may be either because the unbalanced sampling introduces a bias in STRUCTURE, or because the true divergence is different than hypothesised. Subsampling the population samples from the Alps should allow a distinction between these two possibilities.

Materials and methods

Data

The data were used from the IntraBioDiv-project (Gugerli et al. 2008; Alvarez et al. 2009; Taberlet et al. 2012), which contains AFLP data from 39 high-alpine species from the Alps and/or the Carpathians. From this data set, a subset of 12 species was selected that had a sufficient number of samples from both mountain ranges. Details of the sampling and AFLP-protocol can be found in Gugerli et al. (2008) and Alvarez et al. (2009). The greatest strength of this data set is that sampling was performed uniformly for all species: the area was divided into a regular grid with cell sizes of 20° longitude by 12° latitude (~20 by 22.5 km) and every second cell was extensively searched for the presence of all species. When a species was present, plant material was sampled from three individuals from a single location

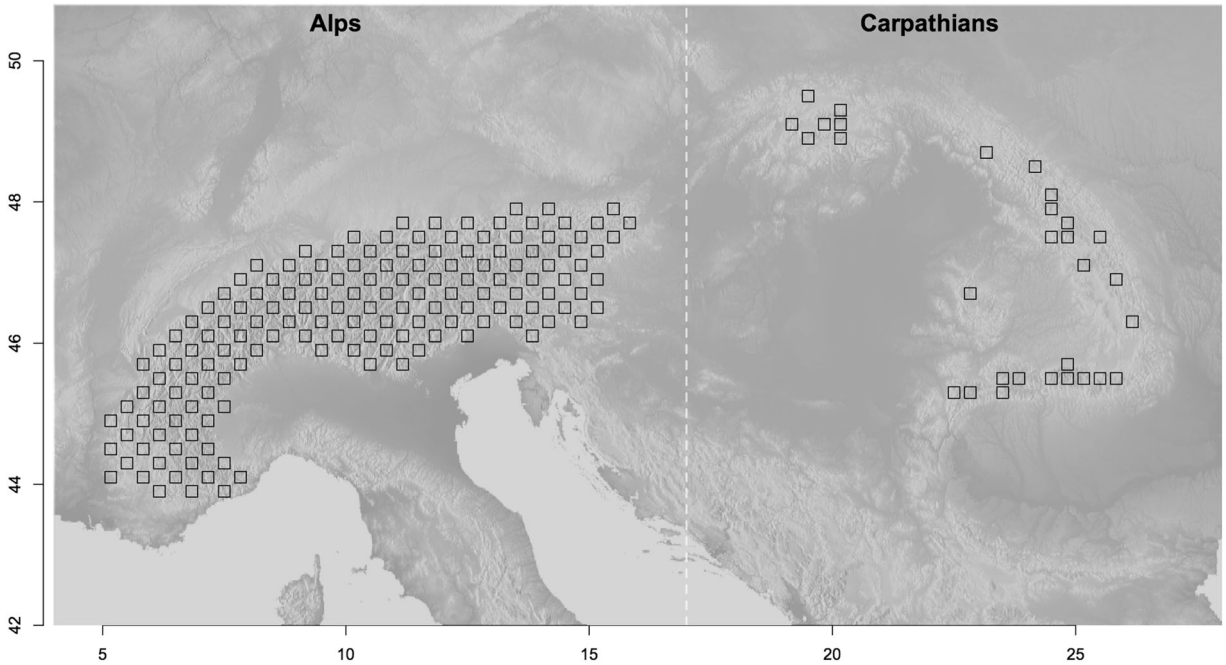

Fig. 1 Map of Central Europe showing the Alps and the Carpathians. Squares indicate all cells of the main IntraBioDiv grid where populations from the 12 species included here have been sampled

within the cell along a horizontal transect with 10 m distance between individuals. The total number of individuals per selected species ranged from 153 to 408 for the Alps and from 18 to 80 for the Carpathians (Table 1). Genotyping of the sampled individuals followed the standard protocol from Vos et al. (1995); bands were visualised either by electrophoresis on 8% polyacrylamide gels or on automatic capillary sequencers. The number of loci ranged from 61 for *Geum reptans* to 234 for *Luzula alpinopilosa*.

The data used was a slightly expanded version of the version stored in the Dryad database at: https://doi.org/10.5061/dryad/s4q6s. The Dryad version was split into separate data sets for the Alps and the Carpathians and contained for each mountain range only those loci that were polymorphic within that range. Since the sample sizes were much smaller in the Carpathians than in the Alps, the result was that the data set for the Carpathians contained fewer loci than that of the Alps, even though originally, both data sets contained the same set of loci. Since loci that are polymorphic in one range but monomorphic in the other are informative about the differentiation between the two mountain ranges, the original data set was used.

STRUCTURE analysis

For every species first a STRUCTURE (version 2.3.4, Pritchard et al. 2000) analysis was run for the full unbalanced data set, with all populations from the Alps and all populations from the Carpathians. The AFLP data were coded as suggested in the manual by including an extra row at the top that contained for every locus the name of the recessive allele (which was set to "0" for all loci). Since I was only interested in the distinction between the Alps and the Carpathians, which is expected to be the highest hierarchical level of clustering, I focused on the results when STRUCTURE was ran with $K = 2$. STRUCTURE was run with the admixture model, with uncorrelated allele frequencies, and without using the sampling locations as prior information. Changing those settings—correlated allele frequencies, no-admixture model, or using the Alps-Carpathian distinction as priors—did not notably change the results. The Monte Carlo Markov Chain was run for 100,000 steps, after a burnin period of 10,000 steps; trial runs suggested that this was enough to reach convergence. Ten replicate analyses were run for every data set, and the results of the run with the highest overall likelihood, according to the ln $Pr(X|K)$ statistic, was used for interpretation.

To assess any bias resulting from sampling more populations in the Alps, I used both the ancestry model settings suggested by Wang (2017) and the subsampling strategy suggested by Puechmaille (2016). The ancestry model was changed by setting the parameter POPALPHAS in the "extraparams" file to a value of 1; in the graphical user interface of STRUCTURE this corresponds to checking the box labelled "separate Alpha for each population" under the "Advanced…" ->"Configure" option of the "Ancestry Model" settings. Two values of the initial value of alpha were used—the default of 1.0, and 0.5–, by setting the

Table 1 Overview and results of the Structure analyses of AFLP data from 12 Alpine species from the Alps and the Carpathians

Species	Abbreviation	# loci	# pops Alps	# pops Carp.	F_{ST}	F_{SC}	F_{CT}	β_{AC} (full data)	β_{AC} (Wang, alpha = 1.0)	β_{AC} (Wang, alpha = 0.5)	β_{AC} (mean of subsamples)
Arabis alpina	Aal	151	129	19	0.72	0.67	0.13	0.46	0.44	0.44	0.69
Carex sempervirens	Cse	125	137	22	0.41	0.33	0.12	0.38	0.38	0.38	0.63
Dryas octopetala	Doc	101	124	15	0.37	0.22	0.19	0.64	0.65	0.65	0.92
Geum montanum	Gmo	93	122	19	0.51	0.35	0.25	0.58	0.59	0.59	0.86
Geum reptans	Gre	61	51	8	0.65	0.42	0.39	0.78	0.82	0.82	0.92
Hedysarum hedysaroides	Hhe	123	76	11	0.86	0.86	0.05	0.17	0.17	0.17	0.37
Hypochaeris uniflora	Hun	102	59	27	0.52	0.26	0.36	0.80	0.77	0.78	0.76
Juncus trifidus	Jtr	88	91	23	0.38	0.28	0.14	0.51	0.51	0.51	0.51
Loiseleuria procumbens	Lpr	121	90	13	0.40	0.31	0.13	0.09	0.07	0.07	0.74
Luzula alpinopilosa	Lal	234	82	19	0.45	0.32	0.19	0.76	0.74	0.74	0.69
Saxifraga stellaris	Sst	199	101	12	0.52	0.43	0.16	0.56	0.56	0.57	0.88
Sesleria caerulea	Sco	70	137	7	0.85	0.85	0.02	0.54	0.50	0.50	0.40

Given are the number of loci, the number of populations sampled in the two mountain ranges, the *F*-statistics from an AMOVA, and the difference in the Structure assignments between the Alps and the Carpathians (β_{AC}). The β_{AC} values are given for the analysis of the full data set, the analysis with the alternative ancestry model (with two initial values of the alpha parameter), and for the analysis where the data set was subsampled to balance the number of population samples in the two mountain ranges (average shown over 500 replicate subsamples)

ALPHA parameter in the extraparams file (corresponding to "Initial Alpha" in the GUI). Besides these parameters, STRUCTURE was run with the same settings as above.

Subsampling was done for every species separately by creating 500 subsampled data sets where the number of sampling locations in the Alps matched the number of locations in the Carpathians. For every subsampled data set all Carpathian populations were included plus a random sample (without replacement) of an equal number of populations from the Alps. For example, for *Arabis alpina* each subsample consisted of 38 populations: 19 from the Carpathians and 19 from the Alps, the latter randomly sampled from the 129 Alpine populations. Each subsampled data set was analysed in STRUCTURE at $K = 2$ with the same parameter settings as the full data set, using the default option to set the random number seed based on the system clock. Comparing the output of multiple STRUCTURE runs can be challenging, as the labelling of clusters is arbitrary in every run. There are algorithmic approaches for solving this (Jakobsson and Rosenberg 2007), but these require that the same individuals are present in every replicate, which is not the case with repeated subsampling. Here, I based the cluster alignment of the subsamples on the results of the analysis with the full data set: for every replicate analysis, I switched the labels in such a way that the sum of squared deviations between the assignments of the subsampled and full data were minimised. I then proceeded by calculating for every location the average assignment to the two clusters over all subsamples in which the location was included. These average assignments were then plotted on a map to provide a visual way to compare them to the assignments from the full, unbalanced, data set.

A simple test statistic was selected to quantify the degree to which STRUCTURE returned separate clusters for the Alps and the Carpathians. This test statistic (β_{AC}) was calculated by taking the absolute value of the variable coefficient ("slope") of an Analysis of Variance with mountain range as the explanatory variable and the STRUCTURE *q*-values as the response variable (calculated using the lm() function in R). The β_{AC}-statistic is equivalent to calculating for every mountain range the mean proportion of individuals assigned to the first cluster, and then taking the absolute difference between the two mountain ranges. When the two mountain ranges harbour genetically completely separated clusters, the value of β_{AC} equals 1; when they contain exactly equal proportions of the two clusters, the value of β_{AC} equals zero.

For every species, the β_{AC} statistic was calculated on the STRUCTURE results for the full data set and for all 500 replicate subsamples. When the STRUCTURE results of the full data set are biased by the unbalanced sampling design, the value of β_{AC} is expected to be substantially lower for the full data set than for the subsampled data sets. This is not a formal statistical test and cannot be used to calculate *p*-values, but should nevertheless give a good indication of whether there is a difference between the analyses of the full and subsampled data sets. Note that the β_{AC} statistic is

meant to compare the strength of clustering compared to a priori defined groups (here the two mountain ranges) and should not be confounded with other summary statistics such as ΔK (Evanno et al. 2005) and ln $Pr(X|K)$ (Pritchard et al. 2000) which are meant to compare the clustering at different values of K. Furthermore, neither the coefficient of determination (r^2) nor the p-values of the LM can be used to assess the strength of the association between the mountain ranges and the STRUCTURE results as the former is affected by unbalanced sampling and the latter by sample size.

For the data analysis, I mostly focused on $K = 2$ as I explicitly wanted to assess how well the Structure results match the a priori expectation of a differentiation between the Alps & the Carpathians. However, it is also of interest to investigate other values of K, and to assess which value of K has the strongest support in each species. To this end, STRUCTURE was run for every species with K values from 1 to 11, for the full data set and for 100 out of the 500 subsampled data sets. The same settings were used as detailed above, so with ten replicate runs per value of K. For every data set, I calculated the ΔK-statistic (Evanno et al. 2005) to select the value of K with the strongest support.

In addition to the STRUCTURE analysis, a hierarchical AMOVA (Excoffier et al. 1992) was performed for each species, with the populations clustered into two groups corresponding to the two mountain ranges. The main objective of this AMOVA was to estimate the F_{CT}-statistic, which quantifies the degree of population divergence between the Alps and the Carpathians. This was done for the full data set and for every subsampled data set, using the function *poppr.amova*() from the R-package POPPR (Kamvar et al. 2014). The STRUCTURE and AMOVA analyses were performed and the results were parsed using custom scripts in R; these scripts can be found in Dryad package https://doi.org/10.5061/dryad.nh4366s.

Results

STRUCTURE with unbalanced data sets

When STRUCTURE was run using the full unbalanced data sets, only a few species showed separate clusters for the Alps and the Carpathians (Fig. 2, top graph for every species). Generally, the populations from the Carpathians were all grouped in the same cluster (sometimes with a bit of admixture), but they shared this cluster with multiple populations from the Alps. It was not always the case that the Carpathian populations were grouped together with the populations from the Alps that are geographically the closest. This is most notable in *Arabis alpina* and *Saxifraga stellaris* where the Carpathian populations cluster together with the western-most populations from the Alps.

One species (*Hypochaeris uniflora*) showed a pattern that was distinctly different in this respect (Fig. 2): here all populations from the Alps formed a cluster together with the populations from the Western Carpathians, with the rest of the Carpathian populations forming the second cluster. Interestingly, this was also the species where the sampling was most balanced with 27 populations sampled in the Carpathians and 59 in the Alps; this represents a ratio of 1:2.2, whereas over all species this is on average 1:7.2. *Hypochaeris uniflora* also showed strong genetic differentiation between the two mountain ranges; it had the second-highest F_{CT} value at 0.36.

STRUCTURE with alternative ancestry model

Changing the ancestry model to infer a separate value of alpha for each population did not notably change the results. Plotting the assignments yielded almost exactly the same patterns (Fig. S1) as the run with the default ancestry model; the value of the β_{AC} statistic was also close to the value obtained with the default setting (Table 1). Changing the initial value of alpha from its default value of 1.0 to a value of 0.5 also did not have any affect on the results (Fig. S1).

STRUCTURE with subsampled balanced data sets

In nine out of the 12 species, subsampling to create balanced data sets increased the separation between the Alps and the Carpathians in the STRUCTURE results as quantified by the β_{AC} statistic (Fig. 3a). Four of those species stood out in that they showed near-complete separation between the two mountain ranges in the subsampled balanced data sets, but not in the full unbalanced data sets (Fig. 2): *Dryas octopetala*, *Geum montanum*, *Loiseleuria procumbens*, and *Saxifraga stellaris*. For these species, the β_{AC} statistic for the unbalanced data set was located in the lower 2.5% percentile of the distribution of β_{AC} scores for the subsampled data sets (Fig. 3a). These species therefore represent cases where the unbalanced sampling design has lead to a consequential difference in the results. However, for the other species the difference between the balanced and unbalanced data sets were only slight. There was also one species, *Luzula alpinopilosa*, where STRUCTURE returned less separation between the two mountain ranges when sampling was balanced. In the subsampled data sets, the populations from the Western Carpathians were clustered together with all populations from the Alps.

Within several species, there was a large degree of variation in the values of the β_{AC} scores among the replicate subsamples (Fig. 3a, showing the percentiles of the distribution of q-values across replicates). This variation was largest in *Arabis alpina*, where β_{AC} ranged from a minimum of 0.0042 (almost equal assignment to the two clusters in

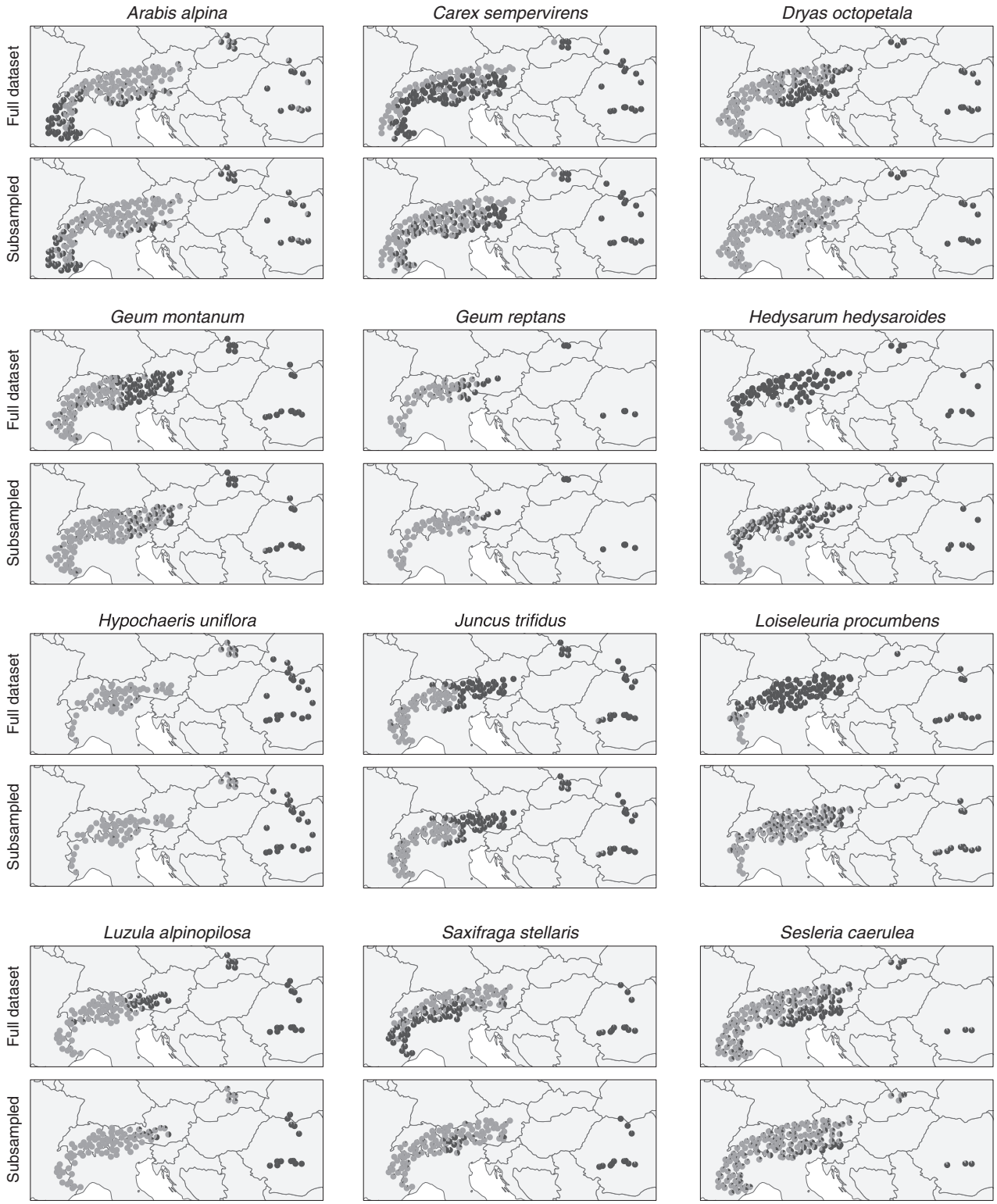

Fig. 2 Maps showing the results of STRUCTURE analyses of AFLP data for 12 alpine species, for both the full unbalanced data sets and subsampled balanced data sets. Pies represent the assignments (q-values) to $K = 2$ clusters, averaged over the three individuals that were sampled at each location. The maps for the subsampled data sets were calculated by averaging the assignments over 500 replicate analyses per species

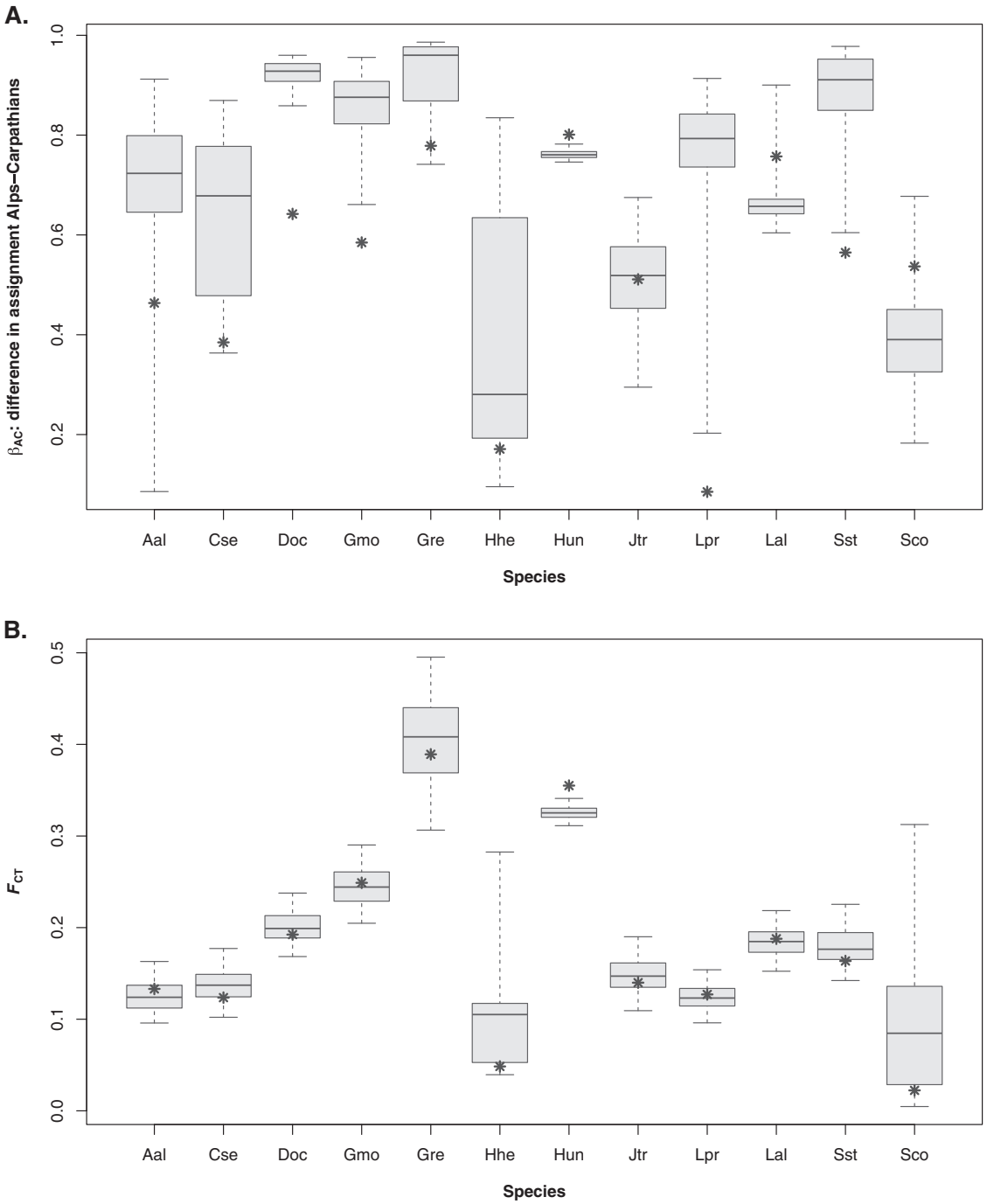

Fig. 3 Divergence between populations from the Alps and from the Carpathians for 12 alpine species based on the results of STRUCTURE (a; β_{AC} statistic) and the results of an AMOVA (b; F_{CT}). Asterisks represent the results of the full data set with unbalanced sampling; boxplots represent the distribution of the results of the 500 replicate subsamples where sample sizes from the Alps matched those from the Carpathians (thick line gives the median; box gives 25% and 75% percentiles; whiskers give 2.5% and 97.5% percentiles)

the Alps and in the Carpathians) to a maximum of 0.96 (clustering almost coincided completely with the two mountain ranges). Other species with notably large ranges in β_{AC} with the subsampled data include *Hedysarum hedysaroides* (0.092–0.94) and *Loiseleuria procumbens* (0.15–0.94). The variation in assignments can also be visualised by calculating the standard deviation across replicates for every population separately (Fig. 4). This shows for some species remarkable geographical patterns. In some species—e.g. *Dryas octopetala*—the standard deviation is uniformly low. In other species—e.g. *Carex sempervirens*—it was low in some parts of the sampling

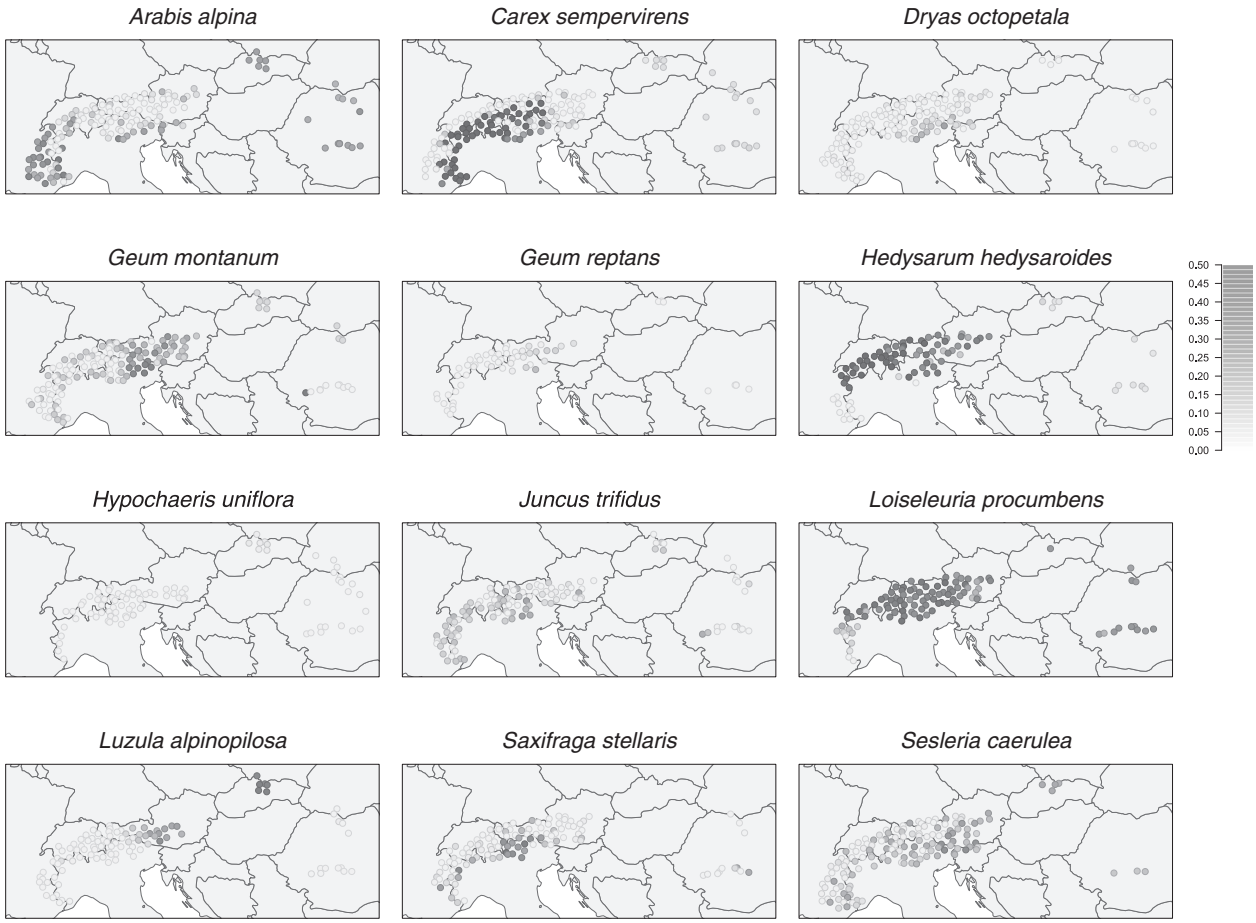

Fig. 4 Maps showing per sampling location the standard deviation in STRUCTURE assignment over 500 replicate subsamples, where the number of sampling locations in the Alps was reduced to match the number of locations in the Carpathians

range but high in other parts. Finally, in some species—most notably *Loiseleuria procumbens*—it was high throughout almost the whole-sampling range.

In contrast with β_{AC}, the variation in F_{CT} across subsamples was generally much smaller (Fig. 3b; note difference in scale with 3a). For F_{CT}, the value for the full data set was also generally very close to the median of the values for the subsampled data sets; with the exception of *Hypochaeris uniflora*.

Other values of K

The ΔK-statistic indicated for ten out of the 12 species an optimal value of $K = 2$ clusters (lines in Fig. 5). The only exceptions were *Geum reptans* and *Juncus trifidus*, which both showed the highest value of ΔK at $K = 3$. In addition, *Carex sempervirens* showed a ΔK value for $K = 3$ that was only slightly lower than that for $K = 2$. Despite the general support for two clusters, most species showed distinct geographical patterns for the clusters at higher values of K (Fig. S2, showing up to $K = 5$), indicating that these may be well worth a biological explanation, despite not having the strongest support. The histograms in Fig. 5 show how frequently the different values of K were inferred to be the optimal value among 100 of the subsampled data sets. These histograms show that in most species, there was considerable variation in the optimal values of K among the subsampled data sets.

Discussion

The results of the analysis of genetic data from 12 alpine species confirm previous simulation results that STRUCTURE (Pritchard et al. 2000) may have a bias when population sampling is unbalanced (Kalinowski 2011; Neophytou 2013; Puechmaille 2016). In four out of the 12 species, the distinction between the Alps and the Carpathians increased drastically when the sample sizes from the Alps were reduced to match those from the Carpathians. Furthermore, there were several other species that showed a more moderate increase in the Alps-Carpathians distinction.

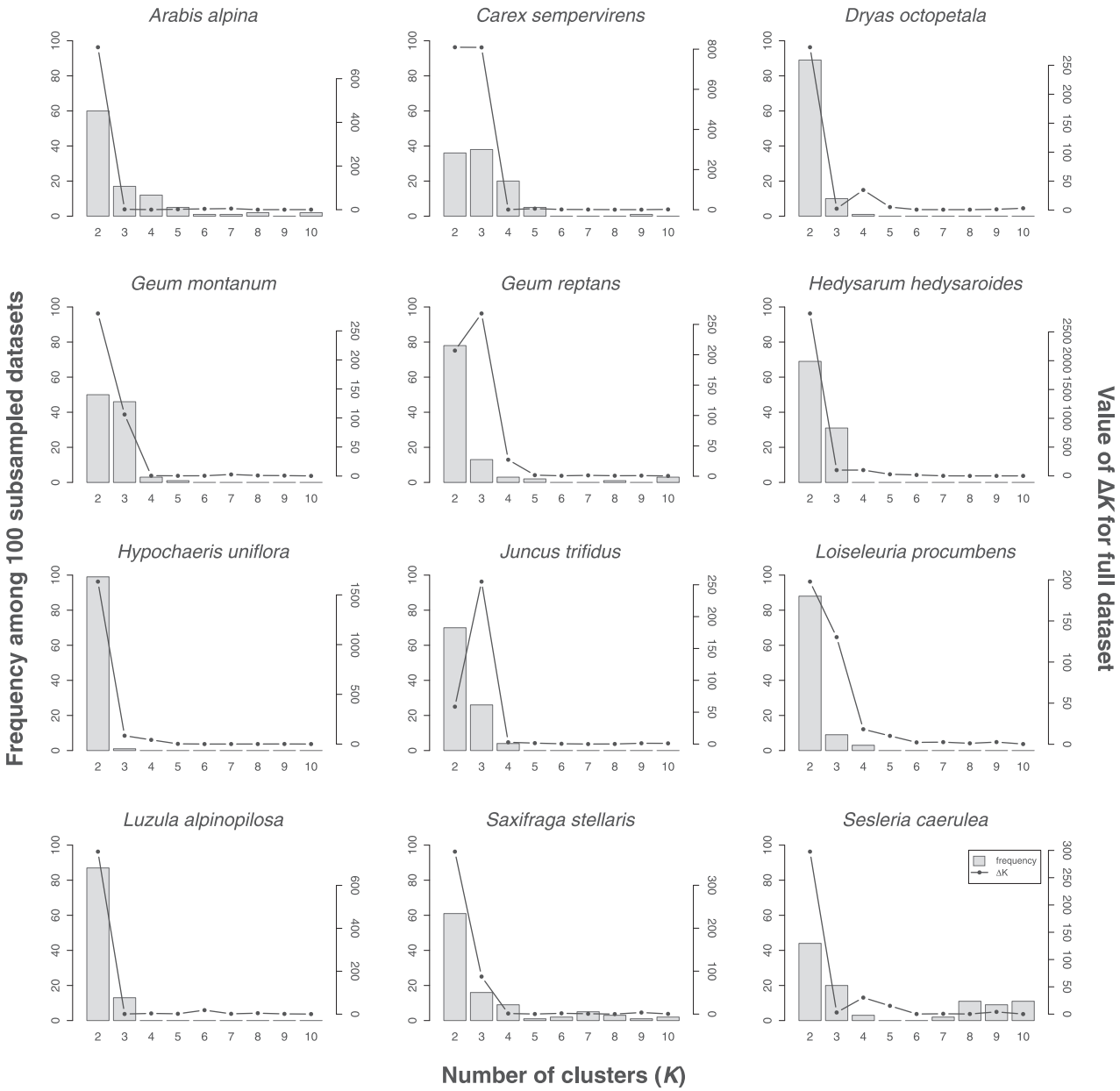

Fig. 5 Inference of the optimal number of clusters according to the ΔK-statistic (Evanno et al. 2005). For each of the 12 species, the red line shows the value of ΔK for the full data set (secondary *Y*-axis); the histograms show the frequency at which each value of *K* was inferred to be the optimal value among 100 of the 500 subsampled data sets (primary *Y*-axis)

Whereas these simulation studies used codominant markers, my analyses used dominant AFLP markers, indicating that the bias is present with both marker types. The underlying cause of this bias is very hard to tell, as determining that would require a very detailed and mechanistic study of how STRUCTURE works, which is something that cannot be done with the data used here. In any case, it is important for researchers to realise that the results from a STRUCTURE analysis should not be taken at face value, especially when the results do not match an a priori expectation.

The subsampling strategy suggested by Puechmaille (2016) proved very useful for uncovering the bias present in the STRUCTURE result of these four species. However, one drawback of this method is that it requires an a priori assumption of what the actual populations are and which populations are underrepresented in the sampling. Of course, if such information is available at the start of the experiment it would be preferable to try to avoid unbalanced sampling in the first place. In practice, however, this may prove to be difficult as access to sampling sites may be restricted or simply beyond the budget of the study.

Nevertheless, in any case where STRUCTURE gives unexpected or biologically difficult-to-explain results the subsampling strategy should be employed.

When there is an a priori expectation of what the population structure looks like, a STRUCTURE analysis should always be accompanied by a direct test of the population structure, for example using an AMOVA (Excoffier et al. 1992). As could be expected, in the data set used here there was a significant positive correlation (Spearman's $r = 0.69$; $p = 0.013$) between the F_{CT} statistic returned by an AMOVA and the β_{AC} statistic that quantified whether STRUCTURE returned separate clusters for the Alps and the Carpathians. Interestingly, the correlation coefficient was slightly higher (Spearman's $r = 0.72$; $p = 0.008$) with the average value β_{AC} values from the subsampled STRUCTURE analyses then with the full data; suggesting that the subsampled STRUCTURE analyses matches the result of the AMOVA slightly better than the STRUCTURE results from the full data set.

The alternative ancestry model setting suggested by Wang (2017), where a separate admixture parameter (alpha) is inferred for each cluster, had very little effect on the results returned by STRUCTURE. In addition, modifying the alpha setting did not improve, or affect, the result of STRUCTURE with unbalanced sampling. The only notable effect was a slight change in the estimation of the number of clusters: *Geum reptans*, which under the default model had three clusters according to ΔK, showed an optimum of $K = 2$ under the alternative ancestry model (Fig. S3). The small affect of the alternative model is surprising since Wang found that this method was very effective with simulated genetic data with unbalanced sampling, and also with a real genetic data set from human populations. One explanation may be that Wang focused on data sets with multiple populations—so with higher values of K—whereas I focused almost exclusively on $K = 2$. Furthermore, the alternative ancestry model assumes simultaneous divergence of the clusters from a unique ancestral pool, which may simply not be applicable to the plant species studied here.

The subsampling analysis also revealed that in some species there is a lot of variation in STRUCTURE results, depending on which populations are included (Fig. 4). This large variation is clear from the large range in β_{AC} values across replicates, which in some species nearly ranged from the minimum value of zero to the maximum value of one. In addition, there were some strong patterns across the sampling range with some populations showing much higher variation in cluster assignment then others. In some species, most notably *Geum montanum*, the populations with a high variation in assignment corresponded to populations that showed admixture in the analysis of the full data set, nicely illustrating the uncertainty associated with the admixture process. However, this was not the case for all species. For example, in *Hypochaeris uniflora* the populations from the Western Carpathians are highly admixed in the analysis of the full data set, but show a low standard deviation in the assignment of the subsampled data sets. Conversely in *Carex sempervirens* (see also Gugerli et al. 2008), the populations from the Southwestern and South-Central Alps showed a high standard deviation in assignment across the subsampled replicates, but little admixture in the analysis of the full data set. In general, visualising the spatial variation in assignments across replicate subsamples may be insightful for pointing out areas where there may be uncertainty in the assignment. For this, the command line version of STRUCTURE can be used to automate the process.

In addition to variation in assignments across replicates, the subsampling analyses also showed a large amount of variation in the estimates of K. For the full data set, ten of the twelve species showed an optimal value of $K = 2$, according to the ΔK-statistic. This corresponds to the observation of Janes et al. (2017) that ΔK has a strong tendency towards $K = 2$, possibly as it tends to return the highest hierarchical level when there are multiple levels of clustering (Evanno et al. 2005). In contrast with this finding for the full data set, the subsampled data sets showed a range of K-estimates for the subsampled data sets for most species. For two species, *Arabis alpina* and *Sesleria caerulea*, the estimates even spanned the whole tested range from $K = 2$ to $K = 10$. This dependence on the exact sampling used for a STRUCTURE analysis reduces the reproducibility of the results (see also Gilbert et al. 2012): two studies on the same species but with slightly different sampling schemes (even when taken from the same part of the species' range) may show strikingly different results.

Though four species showed a clear distinction between the Alps and the Carpathians after subsampling, the other eight species showed partial overlap of clusters between the two mountain ranges. This indicates that the demographic history of these species is more complex than a simple Alps-Carpathians dichotomy. The STRUCTURE clusters also show many different patterns across species, meaning that there are few generalities in the phylogeography of these species. Using partly the same data, Alvarez et al. (2009) already showed for the Alps that the phylogeographic patterns were strongly dependent on the soil requirements of the species, with species from calcareous soils showing different patterns than species from acidic soils. This was hypothesised to be the result of the different locations of pleistocene refugia containing the different soiltypes. In addition, Meirmans et al. (2011) showed how various ecological and life-history traits differently affected different aspects of the genetic population structure of 27 alpine species. Unfortunately, the IntraBioDiv data set (Gugerli et al. 2008; Taberlet et al. 2012) only has 12 species with

sufficient samples in both the Alps and the Carpathians, so tests of the influence of the ecology and life-history of these species on the large-scale genetic patterns would have limited power with $n = 12$.

The complexity of the demographic history of these species is also apparent when looking at higher values of K (Fig. S2). For some species, shared Alp-Carpathian clusters are no longer present at higher values of K; this is most notably the case for the four species where subsampling drastically changed the STRUCTURE results. This reflects patterns that were present in the simulation results of Puechmaille (2016). Of course, unlike with simulated data, for real data sets as were used here, one can never be sure whether results from the subsampled or from the full data are closer to the true situation. Furthermore, though the data set is only a couple of years old, the number of loci used is relatively low compared to today's standards. Since the simulations of Puechmaille (2016) and Wang (2017) used comparable numbers of loci, it remains to be seen whether substantially larger numbers of loci still lead to bias in the STRUCTURE results. Nevertheless, the point remains that for several species STRUCTURE gave consistently different results when subsampling than with the full data set. From a statistical point of view these results are jarring since one wishes different permutations of the same data to give more-or-less the same results. This is the basis of many time-tried statistical approaches such as bootstrapping, jackknifing, and separating data sets into a training set and a validation set. AMOVA's F_{CT} statistic performed much better in this respect as the values for the subsampled data sets were generally nicely centred around the value for the full data sets.

One of the major limitations of STRUCTURE is that it does not take the coordinates of the sampling locations directly into account while clustering. Since in this study the a priori expectation of separation between the Alps and Carpathians is distinctly spatial, there is the possibility that the inclusion of the spatial data could counteract any of the effects of unbalanced sampling is this case. Multiple methods have been developed that explicitly use the spatial data in analyses of population structure (e.g., Dupanloup et al. 2002; Corander et al. 2003; François et al. 2006), and it would be interesting to test whether these methods show similarly biased results as STRUCTURE for this set of species. However, doing this requires a considerable extra amount of calculation and is therefore outside of the scope of the current study.

Recommendations

For one-third of the 12 included species, I found that subsampling the populations from the Alps drastically changed the results of the STRUCTURE analysis. This confirms previous results with simulated data sets (Puechmaille 2016; Wang 2017) that STRUCTURE has difficulties uncovering the true population structure when sampling is unbalanced. To detect such a bias, it is recommended to use the subsampling approach originally suggested by Puechmaille (2016) and expanded upon here. Unfortunately, this method is only applicable when there is an a priori expectation of the population structure that can be used as a basis for the subsampling. Based on the results presented here, using the alternative ancestry model suggested by Wang (2017) is not recommended, as it did not lead to a visible change in the results. This is unfortunate as the alternative ancestry model is much simpler to implement than the subsampling approach and can be applied without any a priori expectation. The results presented here do not mean that the use of STRUCTURE should be discarded: there is abundant evidence that STRUCTURE can return highly insightful results. However, it does mean that STRUCTURE does have its limitations and its results should never be taken at face value. Therefore, the most important recommendation is to always interpret the results with great scrutiny and in the light of available ecological, demographic, and life-history information about the species (Meirmans 2015). Visual inspection of the STRUCTURE results, and comparison with the spurious patterns shown in the paper by Puechmaille (2016) may also be of great aid in this. In fact, it was such visual inspection of the STRUCTURE results for these 12 alpine species that eventually lead to the production of this paper.

Data accessibility

The data used are a slightly extended version of the data present in Dryad packages: https://doi.org/10.5061/dryad.f3rk4 and https://doi.org/10.5061/dryad.s4q6s. The used R-scripts, input files, results files, and associated data can be found in Dryad package: https://doi.org/10.5061/dryad.nh4366s.

Acknowledgements I would like to thank Felix Gugerli, Pierre Taberlet and Michal Ronikier for answering all my questions about the IntraBioDiv data set. I would also like to thank the students of the course "Spatial Processes in Ecology & Evolution" for discussions on genetic data analysis in general, and the analysis of these data sets in particular. Three anonymous reviewers and the associate editor gave valuable comments that helped improve the paper.

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

References

Alvarez N, Thiel-Egenter C, Tribsch A, Holderegger R, Manel S, Schönswetter P et al. (2009) History or ecology? Substrate type

as a major driver of spatial genetic structure in Alpine plants. Ecol Lett 12:632–640

Corander J, Waldmann P, Sillanpaa M (2003) Bayesian analysis of genetic differentiation between populations. Genetics 163:367–374

Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. Mol Ecol 11:2571–2581

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

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial-DNA restriction data. Genetics 131:479–491

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

François O, Ancelet S, Guillot G (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. Genetics 174:805–816

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? J Appl Ecol 46:493–505

Freville H, Justy F, Olivieri I (2001) Comparative allozyme and microsatellite population structure in a narrow endemic plant species, *Centaurea corymbosa* Pourret (Asteraceae). Mol Ecol 10:879–889

Gilbert KJ, Andrew RL, Bock DG, Franklin MT, Kane NC, Moore J-S et al. (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program structure. Mol Ecol 21:4925–4930

Gugerli F, Englisch T, Niklfeld H, Tribsch A, Mirek Z, Ronikier M et al. (2008) Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation – a project synopsis. Perspect Plant Ecol Evol Syst 10:259–281

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

Janes JK, Miller JM, Dupuis JR, Malenfant RM, Gorrell JC, Cullingham CI et al. (2017) The K=2 conundrum. Mol Ecol 26:3594–3602

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

Kamvar ZN, Tabima JF, Grünwald NJ (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2:e281

Lee C-R, Mitchell-Olds T (2011) Quantifying effects of environmental and geographical factors on patterns of genetic differentiation. Mol Ecol 20:4631–4642

Luijten SH, Schidlo NS, Meirmans PG, de Jong TJ (2015) Hybridisation and introgression between *Brassica napus* and *B. rapa* in the Netherlands. Plant Biol 17:262–267

Meirmans PG (2012) The trouble with isolation by distance. Mol Ecol 21:2839–2846

Meirmans PG (2015) Seven common mistakes in population genetics and how to avoid them. Mol Ecol 24:3223–3231

Meirmans PG, Godbout J, Lamothe M, Thompson SL, Isabel N (2017) History rather than hybridization determines population structure and adaptation in *Populus balsamifera*. J Evol Biol 26:229

Meirmans PG, Goudet J, Gaggiotti OE (2011) Ecology and life history affect different aspects of the population structure of 27 high-alpine plants. Mol Ecol 20:3144–3155

Nadeau S, Meirmans PG, Aitken SN, Ritland K, Isabel N (2016) The challenge of separating signatures of local adaptation from those of isolation by distance and colonization history: the case of two white pines. Ecol Evol 6:8649–8664

Neophytou C (2013) Bayesian clustering analyses for genetic assignment and study of hybridization in oaks: effects of asymmetric phylogenies and asymmetric sampling schemes. Tree Genet Genomes 10:273–285

Orsini L, Mergeay J, Vanoverbeke J, De Meester L (2012) The role of selection in driving landscape genomic structure of the waterflea *Daphnia magna*. Mol Ecol 22:583–601

Peijnenburg KTCA, Goetze E (2013) High evolutionary potential of marine zooplankton. Ecol Evol 3:2765–2781

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

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. Mol Ecol Resour 16:608–627

Taberlet P, Zimmermann NE, Englisch T, Tribsch A, Holderegger R, Alvarez N et al. (2012) Genetic diversity in widespread species is not congruent with species richness in alpine plant communities. Ecol Lett 15:1439–1448

Tutin TG, Heywood VH, Burges NA, Moore DM, Valentine DH, Walters SM, Webb DA (eds.) (1980) Flora Europaea, vols. 1–5. Cambridge Univ. Press, Cambridge, 1964–1980

Vos P, Hogers R, Bleeker M, Reijans M, Vandelee T, Hornes M et al. (1995) AFLP: a new technique for DNA-fingerprinting. Nucleic Acids Res 23:4407–4414

Wang J (2017) The computer program structure for assigning individuals to populations: easy to use but easier to misuse. Mol Ecol Resour 14:2611