Molecular Ecology (2005) **14**, 2611–2620 doi: 10.1111/j.1365-294X.2005.02553.x

Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study

G. EVANNO, S. REGNAUT and J. GOUDET
Department of Ecology and Evolution, Biology building, University of Lausanne, CH 1015 Lausanne, Switzerland

Abstract

The identification of genetically homogeneous groups of individuals is a long standing issue in population genetics. A recent Bayesian algorithm implemented in the software STRUCTURE allows the identification of such groups. However, the ability of this algorithm to detect the true number of clusters (*K*) in a sample of individuals when patterns of dispersal among populations are not homogeneous has not been tested. The goal of this study is to carry out such tests, using various dispersal scenarios from data generated with an individual-based model. We found that in most cases the estimated 'log probability of data' does not provide a correct estimation of the number of clusters, *K*. However, using an ad hoc statistic Δ*K* based on the rate of change in the log probability of data between successive *K* values, we found that STRUCTURE accurately detects the uppermost hierarchical level of structure for the scenarios we tested. As might be expected, the results are sensitive to the type of genetic marker used (AFLP vs. microsatellite), the number of loci scored, the number of populations sampled, and the number of individuals typed in each sample.

Keywords: AFLP, hierarchical structure, microsatellite, simulations, STRUCTURE software

Received 5 October 2004; revision accepted 17 February 2005

Introduction

Population genetics deals with the variations of allele frequencies between and within populations. The most widely used measures of population structure are Wright's *F* statistics (Wright 1931). To calculate these indices, one needs first to define groups of individuals and then to use their genotypes to compute variance in allele frequencies. Thus, a fundamental prerequisite of any inference on the genetic structure of populations is the definition of populations themselves. Population determination is usually based upon geographical origin of samples or phenotypes. However, the genetic structure of populations is not always reflected in the geographical proximity of individuals. Populations that are not discretely distributed can nevertheless be genetically structured, due to unidentified barriers to gene flow. In addition, groups of individuals with different geographical locations, behavioural patterns or phenotypes are not necessarily genetically differentiated (for instance, migratory bats from the same breeding roost could be sampled thousands of kilometres apart in winter, see, e.g. Petit *et al*. 2001).

Among the methods not assuming predefined structure, tree-based methods use genetic distance between individuals and tree construction algorithms such as UPGMA or neighbour joining to group them in clusters (e.g. Saitou & Nei 1987). Similarly, multivariate analyses such as multidimensional scaling can help in identifying clusters of individuals. However, these graphical methods are only loosely connected to statistical procedures allowing the identification of homogeneous clusters of individuals.

An alternative model-based method developed recently by Pritchard *et al*. (2000) and implemented in the software STRUCTURE aims at delineating clusters of individuals on the basis of their genotypes at multiple loci using a Bayesian approach. The model accounts for the presence of Hardy–Weinberg or linkage disequilibrium by introducing population structure and attempts to find population groupings that (as far as possible) are not in disequilibrium (Pritchard *et al*. 2000). The estimated log probability of data $\Pr(X \mid K)$ (equation 12 in Pritchard *et al*. 2000) for each value of *K* is given, allowing the estimation of the more likely number of clusters. A quantification of how likely each individual

Correspondence: Jérôme Goudet, Fax: + 41 21 692 42 65;
E-mail: Jerome.goudet@unil.ch

is to belong to each group is also given, information that can be then used to assign individuals to populations. While the authors warn that Pr($X \mid K$) is really only an indication of the number of clusters and an ad hoc guide (p. 949 in Pritchard *et al*. 2000; p. 3 in Pritchard & Wen 2003), the program has been widely used to this end. More generally, it has been used for detection of genetic structure in sample populations for medical purposes (Pritchard & Donnelly 2001; Satten *et al*. 2001), assignment studies (Rosenberg *et al*. 2001), population admixture and hybridization analysis (Beaumont *et al*. 2001; Goossens *et al*. 2002; Randi & Lucchini 2002), migration and dispersal analysis (Arnaud *et al*. 2003; Cegelski *et al*. 2003; Berry *et al*. 2004) and also to detect, with or without success, cryptic genetic structure of natural populations (Rosenberg *et al*. 2002; Caizergues *et al*. 2003). Among the Bayesian clustering methods, STRUCTURE is the most widely used. While other methods have been developed (Banks & Eichert 2000; Dawson & Belkhir 2001; Corander *et al*. 2003) and still other methods for the assignment of individuals to populations exist (but imply the a priori knowledge of source populations: Paetkau *et al*. 1995; Rannala & Mountain 1997; Cornuet *et al*. 1999), we will focus here exclusively on the software STRUCTURE.

Tests and comparative studies using empirical data sets have been performed to assess STRUCTURE's ability in assigning individuals to their known cluster of origin (Pritchard & Donnelly 2001; Rosenberg *et al*. 2001; Manel *et al*. 2002; Turakulov & Easteal 2003). Most of these studies have proven the software to be efficient in assigning individuals to their populations of origin (albeit most are based on simulations with limited number of populations and absence of dispersal between them). However, little is known on the crucial ability of STRUCTURE to detect the real number of clusters (K) which composes a data set. Pritchard *et al*. (2000) showed that STRUCTURE easily detects two to four highly differentiated populations but studies in molecular ecology usually include many more populations and very often these populations are not evenly distributed in space. Many studies have described migration patterns departing from Wright's island model and including several hierarchical levels and/or isolation by distance. For instance, Chapuisat *et al*. (1997), Giles *et al*. (1998), Bouzat & Johnson (2004) or Trouvé *et al*. (2005) have documented situations with a hierarchical pattern of population structure, as groups are themselves clusters of differentiated populations. Another pattern frequently described is a contact zone between otherwise isolated populations. This situation implies a relative genetic isolation between the two groups of populations and sometimes also a pattern of isolation by distance within each group. Such a migration scheme was found for instance by Lugon-Moulin *et al*. (1999) who describe two longitudinal geographical patterns of isolated shrew populations separated by a zone through which dispersal is strongly reduced.

Many of these studies have been conducted using microsatellite markers to assess polymorphism. These DNA markers are widely used because they are both codominant and highly polymorphic (Jarne & Lagoda 1996). However, their development is relatively expensive, time consuming and can be difficult. An alternative family of markers also commonly used in populations studies are the amplified fragment length polymorphism (AFLPs) (Vos *et al*. 1995). AFLPs generate hundreds of polymorphic bands and are easier to develop than microsatellites, but they have the potential inconvenience of being dominant (a DNA band is either present or absent). These two types of markers have different properties. For instance, Gaudeul *et al*. (2004) reported very different levels of population structuring inferred from AFLPs and microsatellite markers. Both AFLP and microsatellites can be used for assignment studies but their respective ability to delineate clusters of individuals has not been compared so far.

The goal of this study is to test the ability of the algorithm underlying the software STRUCTURE to detect the number of clusters in situations including more than two populations. While the program is increasingly used, it is unknown whether it can efficiently detect the real number of clusters in hierarchical systems where migration between populations is uneven. We present an evaluation of the performances of the method under three models of population structure: the island model, a contact zone, and a hierarchical island model. For each model, we simulated AFLP and microsatellite genotypic data sets that were subsequently run in STRUCTURE, and then we analysed the output. We find that ΔK, an ad hoc quantity related to the second order rate of change of the log probability of data with respect to the number of clusters, is a good predictor of the real number of clusters. STRUCTURE identifies groups of individuals corresponding to the uppermost hierarchical level, and performs well with both dominant and codominant markers.

Materials and methods

Simulation of the three migration models

We used the software EASYPOP (Balloux 2001) to generate genotypic data from three different models of population structure: an island model, a hierarchical island model and a contact-zone model (Fig. 1). For all simulations and model of population structure, mutation process followed the K allele model (equal probability of mutations to any allelic state) at a rate of $\mu = 10^{-3}$. The modelled organisms are diploid, hermaphroditic and randomly mating (excluding selfing). Each simulation was run for 10 000 generations to obtain populations at drift, migration and mutation equilibrium. For each model, we generated 10 replicates where each individual genotype was made of 100 microsatellite loci, each with 10 possible allelic states.

Table 1 Parameters of the three migration models

	Number of populations	Number of individuals/ population	Migration rate within set	Migration rate between sets
Island model	5	100	10^{-2}	—
Contact zone	10, 2 sets of 5 pop.	100	10^{-2}	10^{-3}
Hierarchical island model	20, 5 sets of 4 pop.	50	2×10^{-2}	10^{-3}

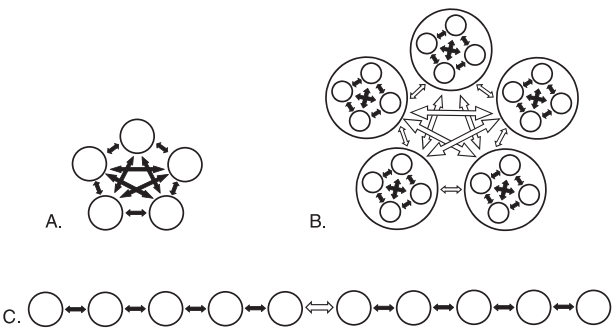

Fig. 1 Schematic representation of the three migration models: (A) Island model. (B) Hierarchical island model. (C) Contact zone. Open arrows represent the migration rates between sets of populations and solid arrows the migration rates within sets (see also Table 1).

The parameters that were varied for the simulations are the number of populations, the number of individuals per population, and the migration rates. These parameters are summarized in Table 1. For the finite island model, five populations of 100 individuals each are exchanging migrants at a rate 0.01. The expected value of F_{ST} for these simulations is 0.15.

The hierarchical island model (Slatkin & Voelm 1991) consists in five sets of four populations, each made of 50 individuals (Fig. 1). Migration occurs at a rate 0.02 within archipelago and 0.001 between archipelagos (Table 1). The expected value of F_{ST} is 0.30 between archipelagos ($F_{\text{Archipelago-Total}}$), 0.16 between islands within archipelagos ($F_{\text{Island-Archipelago}}$), and 0.41 overall ($F_{\text{Island-Total}}$).

The contact zone model is characterized by two sets of five populations, which are organized in a one dimension stepping-stone scheme (Kimura & Weiss 1964). Migration between the two sets occurs through the two central populations at a rate 10 times lower than within each set (Table 1). The expected value of F_{ST} for this model cannot be easily analytically resolved, but global F_{ST} estimated over the 10 replicates (10 times 100 microsatellite loci) is 0.33 and pairwise F_{ST} range from 0.16 to 0.43. The observed value of F_{ST} is 0.17 between the two sets ($F_{\text{Set-Total}}$), 0.25 between populations within sets ($F_{\text{Population-Set}}$), and 0.38 overall ($F_{\text{Population-Total}}$).

EASYPOP generates codominant, microsatellite-like genotypic data. In order to simulate dominant AFLP data, the genotypes generated by EASYPOP were recoded as biallelic loci, in a manner similar to Mariette *et al*. (2002): a randomly chosen half of the microsatellite alleles were coded as '1' and considered dominant while the second half was coded as '2' and considered recessive. Because with dominant data, one cannot distinguish between a dominant homozygote and a heterozygote, dominant phenotypes (obtained from genotypes 1–1 and 1–2/2–1) were recoded as 1–0, where 0 indicates a missing datum. Thus, AFLP data sets bear a proportion of missing data that microsatellite sets do not. This coding of alleles is different from what is recommended in the user's manual of STRUCTURE (Pritchard & Wen 2003), which suggests that dominant markers can be dealt with by coding each phenotype (absence or presence of a band) by a single allele and a missing datum (1–0 for dominant and 2–0 for recessive). We did not use this method because it implies adding a missing value also for recessive homozygotes, which seems unnecessary.

Microsatellite data sets given to STRUCTURE were made of 10 loci as this is a number commonly found in molecular ecology studies. AFLP data sets were made of 100 loci, which seem conservative as AFLP-based studies often include hundreds of markers (Luikart *et al*. 2003). A further reason for this 1:10 ratio of microsatellite loci to AFLP bands comes from a recent simulation-based study (Mariette *et al*. 2002) showing that at least 10 times more AFLP than microsatellite loci are necessary to reach a similar accuracy in the estimation of genetic diversity.

Sampling scheme

To assess the effects of sampling strategies on the method's accuracy, analyses were also carried out on partial data sets. We investigated first the effect of the number of typed loci by sampling only five microsatellites or 50 AFLP bands (Table 2). We also looked at the effect of sampling a subset of individuals from each population (Table 2). Last, for the hierarchical island model, we also looked at the effect of sampling a subset of the populations by randomly omitting one island per archipelago (Table 2). We tested whether partial sampling affected the detection of the true K by comparing results between full and partial data sets.

Table 2 Sampling scheme used for each model. In each situation, all the combinations (full and partial) between the numbers of individuals and loci were tested. For the hierarchical island model the number of populations was also subsampled: 15 out of 20 populations (three populations per archipelago)

	Number of populations		Number of individuals/ population		Number of loci			
	full	partial	full	partial	full		partial	
					AFLP	microsat	AFLP	microsat
Island model	5	—	100	20	100	10	50	5
Contact zone	10	—	100	20	100	10	50	5
Hierarchical island model	20	15	50	20	100	10	50	5

Structure runs

We set most of parameters to their default values as advised in the user's manual of STRUCTURE 2.0 (Pritchard & Wen 2003). Specifically, we chose the admixture model and the option of correlated allele frequencies between populations, as this configuration is considered best by Falush *et al*. (2003) in cases of subtle population structure. Similarly, we let the degree of admixture alpha be inferred from the data. When alpha is close to zero, most individuals are essentially from one population or another, while alpha > 1 means that most individuals are admixed (Falush *et al*. 2003). Lambda, the parameter of the distribution of allelic frequencies, was set to one, as the manual advices. From a pilot study, we found that a length of the burn-in and MCMC (Markov chain Monte Carlo) of 10 000 each was sufficient. Longer burn-in or MCMC did not change significantly the results. As we found that different runs could produce different likelihood values (even with much longer chains, e.g. 1 000 000), for each data set 20 runs were carried out in order to quantify the amount of variation of the likelihood for each *K*. The range of possible *K*s we tested was from 1 or 2 to the true number of populations plus 3.

Statistics used to select K

The model choice criterion implemented in STRUCTURE to detect the true *K* is an estimate of the posterior probability of the data for a given *K*, $\Pr(X \mid K)$ (Pritchard *et al*. 2000). This value, called 'Ln P(D)' in STRUCTURE output, is obtained by first computing the log likelihood of the data at each step of the MCMC. Then the average of these values is computed and half their variance is subtracted to the mean. This gives 'Ln P(D)', the model choice criterion to which we refer as $L(K)$ afterwards. True number of populations (K) is often identified using the maximal value of $L(K)$ returned by STRUCTURE (Zeisset & Beebee 2001; Ciofi *et al*. 2002; Vernesi *et al*. 2003; Hampton *et al*. 2004). However, we observed in our simulations that in most cases, once the real K is reached, $L(K)$ at larger Ks plateaus or continues increasing slightly (a phenomenon mentioned in the STRUCTURE's manual, Pritchard & Wen 2003) and the variance between runs increases (Fig. 2A).

The distribution of $L(K)$ did not show a clear mode for the true K, but we found that an ad hoc quantity based on the second order rate of change of the likelihood function with respect to K (ΔK) did show a clear peak at the true value of K. The rational for this ΔK is to make salient the break in slope of the distribution of $L(K)$ at the true K. It is best explained graphically, as is shown on Fig. 2. First, we plotted the mean likelihood $L(K)$ over 20 runs for each K (Fig. 2A). Second, we plotted the mean difference between successive likelihood values of K, $L'(K) = L(K) - L(K-1)$ (Fig. 2B). This difference corresponds to the rate of change of the likelihood function with respect to K, and is noted $L'(K)$. In a third step we plotted the (absolute value of the) difference between successive values of $L'(K)$, $|L''(K)| = |L'(K+1) - L'(K)|$ (Fig. 2C). This corresponds to the second order rate of change of $L(K)$ with respect to K. Finally, we estimated ΔK as the mean of the absolute values of $L''(K)$ averaged over 20 runs divided by the standard deviation of $L(K)$, $\Delta K = m(|L''(K)|)/s[L(K)]$, which expands to $\Delta K = m(|L(K+1) - 2L(K) + L(K-1)|)/s[L(K)]$ (Fig. 2D). We divided $m(|L''(K)|)$ by $s[L(K)]$ because we found a clear and general trend toward an increase of the variance of $L(K)$ between runs as K increased. We found the modal value of the distribution of ΔK to be located at the real K. We used the height of this modal value as an indicator of the strength of the signal detected by STRUCTURE.

Results

Overall simulation scenarios, we seldom found a mode of the likelihood distribution $L(K)$ at the real K (Fig. 3). In most cases, the likelihood increased until the real K was reached, and then leveled off (often still increasing after the

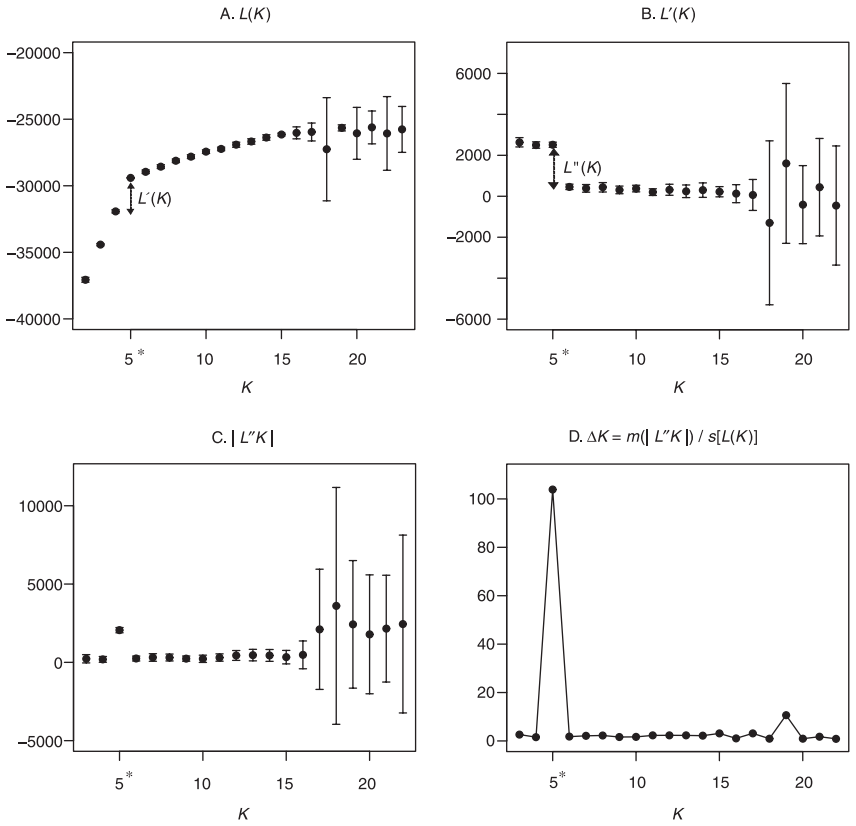

Fig. 2 Description of the four steps for the graphical method allowing detection of the true number of groups K^*. (A) Mean L(K) (± SD) over 20 runs for each K value. The model considered here is a hierarchical island model using all 100 individuals per population and 50 AFLP loci. (B) Rate of change of the likelihood distribution (mean ± SD) calculated as $L'(K) = L(K) - L(K-1)$. (C) Absolute values of the second order rate of change of the likelihood distribution (mean ± SD) calculated according to the formula: $|L''(K)| = |L'(K+1) - L'(K)|$. (D) ΔK calculated as $\Delta K = m|L''(K)| / s[L(K)]$. The modal value of this distribution is the true K(*) or the uppermost level of structure, here five clusters.

real K, Fig. 3). On the other hand, the distribution of ΔK almost always showed a mode at the real K (Fig. 4).

For all three models, and both in full or partial configurations, STRUCTURE identified a number of groups corresponding to the uppermost hierarchical level of genetic partitioning between populations. STRUCTURE primarily highlights the between-sets of populations level for the hierarchical island model and the contact zone, and the between populations level for the island model. Importantly, these results were obtained by using the modal value of ΔK rather than the maximum value of $L(K)$ (Fig. 2A, D). In Fig. 4, the magnitude of ΔK is plotted for each model and sampling scheme, which allows the comparison of results obtained with different parameters sets. Overall, there was some variance among likelihood values $L(K)$ for the different replicates of the same parameter set, but for 29 out of 32 models, all replicates had the same modal value for ΔK.

Island model

For the full data set, as well as for the partial samplings, the modal value of ΔK was $K = 5$, the true number of populations (Fig. 4A, B). The only situations in which STRUCTURE failed to detect the real K were the partial samplings of 20 individuals and five microsatellite markers as well as 20 individuals and 50 AFLPs markers (Fig. 4A, B). For the case with microsatellites which failed to work, we did not see any plateau nor a clear maximum in the likelihood distribution of K for any of the 10 replicates, and the software found a maximal likelihood value at $K = 5$ in 2 replicates, at $K = 2$ twice, at $K = 3$ four times and at $K = 4$ twice. For the case where the true K was not detected by AFLPs, although most replicates had a distribution of $L(K)$ with a break in slope at $K = 5$ followed by a plateau, this pattern was not strong enough to be translated in a high ΔK.

There is a stronger effect of the partial sampling of individuals and loci for microsatellites than AFLP markers (Fig. 4A, B). For the complete data sets, microsatellites seem to perform better than AFLPs markers (the peak is higher) whereas for partial sampling, the results are similar for both types of marker (Fig. 4A, B).

Hierarchical island model

For this model and under exhaustive sampling, the highest likelihood was observed for $K = 11$ for AFLP (Fig. 3C) and $K = 12$ for microsatellites (Fig. 3D) but the modal value of ΔK was at $K = 5$, which corresponds to the number of archipelagos. Using ΔK, we observed that STRUCTURE always found the modal value to be $K = 5$ when all populations were sampled (Fig. 4C, D). When we omitted one island in each of the archipelagoes there was only one case of partial

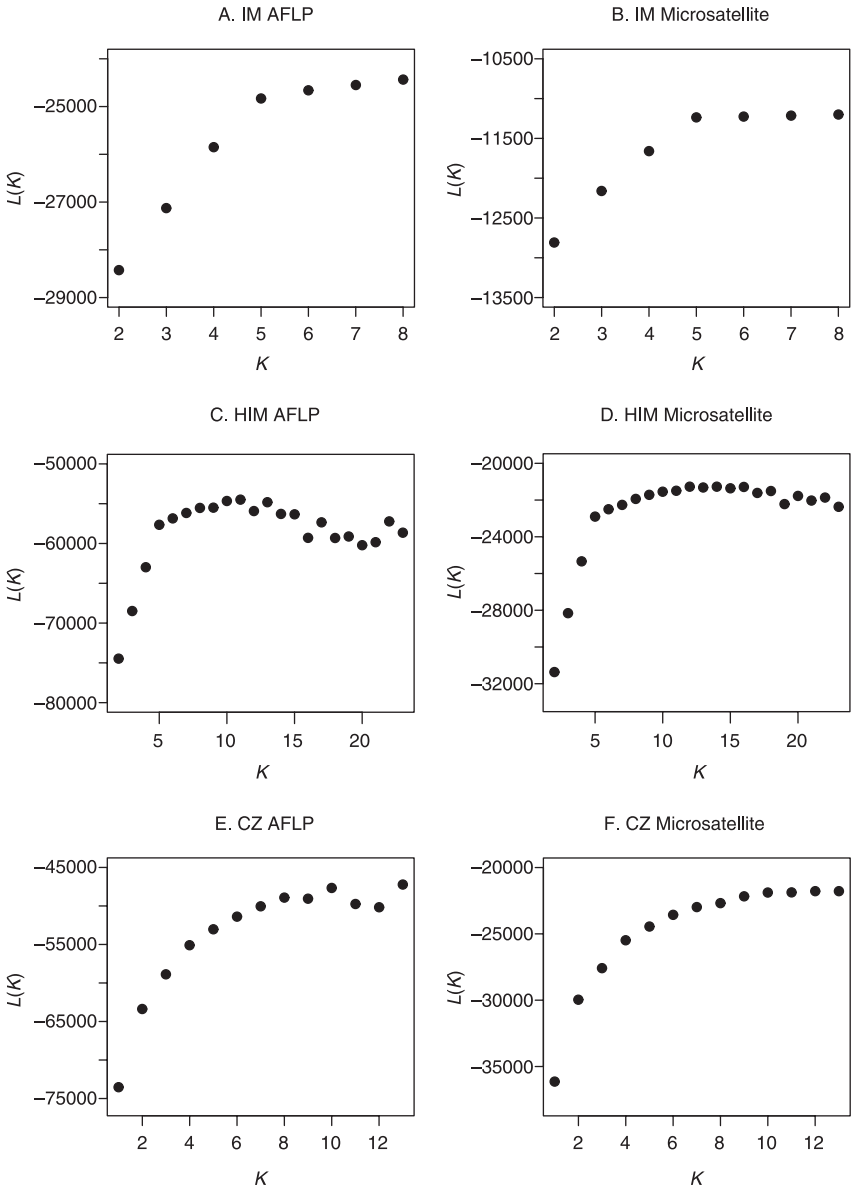

Fig. 3 Log probability of data $L(K)$ as a function of K for the three migration models under exhaustive sampling (averaged over the 10 replicates). Results are shown for AFLPs (panel A, C and E) and microsatellites (panels B, D and F). Panels A and B: island model (IM). Panels C and D: hierarchical island model (HIM). Panels E and F: contact zone (CZ).

sampling where ΔK was not maximum at $K = 5$ (Fig. 4F). This was for the sampling of 20 individuals and 5 microsatellite loci and was actually due to 3 out of 10 replicates. In these replicates, one of the 20 runs of the MCMC gave extremely small $L(K)$ values for $K = 20$ and $K = 21$, which brought the mean of $L(K)$ down. As ΔK is an absolute value of a second order rate of change, it is sensitive to this type of behaviour. However, in real situations, it would have been obvious that these runs did not converge and should be removed (Fig. 4F), in which case the mode at $K = 21$ disappears (data not shown).

For AFLPs, the height of the modal value increased with the intensity of sampling and the number of loci typed, as expected. For microsatellites, the situation was less clear because in several cases ΔK was higher with partial sampling (Fig. 4D). When comparing AFLPs and microsatellites data sets for the same sampling intensity, we found the height of the modal value to be on average higher for microsatellites than forAFLPs, an indication that the signal was stronger in the former. However, AFLPs performed more regularly than microsatellites (Fig. 4C, E).

In order to detect substructuring within archipelagos, we used the best assignment of individuals to one of the five groups to define five subgroups. Each of this subgroup was subsequently analysed with STRUCTURE to detect number of subgroups in each cluster. We did not apply this method to all the subsets of data but for the three subsets we tested, we always found the modal value of ΔK to be $K = 4$, which corresponds to the number of populations within each subset.

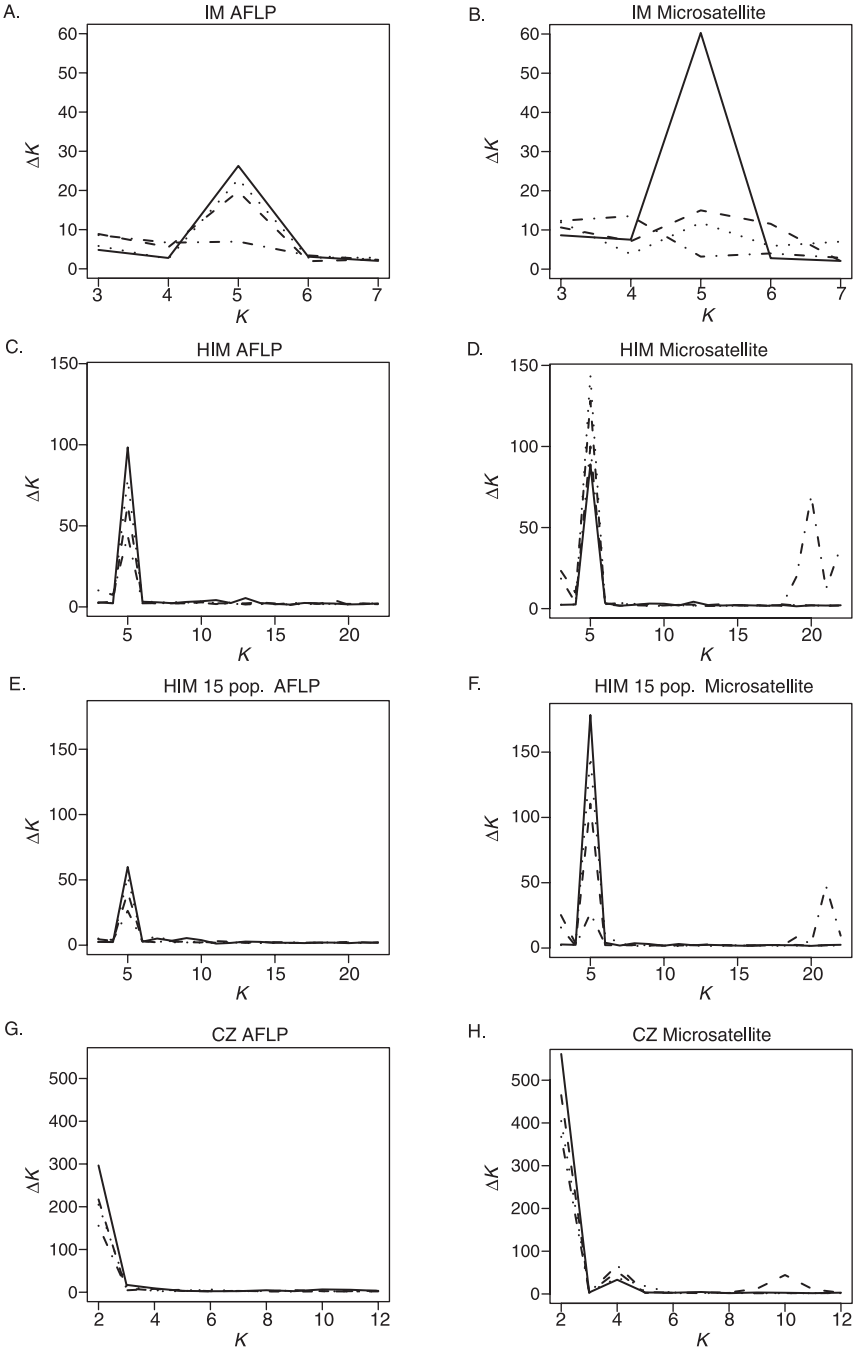

Fig. 4 Magnitude of Δ*K* as a function of *K* (mean ± SD over 10 replicates), calculated for each model using the procedure illustrated in Fig. 2 (A) island model (IM) with AFLP loci; (B) IM with microsatellite loci; (C) Hierarchical island model (HIM) with AFLP loci; (D) HIM with microsatellite loci; (E) HIM with AFLP loci and 15 populations sampled out of 20; (F) HIM with microsatellite loci and 15 populations sampled out of 20; (G) Contact Zone (CZ) with AFLP loci; (H) CZ with microsatellite loci. Solid lines correspond to exhaustive sampling, while dashed, dotted and dotted-dashed lines represent partial sampling. Dashed lines illustrate models with 100 individuals and 50 loci (A, G), 100 individuals and 5 loci (B, H), 50 individuals and 50 loci (C, E) and 50 individuals and 5 loci (D, F). Dotted lines represent cases with 20 individuals and 100 loci (A, C, E, G) and 20 individuals and 10 loci (B, D, F, H). Dotted-dashed lines illustrate models with 20 individuals and 50 loci (A, C, E, G) or 20 individuals and 5 loci (B, D, F, H).

Contact zone

Averaged over replicates (and for exhaustive sampling), the highest likelihood was observed for $K = 13$ (Fig. 3E, F). However, the modal value of Δ*K* was $K = 2$ for all replicates, using either full or partial data sets with 20 individuals out of 100 and 5 microsatellite loci or 50 AFLP loci only (Fig. 4G, H). $K = 2$ corresponds to the uppermost level of structuring in the model, as the 10 demes were partitioned into two sets of five populations by a 'contact zone' of restricted gene flow. Similarly to the hierarchical island model, a division of the data set in two groups corresponding to the best assignment of individual to groups made by STRUCTURE and a subsequent analysis of each subset detected five populations in each subset.

Subsampling of individuals or loci reduced the height of the modal value of Δ*K* (Fig. 4G, H), and 10 AFLPs produced a weaker signal than one microsatellite because the average magnitude of the height of the modal value of Δ*K* was twice lower for the former.

Discussion

Our goal in these simulations was to confront the algorithm underlying the program STRUCTURE with populations organized less simply than the standard island model. We emphasize here that our purpose was not to test the quality of the assignment of individuals to groups, as this has been done (for simpler population structure) by others (e.g. Rosenberg *et al*. 2001; Manel *et al*. 2002). We showed that while $L(K)$, the (ad hoc) estimate for the number of groups given by STRUCTURE often does not correspond to the real number, ΔK, another ad hoc quantity based on the second order rate of change of the likelihood function with respect to K, has a mode at the true K for most of the situations investigated. When the mode of ΔK at the true K was absent, it was either because sample size and marker number was small, leading to an absence of signal, or visual inspection of the values of $L(K)$ would have identified runs of the MCMC with outlying values for $L(K)$. We further found that the algorithm underlying STRUCTURE detects the uppermost level of population structure, and that subgroups created by the best individual assignment produced by STRUCTURE permits to identify sublevels of structuring. We restricted our simulations to cases of moderate to strong structure at different hierarchical levels because our goal was to test the ability of the algorithm to detect the number of groups of individuals in situations when different layers of population structure exist, as is often the case in real situations. Limited simulations for the hierarchical island model with a higher migration rate still detected the (correct) number of archipelagos. This was the case for 10 microsatellites or 100 AFLPs bands with migration rates equal to 0.004 between archipelagos ($F_{\text{Archipelago-Total}} = 0.17$) and 0.02 within archipelagos ($F_{\text{Island-Archipelago}} = 0.14$) and the correct number of archipelagos was still detected with a migration rate of 0.08 between archipelagos ($F_{\text{Archipelago-Total}} = 0.038$) and 0.02 within ($F_{\text{Island-Archipelago}} = 0.035$), but only with the genetic information from 100 AFLPs.

As might be expected, we found that the intensity of sampling both of individuals and markers plays a role in the correct detection of the number of groups. Among the types of markers commonly used for population structure detection, it seems that microsatellites perform slightly better than AFLPs. However, AFLPs gave more regular results in the situations of partial sampling. We note here that the AFLPs coding used (which differs from that advocated by Pritchard & Wen 2003) seems to work quite well despite the presence of numerous and nonrandom missing observations (since the missing allele always comes associated with the dominant), absent from microsatellite data sets.

The quantity ΔK still allows the detection of the real number of groups with five microsatellites or 50 AFLPs. However, for the three models we simulated, the intensity of the signal detected with five microsatellites or 50 AFLP loci was usually lower than when the full set of loci was considered. For the AFLP data sets with 50 loci, the signal was the weakest and thus we suggest a minimum of 100 loci is necessary to insure the detection of the correct number of groups by STRUCTURE. Similarly, partial sampling of individuals led to a lower ΔK at the true K.

In the case of the partial sampling including 15 demes out of 20 in the hierarchical island model (three out of four demes on each island) STRUCTURE still detected a strong signal at $K = 5$ except in one situation of partial sampling. For microsatellites, the height of the modal value of ΔK did not change in comparison with the full hierarchical island model but for AFLPs it decreased by about 50%. While the exhaustive sampling of all potential sources of migrants is crucial if one wants to investigate the comprehensive pattern of migration and structure in an area, our results indicate that the program still works with missing sources, given the level of structure we simulated.

Finally, it must be emphasized that while our simulations provide some indications as to how the STRUCTURE's algorithm reacts to limited sampling, a much more thorough investigation remains to be done. Similarly, the ability of STRUCTURE to detect clusters of individuals at different levels when dispersal among the clusters is more intense is not clear. However, Rosenberg *et al*. (2002) showed empirically on a very large microsatellite data set (377 loci) encompassing 1026 individuals from the five continents that humans cluster in five groups, loosely corresponding to the five continents. They obtain these results despite the notoriously weak genetic differentiation among human populations (F_{ST} among continents around 5%, and lower between populations within continents). Obviously, few nonhuman species could be genotyped with such intensity, but this study indicates that detection of the correct number of clusters can still be found when differentiation is weaker than in our main simulations, and this was confirmed by further limited simulations with F_{ST} among archipelagos as low as 3.8% (see above).

In conclusion, we showed that STRUCTURE is not only able to detect the structure of data sets simulated according to an island model but performs also very well when confronted with more complex hierarchical migration schemes. In such situations, the uppermost hierarchical level of population structure is detected. Subsequent analyses of subsets defined by the best assignment of individuals to groups provided by the program allow finding the hidden within-group structure. Importantly, we showed that the real number of groups is best detected by the modal value of ΔK, a quantity based on the second order rate of change with respect to K of the likelihood function. However, we emphasize that while ΔK helps in identifying the correct number of clusters in most situations, it should not be used exclusively. For instance, ΔK cannot find the best K if $K = 1$. We insist that this criterion is another ad hoc criterion, and

that it should be used together with the battery of other information provided by STRUCTURE: $L(K)$ itself, the value of α and individual assignment patterns (see section 5 in Pritchard & Wen 2003). Last, while STRUCTURE is not profiled to analyse data from dominant markers, our simulations show that AFLPs can give results as accurate as microsatellites.

Acknowledgements

We thank M. Schupbach for his help with the computer partitioning of simulations, and D. Coltman, J. Jaquiéry, N. Juillet, S. Manel, S. Trouvé, an anonymous referee and especially J. Pritchard for helpful comments and advice on the manuscript. J.G. was supported by the Swiss NSF (grants no: 31–59326.99 (PhD of G.E), 31–55945.98 and 31–068325.02).

References

Arnaud JF, Viard F, Delescluse M, Cuguen J (2003) Evidence for gene flow via seed dispersal from crop to wild relatives in *Beta vulgaris* (Chenopodiaceae): consequences for the release of genetically modified crop species with weedy lineages. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **270**, 1565–1571.

Balloux F (2001) EASYPOP (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, 92, 301–302.

Banks MA, Eichert W (2000) WHICHRUN (version 3.2): a computer program for population assignment of individuals based on multilocus genotype data. *Journal of Heredity*, 91, 87–89.

Beaumont M, Barratt EM, Gottelli D *et al.* (2001) Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology*, **10**, 319–336.

Berry O, Tocher MD, Sarre SD (2004) Can assignment tests measure dispersal? *Molecular Ecology*, **13**, 551–561.

Bouzat JL, Johnson K (2004) Genetic structure among closely spaced leks on a peripheral population of lesser prairie-chickens. *Molecular Ecology*, **13**, 499–505.

Caizergues A, Bernard-Laurent A, Brenot JF, Ellison L, Rasplus JY (2003) Population genetic structure of rock ptarmigan *Lagopus mutus* in Northern and Western Europe. *Molecular Ecology*, **12**, 2267–2274.

Cegelski CC, Waits LP, Anderson NJ (2003) Assessing population structure and gene flow in Montana wolverines (*Gulo gulo*) using assignment-based approaches. *Molecular Ecology*, **12**, 2907–2918.

Chapuisat M, Goudet J, Keller L (1997) Microsatellites reveal high population viscosity and limited dispersal in the ant *Formica paralugubris*. *Evolution*, **51**, 475–482.

Ciofi C, Milinkovitch MC, Gibbs JP, Caccone A, Powell JR (2002) Microsatellite analysis of genetic divergence among populations of giant Galápagos tortoises. *Molecular Ecology*, **11**, 2265–2283.

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

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989–2000.

Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, **78**, 59–77.

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

Gaudeul M, Till-Bottraud I, Barjon F, Manel S (2004) Genetic diversity and differentiation in *Eryngium alpinum* L. (Apiaceae): comparison of AFLP and microsatellite markers. *Heredity*, **92**, 508–518.

Giles BE, Lundqvist E, Goudet J (1998) Restricted gene flow and subpopulation differentiation in *Silene dioica*. *Heredity*, **80**, 715–723.

Goossens B, Funk SM, Vidal C *et al.* (2002) Measuring genetic diversity in translocation programmes: principles and application to a chimpanzee release project. *Animal Conservation*, **5**, 225–236.

Hampton JO, Spencer PBS, Alpers DL *et al.* (2004) Molecular techniques, wildlife management and the importance of genetic population structure and dispersal: a case study with feral pigs. *Journal of Applied Ecology*, **41**, 735–743.

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

Kimura M, Weiss GH (1964) Stepping stone model of population structure + decrease of genetic correlation with distance. *Genetics*, **49**, 561.

Lugon-Moulin N, Brünner H, Wyttenbach A, Hausser J, Goudet J (1999) Hierarchical analyses of genetic differentiation in a hybrid zone of *Sorex araneus* (Insectivora: Soricidae). *Molecular Ecology*, **8**, 419–431.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Manel S, Brthier P, Luikart G (2002) Detecting wildlife poaching: identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes. *Conservation Biology*, **16**, 650–659.

Mariette S, Le Corre V, Austerlitz F, Kremer A (2002) Sampling within the genome for measuring within-population diversity: trade-offs between markers. *Molecular Ecology*, **11**, 1145–1156.

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.

Petit E, Balloux F, Goudet J (2001) Sex-biased dispersal in a migratory bat: a characterization using sex-specific demographic parameters. *Evolution*, **55**, 635–640.

Pritchard JK, Donnelly P (2001) Case–control studies of association in structured or admixed populations. *Theoretical Population Biology*, **60**, 227–237.

Pritchard JK, Wen W (2003) *Documentation for STRUCTURE software: Version 2*. Available from http://pritch.bsd.uchicago.edu.

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

Randi E, Lucchini V (2002) Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analyses of microsatellite variation. *Conservation Genetics*, **3**, 31–45.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 9197–9201.

Rosenberg NA, Burke T, Elo K *et al.* (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics*, **159**, 699–713.

Rosenberg NA, Pritchard JK, Weber JL *et al.* (2002) Genetic structure of human populations. *Science*, **298**, 2981–2985.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

Satten GA, Flanders WD, Yang QH (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics*, **68**, 466–477.

Slatkin M, Voelm L (1991) F_{ST} in hierarchical island model. *Genetics*, **127**, 627–629.

Trouvé S, Degen L, Goudet J (2005) Ecological components and evolution of selfing in the freshwater snail *Galba truncatula*. *Journal of Evolutionary Biology*, **18**, 358–370.

Turakulov R, Easteal S (2003) Number of SNPS loci needed to detect population structure. *Human Heredity*, **55**, 37–45.

Vernesi C, Crestanello B, Pecchioli E *et al.* (2003) The genetic impact of demographic decline and reintroduction in the wild boar (*Sus scrofa*): a microsatellite analysis. *Molecular Ecology*, **12**, 585–595.

Vos P, Hogers R, Bleeker M *et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Zeisset I, Beebee TJC (2001) Determination of biogeographical range: an application of molecular phylogeography to the European pool frog *Rana lessonae*. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **268**, 933–938.

Currently a PhD student under the supervision of J. Goudet and E. Castella, G. Evanno is working on community genetics of freshwater snalls. S. Regnaut is involved in conservation genetics research. J. Goudet is Professor in population genetics. His research focuses on theoretical and experimental aspects of mating systems in plants and animals, and more generally, on evolution in structured populations.
