# A note on the ability of STRUCTURE to correctly infer the number of populations for Bering-Chukchi-Beaufort Seas bowhead whales

**Karen K. Martien[1], Geof H. Givens[2], and Eric Archer[1]**

[1]Southwest Fisheries Science Center, 8604 La Jolla Shores Blvd., La Jolla, CA 92038 USA.
[2]Dept. of Statistics, 1877 Campus Delivery, Colorado State University, Fort Collins, CO 80523 USA.

## Abstract

Multiple stocks within the Bering-Chukchi-Beaufort region have been inferred from analyses of microsatellite data from bowhead whales killed on migration using the method STRUCTURE (Kitikado et al. 2007). We show that this conclusion is unwarranted and that STRUCTURE analyses are consistent with a single stock. We discuss model choice for STRUCTURE analyses and conclude that, based on bowhead biology, the appropriate model is one that assumes both recent ancestry and some current gene flow (admixture with correlated allele frequencies). Bowhead biology and history is not consistent with the model used by Kitikado et al. (2007) that assumes strong evolutionary separation and no current gene flow (no admixture and independent allele frequencies). We perform two new analyses to facilitate the appropriate selection of the number of populations (*K*) suggested by STRUCTURE analyses. The inference method currently recommended by STRUCTURE's authors strongly suggests one stock within the BCB region. This is consistent with results using individual-based simulations of a single stock modeled to recreate BCB bowhead population dynamics and history (described in Archer et al. 2007): simulated datasets generated from a single stock and analyzed using STRUCTURE in the manner employed by Kitikado et al. falsely favored *K*=2 about 30% of the time. We conclude that in the analysis of the full dataset, including BCB, Okhotsk and Atlantic samples, that STRUCTURE likely makes two errors: 1) it incorrectly fails to identify BCB and Atlantic whales as separate stocks, and 2) it incorrectly identifies two biologically meaningless groups within the pooled BCB-Atlantic samples if selection of *K* is done in a manner not supported by the current literature.

## Introduction

Considerable effort has gone into using the Bayesian clustering method STRUCTURE (Pritchard et al. 2000; Falush et al. 2003) to investigate population structure in Bering-Chukchi-Beaufort Seas (BCB) bowhead whales. Kitakado et al. (2007) have used STRUCTURE to infer the number of genetically distinct groups within BCB bowheads. They and Givens et al. (2007) have both used STRUCTURE to look for patterns of temporal variation in the genetic composition of animals migrating past Barrow in the fall and spring. Finally, Archer et al. (2007) used an individual-based model of BCB bowheads to determine whether the patterns of temporal variation revealed by Givens et al.'s STRUCTURE analyses were consistent with the hypothesis of a single population that is out of genetic equilibrium due to the rapid reduction in population size due to commercial whaling and subsequent recovery of BCB bowhead whales.

However, the appropriate use of STRUCTURE results to infer the number of populations for BCB bowhead whales, given their unique life history and population history, has not yet been addressed. Furthermore, there remains some disagreement regarding which ancestry model (admixture or no admixture) and allele frequency model (correlated or independent) is most appropriate when analyzing the BCB bowhead data, as well as considerable disagreement as to how STRUCTURE results should be interpreted with regard to the number of populations represented. In this paper, we first discuss the issue of model choice (for ancestry and allele frequencies). We then turn to the statistical question of how best to infer the number of populations represented.

## Model choice

STRUCTURE requires the user to make two decisions regarding the model of population structure used. First, the user must specify an ancestry model, which specifies the degree of genetic isolation of modeled populations. The no-admixture model represents populations between which rates of gene flow are so low that individuals can be treated as having descended exclusively from one population or another. The admixture model, on the other hand, is appropriate for populations that have recently or are currently experiencing gene flow at sufficient rates that individuals may have recent ancestors from more than one population.

The second choice the user must make is whether the allele frequencies in the different populations should be treated as independent or correlated. The independence model assumes that the allele frequencies in one population are in no way related to allele frequencies in other populations. This implies that gene flow between the populations is effectively zero, and has been for quite a long time. The correlated model, in contrast, assumes that the populations diverged from a single ancestral population at some point in the past, and the differences in their allele frequencies are the result of drift that has occurred since their divergence. For this model, the degree of correlation between populations is an estimable parameter: populations that have diverged more recently or are experiencing a higher level of ongoing gene flow will have more similar allele frequencies than those that have experienced a greater degree of isolation.

It is important to note that both ancestry and allele frequency models can be appropriate for situations where there may be multiple biological or management stocks. In particular, a search for multiple stocks does not compel the use of a no-admixture/independent frequencies approach. The best model choices depend on the particular biology and history of the individuals studied.

Kitakado et al. (2007) prefer the no-admixture model with independent allele frequencies. From a biological perspective, this model seems inappropriate. The multi-stock hypotheses that have been proposed for BCB bowhead whales assume that the different stocks breed in close proximity to each other and mix during migration. Given their close physical proximity and the highly dynamic environment in which they live, it is difficult to imagine that two BCB stocks could maintain such a high degree of genetic isolation that a no-admixture/independent allele frequencies model would be appropriate. Even very low levels of ongoing gene flow would be sufficient to result in the presence of admixed individuals and considerable correlation in allele frequencies. Furthermore, there are six known historical instances of possible exchange of individuals between the BCB region and the north Atlantic, including two almost certain cases (Bockstoce and Burns, 1993; Tomlin, 1957). It is hardly plausible that two BCB substocks breeding in, say, the Gulf of Anadyr and the Bering Sea, could maintain greater genetic isolation than two groups that breed in different ocean basins. For all these reasons, the admixture/correlated model is the most biologically appropriate choice for the BCB bowhead dataset.

Kitakado et al. (2007) attempted to compare the two models statistically. They applied both the no-admixture/independent model and the admixture/correlated model to two datasets: the 22-locus and 33-locus reference sets (‘REF22’ and ‘REF33’, respectively) defined during the first 2007 AWMP Intersessional Workshop (IWC 2007). For each dataset, Kitakado et al. compare ‘marginal log-likelihoods’ between the two models and conclude that the no-admixture/independent model fits the data best. However, such a comparison is complicated by the fact that the two models incorporate different numbers of parameters.

Though the two ancestry models (admixture and no admixture) contain the same number of parameters, the two allele frequency models do not. For the independent allele frequencies model, a separate parameter is needed to represent the frequency of each allele in each population. Correlated allele frequencies are modeled by assuming that all groups diverged from a single ancestral population at some point in the past, and that the allele frequencies of each group have diverged from the ancestral allele frequencies due to drift. The amount of drift in each group $k$ is determined by the parameter $F_k$. The only parameters required for the correlated allele frequency model are the frequencies of each allele in the ancestral population and the value of $F_k$ for each group.

Thus, the correlated allele frequency model requires 333+$K$ or 404+$K$ parameters for the REF22 or REF33 datasets, respectively, while the independent allele frequency model requires 333*$K$ or 444*$K$ parameters, where $K$ is the user-specified number of clusters. The model with independent allele frequencies that Kitakado et al. prefer therefore contains at least 400-700 more parameters than the correlated model depending on the choice of dataset and $K$. Empirical analysis to compare the two allele frequency models must therefore be based on methods that are not misled by widely differing numbers of parameters.

## Inferring the number of populations

In addition to clustering samples into groups, some users advocate using STRUCTURE to infer the number of populations, $K$, present in a sample. However, carrying out and interpreting the results of such an inference can be challenging. The authors of STRUCTURE warn that their method of inferring $K$ is an *ad hoc* procedure based on ‘dubious’ assumptions and that ‘the inferred value of $K$ may not always have a clear biological interpretation’ (Pritchard et al. 2000).

We examine three issues with respect to inferring the number of populations: (1) the tendency for STRUCTURE to overestimate K in realistic applications, (2) the statistical method used to infer K, and (3) the failure of STRUCTURE to separate known separate stocks. Each of these issues alone presents a compelling case against hypotheses of multiple BCB stocks; together they impose a high standard of proof on anyone who would use STRUCTURE results to support theories of multiple BCB stocks.

## Tendency for STRUCTURE to overestimate *K* in realistic situations

The signal that STRUCTURE looks for to detect multiple populations is lack of Hardy-Weinberg equilibrium (HWE). However, there are many factors besides population structure that can cause samples to be out of HWE. The STRUCTURE user's manual warns that "even in the absence of population structure, these types of factors can lead to a weak statistical signal for *K*>1" (p. 13). For large sample sizes and populations severely out of HWE, the false signal for *K*>1 can be much stronger.

The BCB bowhead whales are expected to be out of genetic equilibrium due to their recent rapid reduction in population size due to commercial whaling and their subsequent recovery. In order to investigate the impact that this unusual population history has on various genetic analyses, Archer et al. (2007) constructed an individual-based model that mimics the life history and population history of BCB whales. They used this model to generate simulated datasets whose demographic characteristics were as closely matched as possible to the empirical REF33 dataset. The simulated datasets were then used to generate null distributions against which empirical values could be compared for a variety of statistical analyses.

One of the analytical methods that Archer et al. examined was STRUCTURE. They found that the groups defined when STRUCTURE was used to cluster the REF33 dataset into two populations were entirely consistent with their model of a single, non-equilibrium population. However, Archer et al. did not use their simulated datasets to investigate the significance of the number of groups inferred by STRUCTURE.

In order to determine whether the genetic disequilibrium present in BCB bowheads could be causing STRUCTURE to erroneously infer the presence of multiple populations, we used STRUCTURE to infer the number of populations in each of Archer et al.'s simulated datasets, using the 'log probability of the data' to choose between *K*=1 and *K*=2 in the same manner done by Kitakado et al. (2007). We used both the no-admixture/independent allele frequencies model and the admixture/correlated allele frequencies model, with the REF33 dataset. Here *K* refers to the number of BCB clusters or subpopulations.

In these simulated datasets where *K*=1 is the right answer, the no-admixture/independent model results favored *K*=2 about 30% of the time (Table 1). For the admixture/correlated model, *K*=1 was favored in 98.7% of the simulated datasets. The values summarized in Table 1 represent null distributions for the number of populations STRUCTURE would infer given under a null hypothesis of a single non-equilibrium population. Thus, the Kitakado et al. (2007) finding of *K* = 2 under the no-admixture/independent model is not unusual under the null distribution for a single stock. The simulation results also point to the greater utility of the admixture/correlated model, which resulted in correct inference of *K* in 98.7% of the simulated datasets, and suggest that the Hardy-Weinberg disequilibrium being detected by the no-admixture/independent model in STRUCTURE is a result of BCB bowheads' history of depletion and recovery rather than undetected population structure.

Table 1. Results of STRUCTURE analyses of simulated REF33 datasets. Cell values indicate the proportion of the simulated datasets in which a given value of *K* resulted in the highest *log Pr[D|M]*. Shaded boxes indicate the value of *K* for which *log Pr[D|M]* was maximized in the analyses of the empirical REF33 datasets.

| Model | *K* = 1 | *K* = 2 |
|---|---|---|
| No-admixture/independent allele frequencies | 0.70 | 0.30 |
| Admixture/correlated allele frequencies | 0.987 | 0.013 |

The STRUCTURE user's manual also provides the following relevant guidance to determining whether or not population structure detected by the program is real: "when there is no population structure, you will typically see that the proportion of the sample assigned to each population is roughly symmetric (~1/*K* in each population), and most individuals will be fairly admixed" (p. 14). Again, this is what we see in the Kitakado et al. (2007) and Givens et al. (2007) STRUCTURE analyses: when forced to divide the BCB samples, STRUCTURE splits them into groups of roughly equal size. This is an indication that the group assignments for these whales represent noise rather than signal.

## Statistical method used to infer K

Kitakado et al. (2007) examine the quantity *log P[D|M]* reported by STRUCTURE, where *D* refers to the data and *M* refers to the model, to infer the number of clusters (*K*). In this context, the model choices include two different choices for the ancestry/allele frequency model (see above), and five choices of *K* for each genetic model. The Kitakado et al. approach is to choose the value of *K* yielding the highest value for *log P[D|M]*. For example, analyzing all samples with the model with no admixture and independent allele frequencies, they report that *K*=4 yields the highest *log P[D|M]*. Kitakado et al. infer from this that "individuals passing by Barrow…probably came from at least two genetically distinct stocks" (p. 1). This approach and the resulting inference by Kitakado et al. are not consistent with best practices advised by the developers of STRUCTURE and by independent researchers who have investigated how best to infer *K*.

Pritchard et al. (2007) advise against simply choosing the value of *K* that maximizes *log P[D|M]*. Referring to *log P[D|M],* they note that often "in real data, the value of our model choice criterion continues to increase with *K*" (p. 12). They advise users to "aim for the smallest value of *K* that captures the major structure in the data" (p. 14) by identifying what they call "more-or-less plateaus" where the value of *log P[D|M]* begins to level off.

A more formal method is developed by Evanno et al. (2005); this approach is also endorsed by Pritchard et al. (2007). In comprehensive simulation testing, Evanno et al. found that "in most cases the estimated 'log probability of data' does not provide a correct estimation of the number of clusters, *K*" (p. 2611). They developed an alternative measure, denoted *ΔK*, which "has a mode at the true *K* for most of the situations investigated" (p. 2618). Indeed, for 29/32 scenarios they examined, the *ΔK* approach chose the correct *K*, while the approach used by Kitakado et al. usually failed with a positive bias for *K*.

The Evanno et al. approach and the advice of Pritchard et al. (2007) are motivated by the understanding that *log P[D|M]* generally increases with *K*. Therefore it is incorrect to choose the model with the highest *log P[D|M]*. Instead, we must look for the "knee" of the plot of *log P[D|M]* against *K*. The "knee" is the point of diminishing returns where the improvement in *log P[D|M]* levels off with increasing *K*. The Evanno approach finds the "knee" by estimating second derivatives. Detection of the "knee" was precisely the reasoning used by Givens et al. (2007) when they inferred that a single BCB stock was most consistent with STRUCTURE results. They examined a larger choice of *K* purely for exploratory purposes.

Here we present the results from Evanno et al.'s *ΔK* for both models examined by Kitakado et al. Like those authors, we analyze the combined samples from BCB, Canada, and the Sea of Okhotsk (the REF22 dataset). A value of *ΔK* is calculated for each *K*. The maximal value of *ΔK* indicates the preferred value of *K*, particularly when it is substantially larger than other choices. In the examples of Evanno et al., the correct *K* was usually identified with a *ΔK* peak value of 50-100, with incorrect *K* choices showing values much closer to zero.

In this analysis, *K* refers to the total number of clusters in the REF22 dataset. Therefore, if BCB animals constitute a single stock, we should find *K*=3 (One for Okhotsk, one for Canada, and one for BCB). If there are two BCB substocks, *K*=4. The *ΔK* statistic cannot evaluate *K*=1 or the maximal choice of *K* explored in analysis, so we ran *K*=1,2,3,4,5 to allow choice between *K*=2, *K*=3, or *K*=4. This was also our motivation for focusing on the REF22 dataset.

Table 2 shows our results. The evidence is incontrovertible: a single BCB stock (*K*=2 overall) is preferred. The two groups identified by STRUCTURE using *K*=2 with either genetic model are (1) the Sea of Okhotsk samples, and (2) all other samples. This indicates that STRUCTURE is unable to identify any clustering beyond a separation of Okhotsk from other samples. There is no evidence for multiple BCB substocks.

Table 1: Estimates of *Log Pr[D|M]* and *ΔK* for different STRUCTURE models. Values of *Log Pr[D|M]* are taken from Kitakado et al.'s (2007) Table 2, and are shifted by 38607 for ease of interpretation. Here, *K* refers to the total number of populations assigned to the combined samples from BCB, Sea of Okhotsk, and Canada. Since STRUCTURE fails to separate Canada and BCB samples, the number of BCB populations in these analyses is *K*-1.

| **Number of BCB Populations** | | **No Admixture/Independent Allele Frequencies** | | **Admixture/Correlated Allele Frequencies** | |
|---|---|---|---|---|---|
| | ***K*** | ***Log Pr[D\|M]*** | ***ΔK*** | ***Log Pr[D\|M]*** | ***ΔK*** |
| **--** | **1** | 0 | NA | 0 | NA |
| **1** | **2** | 717 | 136.1 | 730 | 116.0 |
| **2** | **3** | 1022 | 18.1 | 813 | 1.8 |
| **3** | **4** | 1246 | 5.1 | 850 | 2.6 |
| **4** | **5** | 1028 | NA | 863 | NA |

### Failure of STRUCTURE to separate known stocks

A final piece of advice that the authors of STRUCTURE offer is that we should be skeptical of the results if there is no clear biological interpretation for them. In the case of the analyses of the REF22 dataset, which includes animals from Canada and the Okhotsk Sea, STRUCTURE was able to correctly identify the Okhotsk samples, but could not differentiate between the Canadian and BCB samples. This is also seen in Table 2.

Though the Canadian and BCB stocks may have had substantial gene flow in the distant past, gene flow between them has presumably been quite low in the past 1,000 years. The resulting genetic differentiation was detected in the $F_{ST}$ and $\chi^2$ analyses presented by Givens et al. (2007), but not by STRUCTURE. Instead, STRUCTURE divides the combined BCB/Canadian samples into multiple groups that have no clear biological interpretation: both BCB and Canadian whales are present in equal proportions in each cluster. Consequently, in order to argue that the STRUCTURE results support the existence of multiple populations within the BCB stock, one must explain how it is that those BCB populations maintain their distinctness within the BCB region yet freely interbreed with animals from the Canadian Arctic.

## Conclusion

When using STRUCTURE to make inferences on the number of populations represented in a sample, errors are liable to be made in both directions: genuine, subtle population structure is likely to be missed by STRUCTURE if rates of gene flow between populations are high enough to prevent the development of substantial genetic differentiation, and non-existent populations are likely to be inferred when a population is out of HWE for some reason other than population structure. In the case of the BCB bowhead analyses, both mistakes appear to have occurred. First, STRUCTURE failed to differentiate the BCB whales from the Canadian whales, despite the fact that other analytical methods ($F_{ST}$ and $\chi^2$) identify them as separate populations. Second, the disequilibrium caused by the depletion of bowheads as a result of commercial whaling misled STRUCTURE into erroneously inferring the presence of multiple populations within the combined BCB and Canadian stocks. The simulation performance testing presented by Archer et al. (2007) and in this paper show that the empirical STRUCTURE analysis results are consistent with the existence of a single BCB population that is out of Hardy-Weinberg equilibrium due the effects of commercial whaling. The STRUCTURE results are not consistent with multiple populations within the BCB stock unless one is willing to accept the notion that those populations move freely between the BCB seas and the Canadian Arctic.

Furthermore, it is important to ensure that the output of STRUCTURE is used appropriately if reliable inference about $K$ is desired. The selection of $K$ by strict reference to the peak in $\log P[D|M]$ is noted by the authors of STRUCTURE and by independent researchers as unreliable because it will select too large a value for $K$. These experts all suggest informal or formal detection of the point of diminishing returns, i.e., the point with a large magnitude second derivative of $\log P[D|M]$. Implementing this advice in the manner shown to be highly reliable by Evanno et al. (2005) yields the inescapable conclusion that the STRUCTURE simulations favor only a single BCB stock.

Finally, like all analytical methods, STRUCTURE is limited by the quality of the data it is provided. The signal STRUCTURE uses to detect population structure (namely departures from HWE) is sensitive not only to the presence of multiple populations but also to factors such as non-random mating, null alleles, genotyping errors, and genetic bottlenecks. At least the last two of these factors, and possibly all four, are issues in the BCB bowhead analyses.

## Acknowledgements

The authors thank Barbara L. Taylor for helpful comments on an earlier draft of this paper. Givens' work was supported by the North Slope Borough (Alaska) and the National Oceanic and Atmospheric Administration (through the Alaska Eskimo Whaling Commission).

## References

Bockstoce, J.R. and Burns, J.J. (1993) Commercial whaling in the North Pacific sector. In *The Bowhead Whale*. J.J. Burns, J.J. Montague, and C.J. Cowles (Eds.). Allen Press, Lawrence. Pp. 563-576.

Evanno, G., Regnaut, S., and Goudet, J. (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., 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.

International Whaling Commission. 2007. Report of the Second Intersessional AWMP Workshop to prepare for the 2007 Bowhead Implementation Review.

Kitakado, T., L.A. Pastene, M. Goto, and N. Kanda. 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.

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

Pritchard, J.K., Wen, X., and Falush, D. (2007) Documentation for *structure* software: Version 2.2. http://pritch.bsd.uchicago.edu/software/structure22/readme.pdf

Tomlin, A.G. (1957) Zveri SSSR I Prilezhasfchikh Stran. Zveri Vostochnoi Evropy I Severnoi Azii. Izdatel'stvo Akademi Nauk SSSR, Moscow. 756 pp. (Translated in 1967 as *Mammals of the USSR and Adjacent Countries. Mammals of Eastern Europe and Adjacent Countries. Vol. IX. Cetacea* by the Israel Program for Scientific Translations, Jerusalem, 717pp.