Simulation-based performance testing of the Bayesian clustering program STRUCTURE

Karen K. Martien, Eric Archer, and Barbara L. Taylor
Southwest Fisheries Science Center, 8604 La Jolla Shores Dr., La Jolla, CA 92037, USA

ABSTRACT

Proper management of wildlife species relies on an accurate understanding of population structure. While many methods exist for identifying population structure, most of them have not tested in the range of parameters that are typically of interest to wildlife managers. We used simulated datasets to test the performance of STRUCTURE, a Bayesian clustering method designed to assign individual samples to groups based on their genetic similarity and infer the number of populations represented in a sample. We used STRUCTURE to analyze datasets consisting of one or two populations. We found that in the two-population datasets, the individual assignments produced by STRUCTURE were no better than random. STRUCTURE also performed poorly at inferring the number of populations. We examined two different population ancestry and allele frequency models, and found that, while neither model resulted in acceptable performance, results produced by the default model recommended by STRUCTURE's authors better reflected the actual level of uncertainty in individual assignment. Our results suggest that STRUCTURE is unlikely to be a useful tool in defining management units for large whales.

INTRODUCTION

Managing human-caused mortality in wild populations to achieve acceptable population levels is aided by adequately defining population structure. Two common standards of acceptable levels are: 1) to maintain populations and connectivity to retain evolutionary potential (often called Evolutionary Significant Units (ESUs)) and 2) to maintain sustainable local populations that are called Demographically Independent Populations (DIPs) (for a review see Taylor 2004). Genetic data are often used to discover and delineate such units and many analytical methods have been created to interpret genetic data with respect to population structure. Most methods were created to address evolutionary questions and hence aim to identify ESUs. However, all methods are also commonly used to identify DIPs. A large performance testing exercise was formulated to evaluate the performance of a range of methods across a range of different types of population structure and levels of connectivity relevant to management of wild populations. The exercise is called Testing of Spatial Structure Methods (TOSSM; IWC 2004) and is detailed further below. Here, we test the performance of a commonly used method, called STRUCTURE (Pritchard et al. 2000; Falush et al. 2003), for inferring the number of populations and assigning individuals to populations. We tested the performance of STRUCTURE in both a single population scenario as well as scenarios with two populations that should be recognized as different DIPs. Although even these few results suggest limitations for using STRUCTURE to identify DIPs, general conclusions should be reserved until a wider range of testing is complete.

The program STRUCTURE v 2.2 (Pritchard et al. 2000; Falush et al. 2003) uses a Bayesian algorithm to cluster individuals into groups based on genetic similarity. The algorithm attempts to define groups that are in Hardy Weinberg and linkage equilibrium. The user must specify the number of groups, *K*, to be defined but does not need to stratify data prior to analysis. The method is ideal when such *a priori* stratification is difficult, as would be the case for potentially mixed stocks of whales sampled on migration. By running the analysis for different values of *K* and comparing the log-likelihood of the resulting models, it is possible to use STRUCTURE to infer the number of genetically distinct groups represented in a sample.

STRUCTURE has proven to be a very valuable tool, and is being widely applied in studies of population structure. However, interpretation of STRUCTURE results is often challenging, particularly in a management context. For example, STRUCTURE has played a pivotal role in this year's Implementation Review of the Bering-Chukchi-Beaufort (BCB) Seas stock of bowhead whales (Archer et al. 2007; Givens et al. 2007; Jorde and Schweder 2007; Kitakado et al. 2007; Martien et al. 2007). Yet there remains a lack of consensus among researchers regarding how STRUCTURE is best applied to the bowhead whale data and what conclusions should be drawn from the results.

Interpretation of STRUCTURE results is hampered by the fact that the method has not been subjected to simulation-based performance testing over the range of population structure scenarios that are typically of interest in applied studies. Most tests of performance that have been conducted have either been based on analyses of empirical datasets (Pritchard et al. 2000; Rosenburg et al. 2001; Manel et al. 2002; Turakulov and Easteal 2003) or have used simulated datasets in which populations are more strongly differentiated than in many cases of management concern (Pritchard et al. 2000; Manel et al. 2002; Falush et al. 2003; Evanno et al. 2005, Waples and Gaggiotti 2006). Furthermore, all previous simulation-based assessments of performance have used simulated datasets that conformed to the assumption of genetic and demographic equilibrium, an assumption that is usually violated by real populations.

We used simulation-based performance testing to assess the ability of STRUCTURE to correctly infer population structure under a variety of circumstances. We used two sets of simulated data, those generated for use in the TOSSM project, and those generated for the bowhead Implementation Review. The TOSSM datasets were used to assess the ability of STRUCTURE to correctly infer the number of populations represented in a sample and to determine the accuracy of individual assignments produced by the method. The bowhead datasets were used to examine the behavior of STRUCTURE when it is used to divide a single population that is out of genetic equilibrium into two groups. The bowhead datasets have also been used by Martien et al. (2007) to assess STRUCTURE's ability to infer the number of populations.

METHODS

TOSSM datasets

We used STRUCTURE to infer population structure in datasets generated for the Testing of Spatial Structure Methods (TOSSM) project. TOSSM is a large-scale modeling effort aimed at conducting comparative performance testing of genetic analytical methods (IWC 2004). The TOSSM project has two goals. The first is to assess the performance of methods relative to the types of analyses for which they were originally intended. STRUCTURE was intended primarily as a method for assigning individuals to populations and secondarily as a means for inferring the number of populations represented in a sample. In this paper, we use the TOSSM datasets to assess STRUCTURE's performance for both of these tasks.

The second goal of TOSSM is to adapt methods so that they can be used to define management units and then test their performance in that context. Such management-based performance tests can be conducted using the TOSSM package (described in Gregovich et al. 2007). The package is a collection of functions, written in the language R (R Development Core Team 2006) that allow the user to simulate the management of a population under the Revised Management Procedure used by the IWC. We have begun to test the performance of STRUCTURE using the TOSSM package. We present a description of and preliminary results from these analyses (Appendix) for the purposes of refining the TOSSM exercise. Results are preliminary and should not be used to judge the performance of STRUCTURE as a management tool, but are given to focus discussion on developing performance testing plans for the upcoming year.

The model used to generate the TOSSM datasets is described in Tallmon et al. (2004) and Martien (2006). Simulations were performed using the R package Rmetasim (Strand 2002), which is a collection of functions for conducting individual-based population genetic simulations in the statistical language R (R Development Core Team 2006). Demographic parameters used in the model were based on estimates of vital rates for eastern gray whales (Martien et al. 2004, Martien 2006). Each individual within the model was assigned a 500 bp mitochondrial sequence haplotype and a genotype for 18 microsatellite markers. The parameter controlling the rate of mutation within the model was tuned to produce haplotype and allele frequency distributions comparable to those observed in empirical gray whale datasets ($\mu = 5\text{x}10^{-3}$ for the full mitochondrial sequence; $\mu = 2\text{x}10^{-3}$ for the microsatellite loci).

We examined datasets representing three different population structure scenarios. The first scenario consisted of a single panmictic population with a carrying capacity of 7,500 (this scenario is referred to as Arch1_1 in Martien 2006). The second scenario consisted of two populations, each with carrying capacities of 3,750, and exchanging dispersers at a rate of $5\text{x}10^{-4}$ per year (Arch2_4 in Martien 2006). The third scenario was identical to the second, except that dispersal was increased by an order of magnitude to $5\text{x}10^{-3}$ (Arch2_5 in Martien 2006). Mean differentiation resulting from these dispersal rates is given in Table 1. For each scenario, between 64 and 100 replicate datasets were analyzed. We drew 300 samples from each dataset for analysis. In datasets containing two populations, samples were evenly divided between the two populations. This sample size was chosen to maximize the utility of our analyses for interpreting results of the STRUCTURE analyses of the BCB bowhead dataset, which contained 282 samples (33-locus reference dataset defined at the second Intersessional Workshop; IWC 2007).

Table 1. Dispersal rates and mean differentiation (F_{ST}) between populations for the two-population scenarios. The number of dispersers per year and per generation are absolute number moving, not effective number.

Scenario	Abundance (per population)	Annual dispersal rate	Number of dispersers per year	Number of dispersers per generation	Mean F_{ST}
2	3,750	$5\text{x}10^{-4}$	1.875	37.5	0.0053
3	3,750	$5\text{x}10^{-3}$	18.75	375	0.00048

When running STRUCTURE, the user must make several decisions regarding model choice. First, an ancestry model must be chosen. The ancestry model specifies the degree of genetic isolation that is expected between populations. In the no-admixture model, populations are assumed to be experiencing such little gene flow that individuals can be treated as having descended entirely from one population or another. In contrast, the admixture model allows for the possibility that individuals may have recent ancestors in more than one population, and is therefore appropriate for populations that exchange dispersers at non-trivial rates.

The user must also specify an allele frequency model. The independent model assumes that allele frequencies are completely independent between populations, implying a level of genetic differentiation between populations consistent with that between sub-species. The correlated model assumes that all populations represented in a sample descended from a single ancestral population at some point in the past (Falush et al. 2003). The degree of divergence in their allele frequencies is a function of the rate at which they have experienced genetic drift since divergence, and is an estimable parameter.

We used both the no-admixture/independent model and the admixture/correlated model to analyze the TOSSM datasets. The latter is the program default, and is most appropriate for applied studies, which typically involved populations between which gene flow is non-zero (Falush et al. 2003).

Finally, the user must specify the number of groups (K) STRUCTURE is to define. For each dataset we analyzed, we examined values of K ranging from 1 to 4. Pritchard et al. (2000) described an *ad hoc* procedure by which the log-likelihoods associated with different values of K can be compared in order to infer the number of genetically distinct groups represented in a sample. We used this approach to infer the number of populations in each of the TOSSM datasets. The alternative method of inferring K described by Evanno et al. (2005) was not applied to the TOSSM datasets.

We used two different measures to judge the performance of STRUCTURE. First, we kept track of the frequency with which the correct number of populations was inferred. Second, for those scenarios that consisted of two populations, we examined the probability with which an individual was assigned to the correct population when STRUCTURE was used to define two groups. We calculated three measures of probability of correct assignment. For the first measure, we assigned every individual to the population for which it had the highest assignment probability. Thus, an individual that assignment probabilities of 0.51 and 0.49 for populations one and two, respectively, was assigned to population one. For the latter two measures, we only assigned individuals to populations for which their assignment probability was at least 0.8 or 0.9, respectively.

Bowhead datasets

Archer et al. (2007) used Rmetasim (Strand 2002) to conduct individual-based simulations of the population dynamics and genetics of bowhead whales from the Bering-Chukchi-Beaufort (BCB) Seas region. Their model was structurally very similar to the model used to generate the TOSSM datasets. Both models used the same initialization and burn-in mechanism, though the life history and genetic parameters differed between the two; the bowhead model was tuned to the life history and genetic data currently available for BCB bowhead whales, while the TOSSM datasets were parameterized using data from eastern Pacific gray whales (see Martien et al. 2004 and Martien 2006).

The major innovation of the Archer et al. (2007) model is that it included a harvesting routine that closely mimics the historic harvest of BCB whales. The number of animals killed in each year of the simulation was equal to the number actually taken by whalers. Where such data were available, the age and sex of animals killed in the model was matched to the empirical data. Similarly, the demographic characteristics of simulated animals from which genetic samples were taken were matched as closely as possible to the empirical bowhead genetic dataset.

Archer et al. (2007) used their model to determine whether the genetic heterogeneity that has been observed in BCB bowheads (Givens et al. 2004, Givens et al. 2007, Jorde et al. 2006, Jorde et al. 2007, LeDuc et al. 2007) is consistent with a single population that is out of genetic and demographic equilibrium due to the effects of commercial whaling. They analyzed the simulated datasets using a variety of methods, including STRUCTURE. They used STRUCTURE to cluster the simulated samples into two groups, and compared the patterns of assignment in the simulated datasets to those of the empirical dataset.

We expanded upon Archer et al.'s analysis by testing to see whether the two groups defined by STRUCTURE were in Hardy Weinberg equilibrium (HWE). Recall that STRUCTURE interprets lack of HWE as a signal of population structure and attempts to define groups that are each in HWE. However, in a dataset where departure from HWE is due to something other than population structure, as is the case in the bowhead datasets, it may not be possible to divide the samples into groups that are in HWE. We wished to determine whether the two groups defined by STRUCTURE would exhibit comparable levels of departure from HWE, or if STRUCTURE tends to define one group that conforms to HWE and another that does not. We first tested the groups defined by STRUCTURE to determine how many loci showed significant differentiation from HWE in each group. We refer to the difference between the two groups in this quantity as the observed 'imbalance.' We conducted a single random permutation of individuals between groups and repeated the calculations to generate an expected imbalance for groups that are defined without regard to HWE. We repeated the entire process for each of 100 replicate datasets, resulting in 100 pairs of observed and expected imbalance. We used a one-tailed paired t-test to determine whether the imbalance in the number of loci out of HWE between the two groups defined by STRUCTURE was larger than expected if the groups were defined randomly.

We used Genepop v3.3 (Raymond and Rousset 1995) for all of the tests for HWE. We ran the test for heterozygote deficiency for each locus using an MCMC burn-in of 30,000 iterations and a final chain length of 2,000 iterations in 100 batches. For the groups defined by STRUCTURE, we calculated Hardy-Weinberg disequilibrium across all loci for a given replicate using Fisher's method (Ryman and Jorde 2001).

RESULTS

TOSSM datasets

The number of genetically distinct groups inferred by STRUCTURE was strongly dependent on the ancestry and allele frequency model used (Table 2). With the admixed ancestry/correlated allele frequencies model (admix/corr), STRUCTURE usually indicated the presence of a single population, regardless of how many populations were actually present in the dataset. In contrast, the no-admixture/independent allele frequencies model (noadmix/indep) favored values of *K* larger than one for scenarios 1 and 2, but favored *K* = 1 for scenario 3.

*Table 2. Results of using STRUCTURE to infer the number of populations (*K*) in the TOSSM datasets. The correct value of* K *is given for each scenario (Correct K), followed by the frequency with which different values of* K *were chosen using the method described in Pritchard et al. (2000). 'Admix/corr' refers to a model with admixed ancestry and correlated gene frequencies, while 'Noadmix/indep' refers to a model with no admixture and independent allele frequencies. '# of replicates' indicates the number of simulated datasets that were analyzed for a given scenario and model. The annual dispersal rate between populations was 0.0005 in scenario 2 and 0.005 in scenario 3.*

Scenario	Model	# of replicates	Correct *K*	Chosen *K*			
				1	2	3	4
1	Admix/corr	71	1	1.00	0	0	0
	Noadmix/indep	97	1	0.09	0.52	.25	0.14
2	Admix/corr	99	2	0.68	0.32	0	0
	Noadmix/indep	100	2	0.07	0.62	0.17	0.14
3	Admix/corr	64	2	0.98	0.02	0	0
	Noadmix/indep	100	2	0.77	0.13	0.04	0.06

Errors in assigning individuals to their correct population were large for the scenarios that contained two populations (Table 3). When individuals were assigned to the population for which they had the highest probability, the assignment success was no better than expected by chance, regardless of which STRUCTURE model was used. Restricting assignments to only those individuals that had assignment probabilities greater than 0.8 or 0.9 did not improve assignment success. For scenario 3 with the admix/corr model, using the stricter assignment criteria did appear to affect assignment success. However, this variability simply reflects the fact that in these cases very few individuals (less than 2 individuals per replicate, on average) were assigned to populations under the strict assignment criteria.

Table 3. Probability of assigning individuals to the correct population when STRUCTURE is used to define two groups. 'Highest probability' means that all individuals were assigned to the population for which their assignment probability was highest. '≥80% probability' and '≥90% probability' mean that an individual was only assigned to a population if its assignment probability was at least 0.8 or 0.9, respectively. The numbers in parentheses indicate the average proportion of individuals that were assigned to a population under the restricted assignment criteria.

Scenario	Model	Correct assignment		
		Highest probability	≥80% probability	≥90% probability
2	Admix/corr	0.518	0.540 (0.43)	0.534 (0.22)
	Noadmix/indep	0.524	0.530 (0.81)	0.530 (0.69)
3	Admix/corr	0.500	0.708 (<0.01)	0.236 (<0.01)
	Noadmix/indep	0.523	0.524 (0.97)	0.525 (0.94)

Figure 1 shows the assignment probability of individuals for the two 2-population scenarios and the two STRUCTURE models. The admix/corr model and noadmix/indep models differed with respect to the confidence with which STRUCTURE assigned individuals to populations. When the noadmix/indep model was used, most individuals assigned strongly to one group or the other. For scenario 2, which had a lower dispersal rate, individuals were assigned to the two groups in nearly equal proportions. For scenario 3, however, most individuals assigned strongly to the same group, which is consistent with the fact that STRUCTURE usually favored the definition of a single population for this scenario (Table 2). Note, however, that though individuals assigned strongly to groups under the no-admixture/independent model, they were not assigned correctly; individuals that had assignment probabilities greater than 0.8 were still only assigned to the correct group about half of the time.

Under the admixture/correlated model, assignment probabilities were much closer to 0.5 (Figure 1). There was again a tendency for individuals from scenario 2 to assign more strongly to a group than individuals from scenario 1.

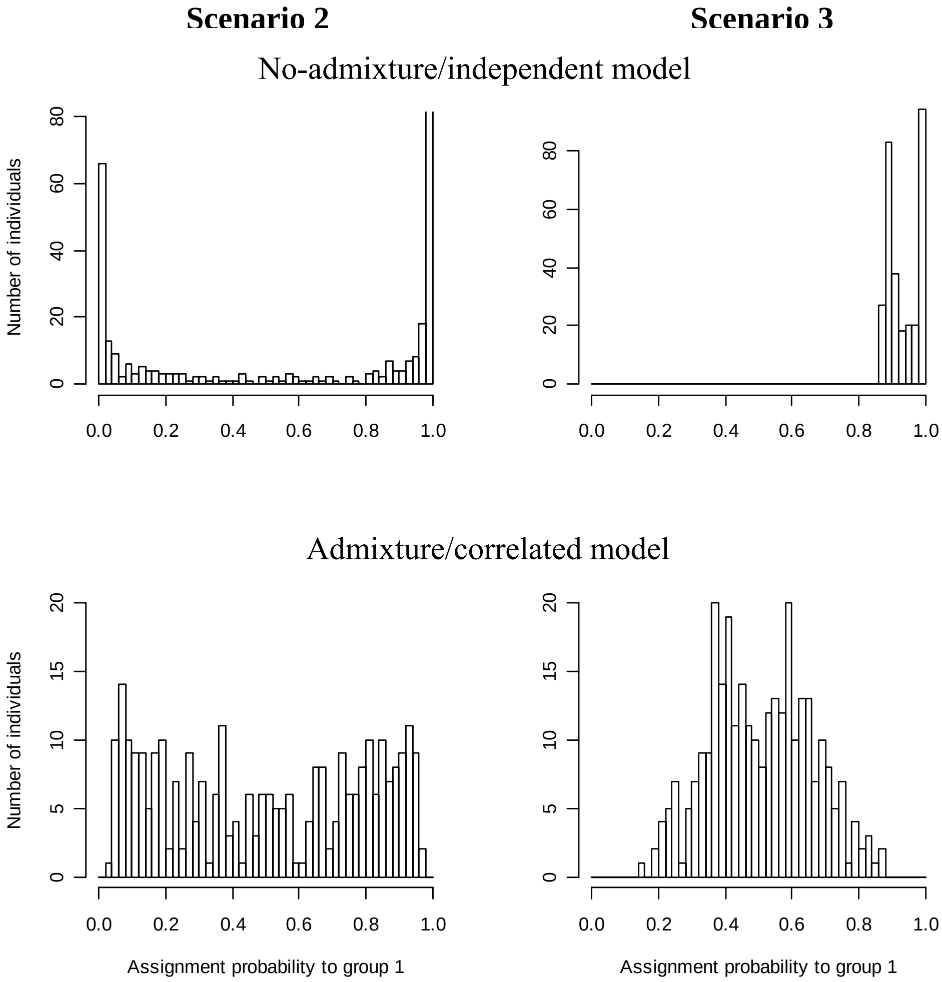

Figure 1. Frequency distribution of assignment probability to group 1. The no-admixture/independent model is shown in the top row, the admixture/correlated model is in the bottom row. Scenario 2 (d = 0.0005) is on the left and scenario 1 (d = 0.005) is on the right. Results shown are from the first replicate dataset for each scenario.

Bowhead datasets

The number of loci that were out of HWE in each group defined by STRUCTURE for the bowhead datasets ranged from 0 to 6, with a median of 1. The difference in the degree of genetic disequilibrium between the two groups defined by STRUCTURE was not significantly greater than would be expected for randomly defined groups (t-test *p*-value = 0.097).

Fourteen replicates exhibited a significant departure from HWE when *p*-values were summarized across loci using Fisher's method (Ryman and Jorde 2001). In each of these cases, only one of the two groups defined by STRUCTURE was out of HWE.

DISCUSSION

We examined three aspects of the performance of STRUCTURE: 1) the accuracy of individual assignments, 2) its ability to infer the number of populations, and 3) its ability to define groups that conform to HWE. We examined two dispersal rates that would both be sufficiently low to warrant separate management of DIPs: d = 0.0005 and 0.005. For both rates, the assignment success of STRUCTURE was no better than expected by random chance. Previous studies have found greater assignment accuracy (Manel et al. 2002; Falush et al. 2003). However, these studies focused on populations that were much more strongly differentiated than any considered in our study. Waples and Gaggiotti (2006) conducted one test of STRUCTURE in which populations exchanged dispersers at a rate comparable that in our scenario 2 and also found that assignment success was no better than random. Populations of large whales typically number in the thousands, meaning that even demographically trivial levels of dispersal are sufficient to prevent them from becoming strongly differentiated. Thus, the magnitude of differentiation for most populations of large whales is likely to be more similar to that in the TOSSM datasets than in those used by Manel et al. (2002) and Falush et al. (2003).

We compared the performance of STRUCTURE under two different ancestry/allele frequency models (no-admixture/independent and admixture/correlated). Though assignment success for the two models was the same, the no-admixture/independent model was much more misleading, in that it often resulted in very high, though incorrect, assignment probabilities for scenario 2 (Figure 1). The admixture/correlated model, on the other hand, tended to result in assignment probabilities closer to 0.5, better reflecting the uncertainty in the assignments.

STRUCTURE performed slightly better with respect to inferring the correct number of populations. The admixture/correlated model performed perfectly when inferring the presence of a single population in the scenario 1 datasets (Table 2). However, it tended to underestimate the number of populations in the two-population scenarios. This bias toward finding too few populations contradicts the warning given by Falush et al. (2003) that this model is 'more permissive of additional populations being fitted to the data set' and therefore may tend to overestimate the number of populations.

The only case in which Falush et al. found the admixture/correlated model to overestimated *K* in their performance tests was when there was a single population with a 50% selfing rate, resulting in very strong departure from Hardy-Weinberg and linkage equilibrium. The bowhead datasets (Archer et al. 2007) also represent a single population that does not conform to HWE, though the degree of disequilibrium in the bowhead datasets is likely much less than that in Falush et al.'s simulated selfing populations. Nonetheless, Martien et al. (2007) found that STRUCTURE accurately inferred the presence of a single population in these datasets 98.7% of the time when the admixture/correlated model was used, suggesting that departures from equilibrium of the magnitude likely to be seen in most large whale populations are unlikely to lead to the overestimation of *K* using this model.

Evanno et al. (2005) also evaluated the probability of correctly inferring *K* using the admixture/correlated model. They found that the method of inferring *K* described in Pritchard et al. (2000) (the same one we applied to our simulated datasets) frequently resulted in an overestimate of *K*. The difference between their results and ours is likely due to the fact that their simulations all involved populations that were very strongly differentiated ($F_{ST} > 0.15$). Consequently, the likelihood of underestimating the number of populations in their simulations was very low. Waples and Gaggiotti (2006) compared the performance of the Evanno et al. method to that described in Pritchard et al. (2000) using simulated datasets in which differentiation was intermediate between those examined by Evanno et al. and those presented in this paper. Waples and Gaggio (2006) found little difference in performance between the two methods of inferring *K*.

The no-admixture/independent model did a much poorer job of inferring the number of populations. It overestimated the number of populations represented by the scenario 1 datasets more than 90% of the time. For scenario 2, the correct value of *K* was chosen for most replicates, though *K* was overestimated 31% of the time. Martien et al. (2007) found a similar bias toward overestimation for this model when they used STRUCTURE to infer the number of populations in the bowhead datasets. These overestimation errors may have been avoided if *K* were inferred using the method proposed by Evanno et al. (2005).

The no-admixture/independent model displayed the opposite bias for the scenario 3 datasets, for which it underestimated *K* in 77% of the replicates. Because the dispersal rate is an order of magnitude higher in scenario 3 than in scenario 2, it is not surprising that STRUCTURE tended to define fewer populations in scenario 3. However, it is unclear why STRUCTURE is more likely to favor *K*=1 for scenario 3, which contained 2 populations, than it is for scenario 1, which only contained a single population.

Finally, our HWE analyses did not support the hypothesis that, when used to analyze data that depart from HWE for reasons other than undetected population structure, STRUCTURE will tend to define groups that differ with respect to HWE. In our simulations, the groups defined by STRUCTURE tended to show a greater 'imbalance' in the number of loci out of HWE than would be expected for randomly defined groups. However, this tendency fell short of statistical significance.

ACKNOWLEDGMENTS

We would like to thank Geof Givens for statistical advice and Dave Gregovich for comments on this paper. Funding for this work was provided by the National Marine Fisheries Service.

REFERENCES

Archer, E., Martien, K.K. And Taylor, B.L. 2007. Use of an individual-based simulation of BCB bowhead whale population dynamics to examine empirical genetic data. Paper SC/59/BRG17 submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, May, 2007. 28pp.

Edwards, C.T.T. and D. S. Butterworth. 2007. Development of a boundary setting algorithm based on migration rates estimated using BayesAss and its preliminary application to TOSSM datasets. Paper SC/59/SD6 submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, May, 2007. 16pp.

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

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567-1587.

Givens, G., Bickham, J.W., Matson, C.W., Ozaksoy, I., Suydam, R.S., and George, J.C. 2004. Examination of Bering-Chukchi-Beaufort Seas bowhead whale stock structure hypotheses using microsatellite data. Paper SC/56/BRG17 submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, July, 2004.

Givens, G., R.M. Huebinger, J.W. Bickham, J.C. George, and R. Suydam. 2007. Patterns of genetic differentiation in bowhead whales (*Balaena mysticetus*) from the western Arctic. Paper SC/59/BRG14 submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, May, 2007. 28pp.

Gregovich, D.P., K.K. Martien, and M.V. Bravington. 2007. A champion's guide to the TOSSM package. Paper SC/59/SD4 submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, May, 2007. 8pp.

IWC. 2004. Report of the Workshop to design simulation-based performance tests for evaluation methods used to infer population structure from genetic data. Journal of Cetacean Research and Management. **6**(Suppl.):469-485.

IWC. 2007. Report of the second intersessional workshop to prepare for the 2007 bowhead whale Implementation Review. Paper SC/59/Rep3 submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, May, 2007. 21pp.

Jorde, P.E., Schweder, T., Bickham, J.W., Givens, G.H., Suydam, R., and Stenseth, N.C. 2006. Detecting genetic structure in migrating bowhead whales off the coast of Barrow, Alaska. In Prep. for Molecular Ecology.

Jorde, P.E. And Schweder, T. 2007. Further analysis of stock structure for BCB bowhead whales using microsatellite DNA data. Paper SC/59/BRG27 submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, May, 2007. 15pp.

Kitakado, T., Pastene, L.A., Goto, M., and Kanda, N. 2007. Updates of stock structure analyses of B-C-B stock of bowhead whales using microsatellites. Paper SC/59/BRG30 submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, May, 2007. 13pp.

LeDuc, R.G. and B.L. Taylor. 2004. Using 'structure' to assess microsatellite loci quality and to address the implication of the suggested populations. SC/56/BRG31.

LeDuc, R.G., K.K. Martien, P.A. Morin, N. Hedrick, K. Robertson, B.L. Taylor, N.S. Mugue, R.G. Borodin, D.A. Zelnnia, and J.C. George. 2007. Mitochondrial genetic variation in bowhead whales in the western Arctic. Paper SC/59/BRG9 submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, May, 2007. 11pp.

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

Martien, K.K. 2006. Progress on TOSSM dataset generation. Paper SC/58/SD2 submitted to the Annual meeting of the Scientific Committee of the International Whaling Commission, June, 2006. 17pp.

Martien, K.K., G.H. Givens, and E. Archer. 2007. A note on the ability of STRUCTURE to correctly infer the number of populations for Bering-Chukchi-Beaufort Seas bowhead whales. Paper SC/59/BRG34 submitted to the Scientific Committee of the International Whaling Commission, May 2007. 8pp.

Martien, KK, Tallmon DA, and Tiedemann R. 2004. Life history matrices for the TOSSM model. Paper SC/56/SD5 submitted to the Scientific Committee of the International Whaling Commission. Sorrento, Italy, June 2004. 4pp.

Pritchard, J.K, M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raymond, M. and F. Rousset. 1995. GENEPOP (version 1.2): population genetics software for exact tests and exumenicism. Journal of Heredity, 85:248-249.

Rosenburg, N.A., T. Burke, K. Elo et al. 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics, 159:699-713.

Ryman, N. and P.E. Jorde. 2001. Statistical power when testing for genetic differentiation. Molecular Ecology 10:2361-2373.

Strand, A. (2002) METASIM 1.0: an individual-based environment for simulating population genetics of complex population dynamics. Molecular Ecology Notes 2(3): 373-376.

Tallmon, D., K.K. Martien, and R. Tiedemann. 2004. Progress in the development and validation of a genetic model for use in the Testing of Spatial Structure Methods (TOSSM) project. Paper SC/56/SD7 SD5 submitted to the Scientific Committee of the International Whaling Commission. Sorrento, Italy, June 2004. 18pp.

Taylor, B. L. 2005. Identifying Units to Conserve. Pages 149-164 In Marine Mammal Research: Conservation Beyond Crisis. J. E. Reynolds III, W. F. Perrin, R. R. Reeves, S. Montgomery, and T. J. Ragen, Eds. The John Hopkins University Press, Baltimore.

Turakulov, R. and S. Easteal. 2003. Number of SNPs loci needed to detect population structure. Human Heredity, 55:37-45.

Appendix: Testing STRUCTURE in the TOSSM context

INTRODUCTION

The TOSSM package (described in Gregovich et al. 2007) is a collection of functions written in the language R (R Development Core Team 2006). These functions allow the user to simulate the management of a population under the Revised Management Procedure (RMP) used by the IWC. The package accepts as input one of the TOSSM datasets. The simulated populations in the dataset are then subjected to a period of un-managed historic harvest (the pre-RMP phase) followed by a period of harvest under the RMP (the RMP phase). An optional recovery phase, during which no harvest occurs, can be included at the end of the simulation. Each year, the simulated populations are projected forward one year using Rmetasim (Strand 2002).

The TOSSM package allows for management units to be defined by a user-provided Boundary Setting Algorithm (BSA). We wrote a BSA in which genetic data sampled from the simulated populations are analyzed by STRUCTURE, and the results of the STRUCTURE analyses are used to define management units. We tested the BSA by using it and the TOSSM package to analyze datasets from two of the scenarios described above (scenarios 1 and 3). While the limited number of analyses we have conducted thus far with the STRUCTURE BSA are insufficient to draw conclusions regarding the performance of STRUCTURE in a management context, they do demonstrate the feasibility of the STRUCTURE BSA and provide some insight into some of the issues that will have to be addressed when using the TOSSM package to conduct performance testing.

METHODS

STRUCTURE BSA

The STRUCTURE BSA uses genetic data passed to it by the TOSSM package to define management units. The user can specify the ancestry and allele frequency models to be used by STRUCTURE, the values of *K* to be evaluated, the length of the burn-in, and the chain length to be used in the analysis. The number of management units defined by the BSA is determined by comparing the posterior probability of the data for different values of *K* and choosing the value of *K* for which this probability is maximized (Pritchard et al. 2000). Each Fully-Integrated Mixed Area (FIMA; see below) is assigned to the management unit for which the average assignment probabilities of all samples from that FIMA is highest. Note that this will not necessarily result in the definition of contiguous management units.

Testing the BSA

To test the STRUCTURE BSA, we used it in conjunction with the TOSSM package to analyze TOSSM datasets for scenarios 1 and 3 described above. We analyzed 50 replicate datasets for each scenario. Scenario 1 consisted of a single population with a carrying capacity of 7,500, while scenario 3 contained two populations, each with carrying capacities of 3,750, that exchange dispersers at an annual rate of 0.005. In the TOSSM package, all harvest and genetic sampling occurs in FIMAs. FIMAs are the smallest spatial unit at which data can be collected and management units can be defined. From a genetic perspective, they represent sampling sites (see Gregovich et al. 2007 for a more detailed description of FIMAs). We defined four FIMAs in our analyses. In the simulations using scenario 3 datasets, FIMAs 1 and 2 consisted entirely of individuals from population 1, while FIMAs 3 and 4 consisted entirely of individuals from population 2.

We simulated a pre-RMP phase that consisted of a single year during which 3,000 animals were killed. The animals killed were taken equally from FIMAs 1 and 2. Thus, in scenario 3, all 3,000 animals were taken from population 1. This resulted in population 1 being reduced to approximately 20% (750/3750) of its pristine abundance. Since scenario 1 contained only a single simulated population, this population was reduced to approximately 60% (4500/7500) of its pristine abundance.

The pre-RMP phase was followed by 100 years of managed harvest. In the first year of the RMP phase, 300 genetic samples were non-destructively sampled from the simulated dataset. For the scenario with two populations, these samples were divided evenly between the two. The STRUCTURE BSA was then used to analyze the genetic samples and define management units. The same management units were used for the entirety of the 100 year RMP phase. The catch-limit algorithm (CLA) was used to calculate quotas in the first year of the RMP phase and every 5 years thereafter. A new abundance estimate was generated each time the CLA was called. We used the package defaults for all other parameters, including the argument mult.CLA=5, which results in the number of animals harvested from the simulated populations being increased five-fold over the quota calculated by the CLA.

We only examined the admixture/correlated model, as that has proven to perform best in our other tests of performance (see above). Because our simulations consisted of only four FIMAs, and a management unit can never be smaller than a FIMA, we evaluated values of *K* ranging from 1 to 4.

For each year of the simulation, the TOSSM package outputs the abundance of the simulated populations and the number of animals harvested. It also outputs the number and location of management unit boundaries defined by the BSA. For each scenario we summarized the average number of management units defined, the average number of whales killed during the RMP phase of the simulation, and the average final depletion of the simulated populations.

RESULTS AND DISCUSSION

The STRUCTURE BSA always resulted in the definition of a single stock, regardless of which scenario was analyzed (Table 1). For scenario 1, average depletion did not change during the RMP phase. Thus, while the population did not recover, it also did not decline. Recall, however, that these analyses were based on multiplying the CLA quota by 5. If this quota multiplier had not been used, the population almost certainly would have grown over the course of the simulation.

In scenario three, the exploited population (population 1) did decline during the RMP phase (Table 1). The variance in depletion was also much higher at the end of the RMP phase than at the beginning, reflecting the fact that in some replicates, the population fared well while in others it was nearly extirpated.

Table 1. Summary performance statistics for analyses of scenarios 1 and 3 using the TOSSM package with the STRUCTURE BSA. Values given are the mean (± s.d.) across replicates.

Scenario	# MUs defined	Initial depletion		Total catch	Final depletion	
		Population 1	Population 2		Population 1	Population 2
1	1 (± 0.0)	0.602 (± 0.010)	na	18,504 (± 574)	0.603 (± 0.038)	na
3	1 (± 0.0)	0.250 (± 0.015)	1.04 (± 0.013)	10,360 (± 707)	0.185 (± 0.137)	1.01 (± 0.018)

While the analyses rpresented in this Appendix are insufficient to draw conclusions about the performance of STRUCTURE in a management context, this exercise has highlighted some of the issues that must be addressed before large scale performance testing of any methods can commence. First priority should be to develop a standardized set of performance trials that should be run for all methods. The parameters that must be specified include sample size, historic catch levels and their allocation to FIMAs, number of FIMAs, the multiplier for the CLA quota (mult.CLA; package default is 5), and the length of the simulation phases.

Because many of the methods to be tested take a long time to run, we must limit the number of performance trials. We will need to be particularly judicious in choosing values for parameters that require additional runs of the BSAs. Since the population projections are relatively fast, parameters that only affect the RMP and post-RMP phases of the simulations can be varied more widely.

The task of defining a standardized set of trials could be made easier if we first determine what sets of parameters have the potential to pose a risk to exploited populations. This could be done by performing a series of analyses in which a single management unit is always defined. This would enable us to determine the parameter combinations under which failure to detect population structure will pose a conservation risk.