

Genesis of simulated genetic data and sampling to emulate empirical bowhead whale samples

Karen Martien, Eric Archer and Barbara Taylor
Southwest Fisheries Science Center, 8604 La Jolla Shores Blvd., La Jolla, CA 92038 USA

Introduction

Analysis of bowhead whale genetic data presents unique difficulties in interpretation. Bowhead whales were greatly reduced in number very rapidly and recovered in only a few generations, guaranteeing the population or populations to be strongly out of genetic equilibrium. Sampling is also not random, with some villages preferring to kill large (and hence older) whales, while others prefer smaller (younger) whales. Further, kills primarily occur during migration and often in short time periods and whales are known to segregate by size and reproductive condition during migration. Our simulations attempt to capture both the population dynamics that lead to non-equilibrium genetic compositions and the sampling that matches to empirical samples. Our aim is to provide insight to better interpret the results from standard genetic statistics and tests that assume that populations are in equilibrium. These analyses range from simple metrics, like the level of genetic diversity, F_{st} and whether markers are in Hardy-Weinberg equilibrium, to more complex methods like STRUCTURE (Pritchard et al. 2000, Falush et al. 2003). For the latter, using data from simulations with known dynamics will be particularly helpful in interpreting the magnitude of population differentiation the method can detect given the specifics of the bowhead genetic dynamics and data. We will make the simulation data sets and the sampled data sets available to others wishing to do their own genetic analyses. The objective of this paper is to explain new features added to the genetic simulations since last year's meeting: sampling the simulated data sets to match the empirical sample, which requires removing catches as realistically as possible, and going from only modeling a single stock to modeling two-stock scenarios.

Methods

We use the R-based package Rmetasim, which is a library of functions to perform individual-based population genetic simulations (Strand 2002). Each individual has a multilocus genotype and a mtDNA haplotype. Individuals are structured demographically with an age- or stage-based matrix population model (see '*Demography*' section below; Caswell, 2001). At each time step individuals are randomly assigned their births, stage transitions, and deaths according to the rates specified in the matrix model (used as distributions to incorporate demographic stochasticity). Offspring genotypes are determined by parental genotypes assuming random mating (unless another mating system is specified), independently segregating alleles, and neutrality of markers. R and Rmetasim are available freely from www.cran.r-project.org. This work is being done using R v. 2.4.1 and a modified version of Rmetasim v. 1.1.008 (see *Demography*). For all parameters not explicitly defined here we use the program default values.

Stock Structure Scenarios

The baseline stock structure scenario used in the AWMP control program assumes that the BCB bowheads comprise a single panmictic population. We will simulate that scenario first. There are three primary two-stock scenarios under consideration: the Chukchi Circuit hypothesis, Spatial

Segregation hypothesis, and Temporal Segregation hypothesis (Givens et al. 2006). Though we hope to simulate all three of these two-stock scenarios, it is unlikely we will be able to complete all of them in the time remaining. We therefore request guidance from the participants at the Intersessional meeting of the AWMP sub-committee regarding how these three two-stock scenarios should be prioritized.

Demography

We are using Rmetasim version 1.1.008, which incorporates density dependent population growth, as described in Martien et al. (2006). Density dependence is implemented by interpolating between matrices that represent survival and reproduction rates at carrying capacity and near zero population density. Rmetasim v1.1.008 only allows for linear interpolation between these matrices. However, we have modified the program to allow for non-linear density dependence. The value of a given element of the life history matrix in year t is given by:

$$x_t = x_0 + (x_{\max} - x_0) \left(1 - \left(\frac{N_t}{K} \right)^z \right)$$

where:

- x_t is the value of the element in year t
- x_0 is the value of the element at carrying capacity
- x_{\max} is the maximum value of the element (near zero population size)
- N_t is the size of the population at the start of year t
- K is the carrying capacity of the population
- z is the shape parameter.

The demographic matrices used for this study are the same as those in Ripley et al. (2006). They are for a stage-based model with the following 8 stages: 5 juvenile stages (J1-J5), adult females (AF), adult males (AM), and supermales (SM). Stage transition probabilities were calculated using the fixed stage duration method (Caswell 2001). The life history parameter estimates presented in Brandon and Wade (in press) were used to develop two matrices, one for which $\lambda = 1.00$, the other for which $\lambda = 1.042$ (Table 1). These matrices were used to represent vital rates at carrying capacity and near zero population size, respectively. Ripley et al. (2006) also presented a matrix for which $\lambda = 1.026$. That matrix will not be used in this study.

Table 1. Demographic parameters at carrying capacity ($\lambda=1.00$) and near zero population size ($\lambda=1.042$). For each stage, stage duration (T) and age-specific survival (σ) are used to calculate the matrix model parameters P (survival in stage) and G (stage transition probability) according the fixed stage duration model (Caswell 2001; Ripley et al. 2006).

Stage	$\lambda = 1.00$					$\lambda = 1.042$				
	T	σ	γ	P	G	T	σ	γ	P	G
J 1	4	0.800	0.173	0.661	0.139	2	0.925	0.470	0.490	0.435
J 2	4	0.978	0.242	0.741	0.236	3	0.985	0.315	0.675	0.310
J 3	4	0.978	0.242	0.741	0.236	3	0.985	0.315	0.675	0.310
J 4	4	0.978	0.242	0.741	0.236	3	0.985	0.315	0.675	0.310
J 5	4	0.978	0.242	0.741	0.118	3	0.985	0.315	0.675	0.155
AF	50	0.978	0.011	0.967	0.011	50	0.985	0.004	0.981	0.004
AM	50	0.978	0.011	0.967	0.011	50	0.985	0.004	0.981	0.004

SM	25	0.978	0.030	0.948	0.029	25	0.985	0.019	0.966	0.019
----	----	-------	-------	-------	-------	----	-------	-------	-------	-------

Genetic initialization

We used the software SIMCOAL v2.1.2 (Laval and Excoffier 2004) to generate mitochondrial DNA (mtDNA) and microsatellite distributions from the coalescent which were in turn used to initialize the simulated populations. In order to initialize SIMCOAL, we estimated the average effective population size (N_e) at carrying-capacity as,

$$N_e = \frac{m}{2 \cdot \left(1 - \left(\frac{H_t}{H_0} \right)^{\frac{g}{t}} \right)}$$

where:

H_0 = initial heterozygosity

H_t = heterozygosity at time t

t = elapsed time in years

g = generation time (= 37 years)

m = multiplier chosen to start the simulation burn-in phase close to equilibrium (= 1.2)

N_e was estimated for mtDNA and microsatellites separately. For microsatellites, the above equation actually estimates $2 \times N_e$, which is the value required by SIMCOAL. We calculated the average effective population size (\bar{N}_e) as the harmonic mean of N_e from 20 population projections lasting 2000 years (t), each initialized with the same survival and reproduction matrices as in the full simulation for a given scenario. The sample size generated by SimCoal was \bar{N}_e for the mtDNA sequences and the smaller of \bar{N}_e and 1000 for the microsatellite loci. The mtDNA sequence was specified to be 397 bp, with a Ts:Tv of 10:1, and a mutation rate of 3.9×10^{-3} . For the microsatellites, 35 loci with an average mutation rate of 1.5×10^{-3} were simulated. Mutation parameters were tuned to produce diversity comparable to that observed as has been done previously (Taylor et al. 2000). For each simulation replicate to be run, a sample of both markers was independently generated.

Burn-in

In order to ensure that the simulated populations were in equilibrium, a burn-in phase of 2000 years was conducted following initialization. Previous examinations of the trajectories of the number of mtDNA haplotypes, microsatellite alleles, and heterozygosity in both markers indicated that this was a sufficient amount of time to ensure that these values were relatively stable.

In scenarios assuming two populations (stocks), we assume that the two populations diverged from a single population at some point in the past and have experienced no gene flow since diverging. We again used SIMCOAL v2.1.2 to simulate the divergence event and generate haplotype and allele frequency distributions for the two populations for use in initializing Rmetasim.

The time since two populations diverged and their effective population sizes will determine the degree of population differentiation (F_{ST}) between them. We chose a divergence time such that the current level of differentiation expected between the two populations is equal to that between BCB bowheads and those from the Sea of Okhotsk ($F_{ST} = 0.062$; LeDuc et al. 2005). This level of differentiation serves as an upper-bound on what is plausible between putative stocks of BCB bowheads, as the populations we are modeling are in much closer geographic and temporal proximity (even mixing under some scenarios) than are the geographically disjunct BCB and Sea of Okhotsk populations.

Post-burn-in (whaling phase)

At the end of the burn-in phase, the populations are subjected to an annual removal of whales designed to mimic the historical kill from the commercial harvest, and Russian and Alaskan subsistence catches from 1848 to 2006. The harvest data were provided by C. George and J. Zeh and are the same data being used in the AWMP (George and Zeh, pers. comm.). In each year catches are allocated to one of two stocks based on the catch allocation matrices for the particular stock structure scenario being simulated (Givens et al., 2006). To do this, catches first have to be assigned to an area (W, E, or O). All animals killed in the Russian subsistence harvest are assumed to have come from area W, while those from the Alaskan subsistence harvest are assumed to be from area E.

Animals killed in the commercial harvest are assigned to areas probabilistically according to the catch allocation matrix. For each year in which there was a commercial harvest, we first calculated the catch allocation by area for the entire year, rather than each season, by weighting the catch allocation in each season by the seasonal split. We then multiplied by the resulting (proportional) catch allocation by the total harvest for that year in order to determine the number of animals killed in each area. The Russian subsistence was subtracted from the animals taken from area W and the Alaskan subsistence was subtracted from the catch in area E. The remaining catch in each area was divided by the total commercial catch for the year in question in order to determine the proportion of the commercial harvest that came from each area. These values were used to assign animals killed in the commercial harvest to an area in a probabilistic manner. In years for which the estimated subsistence harvest for an area exceeded the total catch in that area, the commercial catch for the area was set to zero.

We next needed to determine the probability that an animal killed in a given area, year, and season came from stock 1. To do this, we multiplied the exposure matrices, which specify the proportion of each stock that is in each area in a given year and season, by the abundances of the two stocks in the year in question. These historic abundance estimates were taken from the AWMP Lite spreadsheet when run for the appropriate scenario and parameters. This multiplication resulted in an estimate of the number of animals from each stock that are in each area. We then divided the number of stock 1 animals in each area in a given season by the total number of animals in the area and season. The result is an estimate of the probability that an animal killed in the area, year, and season in question comes from stock 1.

The first whales removed from the simulated populations in a given year are those for which biological samples and measurements were collected from the Alaskan subsistence catch (available from 1974 to 2006). Each sampled whale is assigned to one of the two putative stocks based on the year and season in which it was killed using the methods described above.

In order to match sampled whales to simulated individuals within their assigned populations as closely as possible, we estimated the age and gender for sampled whales when that information was

unavailable. For whales that did not have age estimates (Lubetkin and Zeh 2006), morphological characteristics were used to classify whales into one of three age bins (< 60 years, 60 to 90 years, > 90 years) using the algorithm given in Morita and Zeh (Pers. Comm.). An age was then chosen at random from all individuals in the simulated population within the chosen age bin. If there was insufficient morphometric data for classification, then an age was chosen at random from the entire population. In either case, the base age distribution used was that of the assigned population in the year the whale was killed. If the gender of the sampled whale was unknown, then it was randomly selected from the ratio of known-gender whales killed in that year. Age and gender were estimated as above in each simulation replicate in order to account for the error in the techniques.

For each sampled whale, an individual from the simulated population to which that whale was assigned was chosen out of all individuals of the same sex and age. If no individuals in the simulated population were found to match exactly, a match was sought among individuals that were one year older or younger than the whale under consideration. This age window was continually expanded until at least one whale was available to be killed. In this manner, all sampled whales were matched to a unique simulated individual. If genetic data were available for the sampled whale, the genetic data of its corresponding simulated match were saved.

Following the removal of the biologically sampled Alaskan whales, the unsampled portion of the Alaskan subsistence catch (total reported Alaskan catch minus sampled catch) was then allocated to stock and removed from the simulated populations. This was followed by allocation and removal of the commercial harvest and Russian subsistence catch. For these final three components of the catch, individuals were selected at random with respect to age and sex. Following this simulated whaling, the populations were then projected forward one year and the whaling for the next year would occur again as described above.

The final output was a simulated genetic sample representing the demographic composition of the empirical harvest sample and all individuals surviving in each of the simulated populations. The total number of individuals killed in each stock was saved each year for comparison with the catch allocation matrices. Annual population abundances are also saved for comparison with trajectories from historical trend analyses (Brandon and Wade 2007).

Discussion

The ability to compare analyses from empirical genetic data to simulated data should improve our understanding of bowhead whales and hence deliberations about their management. A prioritized list of AWMP-Lite scenarios that are both plausible and highest risk would aide in providing the most critical scenarios for the upcoming Scientific Committee meeting.

References

Brandon, J. and Wade, P.R. in press. Assessment of the Bering-Chukchi-Beaufort Seas stock of bowhead whales using Bayesian model averaging. *Journal of Cetacean Research and Management*.

Caswell, H. (2001) *Matrix Population Models: Construction, Analysis and Interpretation*. 2nded. Sinauer Associates, Sunderland, Massachusetts, USA.

- 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.
- George, J.C. and S.E. Moore. 2006. Hypothetical stock structure archetypes for the Bering-Chukchi-Beaufort Seas bowhead whale population. Paper SC/58/BRG27 submitted to the Annual meeting of the Scientific Committee of the International Whaling Commission, June, 2006. 20pp.
- Givens, G., A.E. Punt, and J. Zeh. 2006. The scenario space for the *Bowhead SLA implementation review*: a search for plausible trials exhibiting management risk. Paper SC/58/AWMP8 submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, June, 2006. 23pp.
- Laval, G. and L. Excoffier. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* 20(15):2485-2487.
- LeDuc, R.G., A.E. Dizon, A.M. Burdin, S.A. Blokhin, J.C. George, and R.L. Brownell, Jr. 2005. Genetic analyses (mtDNA and microsatellites) of Okhotsk and Bering/Chukchi/Beaufort Seas populations of bowhead whales. *J. Cetacean Res. Manage.* 7(2):107–111.
- Lubetkin, S.C. and J.E. Zeh. 2006. Deriving age-length relationships for bowhead whales (*Balaena mysticetus*) using a synthesis of age estimation techniques. Paper SC/58/BRG14 submitted to the Annual meeting of the Scientific Committee of the International Whaling Commission, June, 2006. 20pp.
- 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.
- Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.
- Ripley, B.J., K.K. Martien and B.L. Taylor. 2006. A simulation approach to understanding non-equilibrium dynamics in a recovering long-lived species: the bowhead whale. Paper SC/58/BRG13 submitted to the Annual meeting of the Scientific Committee of the International Whaling Commission, June, 2006. 12pp.
- 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.
- Taylor, B.L., S. J. Chivers, S. Sexton and A. E. Dizon. 2000. Estimating dispersal rates using mitochondrial DNA data and incorporating uncertainty. *Conservation Biology*:1287-1297.