Use of an individual-based simulation of BCB bowhead whale population dynamics to examine empirical genetic data

Archer, E.[*], Martien, K., Taylor, B.L., LeDuc, R.G.
Southwest Fisheries Science Center, 8604 La Jolla Shores Blvd., La Jolla, CA 92038 USA.
[*]eric.archer@noaa.gov

Givens, G.H.
Dept. of Statistics, 1877 Campus Delivery, Colorado State University, Fort Collins, CO 80523 USA.

and George, J.C.
North Slope Borough Dept. of Wildlife Management, PO Box 69, Barrow, AK 99723 USA.

Abstract

To better interpret whether genetic data from bowhead whales could have resulted from a single stock, we created an individual-based model of bowhead whale population dynamics and genetics using the R package *rmetasim*. The model re-created as closely as possible all aspects of the demography, genetics, and whaling history of bowheads. Simulated datasets were generated by sampling from the simulated population in a way that matched the age, sex, and geographic distribution of the empirical samples. These simulated datasets were used to generate null distributions for a variety of genetic analyses, against which we compared the empirical bowhead dataset. In most respects, the results of our analysis indicate that the empirical genetic data sampled from BCB bowhead whales are consistent with our model of a single, randomly-mating population. Of the 55 spatial, temporal, and cohort comparisons we examined (11 stratifications for 5 measures of genetic differentiation), only the mitochondrial F_{ST} between fall and spring St. Lawrence Island, and the microsatellite F_{ST} between Barrow and St. Lawrence Island exhibited a significant difference between the simulated and empirical datasets. Our results show that the empirical STRUCTURE analyses are entirely consistent with a single population that is out of genetic equilibrium due to the effects of commercial whaling.

Introduction

Analyses of bowhead whale genetic data present unique difficulties in interpretation. Bowhead whales were greatly reduced in number very rapidly and recovered (Brandon and Wade 2007) in only two and a half generations (Taylor *et al.* 2006), 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 (Suydam *et al.* 2004). 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 (Angliss *et al.* 1995).
In this paper, we describe an individual-based simulation that attempts to capture both the population dynamics that lead to non-equilibrial genetic compositions and match the non-random empirical

samples as closely as possible with respect to birth year and sex. Our aim is to interpret the results from standard genetic statistics and analyses by generating null distributions based on a single randomly mating population that does not assume equilibrium conditions. These analyses range from simple metrics, like whether markers are in Hardy-Weinberg equilibrium, and measures of population subdivision (F_{st}, χ^2, and Φ_{st}), to more complex methods like STRUCTURE (Pritchard *et al.* 2000, Falush *et al.* 2003).

Methods

The simulation is based on the *rmetasim* package (version 1.1.008 - Strand 2002), run in the R statistical environment (version 2.4.1 - R Development Core Team 2006). *Rmetasim* is a library of functions which performs individual-based population genetic simulations. Each individual has a multi-locus genotype and a mitochondrial DNA (mtDNA) haplotype. Individuals are structured demographically with a 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, independently segregating alleles, and neutrality of markers. For all parameters not explicitly defined here we use the program default values.

Demography

Rmetasim version 1.1.008 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. Although this version of *rmetasim* only allows for linear interpolation between these matrices, 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:

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

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 for a stage-based model with the following 7 stages: 5 juvenile stages (J1-J5), adult females (F), adult males (M) (Ripley *et al.* 2006). Stage transition probabilities were calculated using the fixed stage duration method (Caswell 2001). The life history parameter estimates presented in Brandon and Wade (2007) 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. We set *z* to 4, which is the posterior median from Brandon and Wade's (2007) backward projection model (referred to as '1848DD' in their paper).

Genetic initialization and burn-in

We initialized the simulated populations using mitochondrial haplotype and microsatellite allele frequency distributions generated by the coalescent program SIMCOAL v2.1.2 (Laval and Excoffier 2004). Initializing from a coalescent rather than with random allele and haplotype frequencies greatly reduced the number of generations required for the simulations to reach equilibrium. 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 (= 1.45)

N_e was estimated for mtDNA and microsatellites separately. For microsatellites, the above equation actually estimates $2N_e$, which is the value required by SIMCOAL. We calculated the average effective population size ($\overline{N}_e$) as the harmonic mean of N_e from 20 population projections lasting 4000 years (t), each initialized with the same survival and reproduction matrices as in the full simulation. The sample size generated by SIMCOAL was $\overline{N}_e$ for the mtDNA sequences and the smaller of $\overline{N}_e$ 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 9.4×10^{-3}. For the microsatellites, two groups of loci were simulated representing the 11 "original" and 22 "new" loci used in Givens *et al.* (2007). Average mutation rates were set at 3.0×10^{-4} and 1.5×10^{-3} for the original and new loci respectively. Mutation parameters were tuned to produce diversity comparable to that observed as has been done previously (Taylor *et al.* 2000).

In order to ensure that the simulated populations were in equilibrium, a burn-in phase was conducted following initialization. Previous examinations of the trajectories of the number of mtDNA haplotypes, microsatellite alleles, and heterozygosity in both markers indicated that 4000 years was a sufficient amount of time to ensure that these values were relatively stable (Figure 1). A sample of all markers was independently generated from SIMCOAL for each burn-in replicate.

Simulated whaling and sampling

For each burn-in replicate, multiple replicates of an annual removal of whales designed to mimic the historical kill were conducted. The historical kill encompasses the commercial harvest and Russian and Alaskan subsistence catches from 1848 to 2006. The harvest data used in the model are the same data being used in the AWMP (George and Zeh, pers. comm.). In each year of a whaling replication, the first whales removed from the simulated populations are those for which biological samples and measurements were collected from the Alaskan subsistence catch (available from 1974 to 2006), followed by a random removal of the remainder of the recorded catch. Age and gender were independently estimated as described below for each harvested whale in each simulation replicate in order to account for error in the aging techniques and variance in sampling.

If the gender of the sampled whale was unknown, then it was randomly selected using the ratio of known-gender whales killed in that year. A 50:50 ratio was assumed if no known-gender whales were available in a particular year (empirical data for all whales from 1974-2006 for which sex was identified = 487(F):468 (M), very close to parity). In order to match harvested whales to simulated individuals, the age of each harvested whale was determined in a hierarchical fashion based on the quality of data available. For whales that were aged and had estimates of standard errors from one of the methods given in Lubetkin and Zeh 2007, ages were randomly sampled from a normal distribution and rounded to the nearest whole age.

For whales that were not aged, a Classification and Regression Tree (CART – Breiman *et al.* 1984), as implemented in the R package *rpart* v3.1-34 (Therneau and Atkinson 2006), was used to estimate the age bin to which they belonged based on morphological characteristics (gender, body length, baleen length, anterior flipper length, peduncle girth, and length of the peduncle white patch). The CART tree was created from 177 known-aged samples using ten age bins (Figure 2), which were selected from an exploratory series of CART regression trees. Bins were selected probabilistically based on the distribution of bin membership from the training data in the node to which an unknown sample was assigned (Table 2). The age for each sample was then chosen at random from all individuals in the simulated population within the chosen age bin. If the morphometric data necessary to classify a sample at a particular split in the CART tree was missing, surrogate variables were used if available. If there was insufficient morphometric data for the CART algorithm, then an age was chosen at random from the simulated population. In all cases, the age distribution being chosen from was that of the simulated population immediately following burn-in, which was considered a stable age distribution.

For each sampled whale in the empirical dataset, an individual of the same age and gender was randomly selected from the simulated population. If no individuals in the simulated population were found to match exactly, all individuals within a progressively increasing age window around the whale under consideration were examined. A match was selected from within the first age window which contained at least one individual of the same gender as the whale under consideration, with probabilities assigned to each individual based on their gender and the size of the window. The probability of choosing an individual of the same sex as the sample under consideration ranged from one for an age window of zero (all simulated individuals were of the same age as the sampled whale) to 0.5 when all individuals in the population were considered. The probability of choosing an individual of the opposite sex was one minus this value. In this manner, all sampled whales were matched to a unique simulated individual.

Following the removal of any biologically sampled whales, the un-sampled portion of the harvest was then removed from the simulated populations. In all cases, whaling was restricted to individuals older than one year. The genetic data of the simulated whales selected to be killed each year were saved if genetic data were available for their matched harvested counterparts. If biopsies were collected in a given year, the genetic data of an equivalent number of randomly selected simulated individuals still alive in the population were also saved. Following a simulated year of whaling and sampling, the populations were then projected forward one year and the whaling for the next year would occur again as described above.

In order to ensure that the abundance trajectories from the simulations were similar to those of historical trend analyses (Brandon and Wade 2007), two abundance "gates" were established that replicates had to pass through. Replicates that had trajectories outside of the 99% confidence intervals of the first and last years of estimated population abundance (1978 and 2001), including those that went extinct, were discarded. For each of 50 burn-in replicates, the first ten successful whaling replicates were saved producing a total of 500 replicates. 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. Annual population abundances were saved for comparison with trajectories from historical trend analyses (Brandon and Wade 2007).

Introduction of errors

Microsatellite datasets inevitably contain genotyping errors. Error rates reported in the literature range from 0.1% to 48% (Morin *et al.*, 2007). To examine the effect of genotyping errors on the analytical methods applied to the bowhead dataset, we introduced genotyping errors into our simulated dataset. By comparing genotypes for duplicate samples included in the original empirical dataset, Morin *et al.* (2007) estimated an overall error rate of 0.01 for the bowhead microsatellite data. Of these, 40% were apparent cases of allelic dropout, i.e., the individuals were scored as homozygotes in one genotyping attempt and as heterozygotes in another.

The number of genotyping errors introduced into a simulated dataset was determined by drawing a random deviate from a binomial distribution given the overall error rate (0.01) and the number of alleles in the dataset (18,314). The alleles to which the errors were applied were chosen at random from the entire dataset. When an error occurred, it had a 0.4 probability of being an instance of allelic dropout, in which case the allele in question was set equal to the other allele the individual possessed at that locus, making the individual homozygous at that locus. Otherwise, the allele was replaced by a different allele chosen at random from the allele frequency distribution for the appropriate locus.

Standard genetic analyses

We used a suite of standard population genetic algorithms to analyze both the genetic samples from the simulations as well as the matching empirical genetic data. Genepop v3.3 (Raymond and Rousset 1995) was used to run the Hardy-Weinberg test of heterozygote deficiency on the 213 samples from Barrow using both the 11 "original" and 22 "new" loci. For this test, an MCMC burn-in of 30,000 iterations was used, with a final chain length of 10,000 and batch size of 100. We also calculated Hardy-Weinberg disequilibrium across all loci using Fisher's method (Ryman and Jorde 2001).

The F_{st} for mtDNA and microsatellites was calculated following Weir and Cockerham (1984), and χ^2 for both sets of markers was calculated following Roff and Bentzen (1989). The AMOVA Φ_{st} metric was calculated for mtDNA data using the R package *ade4* (Chessel *et al.* 2004). Temporal, spatial, and age cohort stratifications of samples for F_{st}, χ^2, and Φ_{st} are the same as those given in LeDuc et al (2007). Unless otherwise noted, *p*-values reported in this paper are the proportion of replicates with test-statistics ≥ the value obtained from the empirical data using the same (matched) samples. Comparisons with *p*-values ≤ 0.05 were considered to indicate empirical results inconsistent with our model.

STRUCTURE analyses

We used the Bayesian clustering program STRUCTURE v. 2.1 (Pritchard *et al.* 2000; Falush *et al.* 2002) to analyze the simulated datasets. STRUCTURE clusters samples into groups so as to minimize the amount of Hardy-Weinberg disequilibrium within the groups. Each sample is assigned probabilistically to a group. The number of groups (*k*) is determined by the user. By comparing the probability of the groupings defined for different values of *k* ($\Pr(X|K)$), STRUCTURE can be used to gain insight into the number of populations present in the sample set.

Analyses of the empirical dataset (Givens *et al.* 2007) have suggested that the model that fits the data best is one with three groups – the Russian samples make up one group, while the Canadian and BCB samples are split approximately evenly between the last two groups. There is no clear geographic or temporal concordance within the two Canadian/BCB groups, though there is some suggestion of temporal pulsing of the two groups within the samples collected in Barrow in the fall, and a significant linear trend in stock membership at Barrow in the spring (Givens *et al.* 2007).

To determine whether the patterns seen in the empirical dataset are consistent with a single stock, we used STRUCTURE to group the empirical dataset and each of the simulated datasets into two groups (*k*=2). It was necessary to re-analyze the empirical dataset because Givens *et al.*'s (2007) analyses included data from Russia and Canada, which were not included in our simulations. All of our analyses were based on microsatellite data only, using all 33 microsatellite loci and only those samples included in the 33 locus microsatellite reference dataset. In contrast, Givens *et al.* use only the 22 new loci and the corresponding reference sample set. Our STRUCTURE analyses all used an admixture model with correlated frequencies, as was done by Givens *et al.* Given the low level of genetic differentiation expected between separate BCB stocks should they exist, this model is the most appropriate.

We determined the appropriate chain length for our STRUCTURE analyses by analyzing the same simulated dataset ten times for a given chain length. We calculated the variance across the ten analyses in the assignment probability for each individual, and then averaged the variance across individuals. We repeated this procedure for a variety of chain lengths to determine the best trade-off between runtime and the precision of the assignment probabilities. Unless indicated otherwise, all analyses on assignment probabilities were conducted on the logit-transform of these values.

We compared the results obtained for the simulated datasets to those from the empirical dataset using a variety of summary statistics. For each STRUCTURE analysis, we calculated the variance in the assignment probability across individuals. In cases where many individuals assigned strongly to one population or the other, this variance will be much higher than in those cases where most individuals had a roughly equal probability of coming each population. We also calculated F_{st} between the two groups defined by STRUCTURE. For this calculation, each sample was assigned to the group for which it had the highest assignment probability.

Givens *et al.* (2007) found evidence of temporal pulsing in their analyses of the empirical data – the animals passing by Barrow in the fall seemed to assign predominantly to one group during the beginning and end of the migration, and predominantly to the other group during the middle of the migration. We used two measures to quantify 'pulsiness' in our STRUCTURE analyses. We first fit a three-phase step function to a plot of the day of the year a whale (simulated or real) was harvested versus its assignment probability to group 1. For the step function, we did not perform a logit-transformation of assignment probability. We applied two constraints to the step function: 1) each phase of the function was at least five days long and 2) the sign of the change in the value of the function was opposite for the first and second steps. This second constraint ensured that the step function reflected a 'bump' or 'dip' in assignment probabilities rather than a monotonic increase or decrease. We compared the step function to a horizontal line equal to the mean assignment probability for all fall Barrow whales. We kept track of the deviance of the step function from the horizontal line (calculated as the area between the two functions), the mean assignment probability for each phase of the step function, and the days on which the steps occurred.

The step-function allowed us to examine the simulated datasets for the specific temporal pattern observed in the empirical dataset, namely, 'pulses' of one group early and late in the fall migration, and a 'pulse' of the other group in the middle. In order to quantify more general pulsing patterns in both the simulated and empirical datasets, we calculated a nearest neighbor correlation coefficient. To do this, we calculated the average assignment probability to group 1 for all whales harvested on a given day of the year, again restricting our analysis to animals harvested in Barrow in the fall. We used the R package *spatstat* (Baddeley and Turner, 2005) to calculate the nearest neighbor correlation in the assignment probability between days. Positive nearest neighbor correlations indicate that the assignment probability on a given day is more similar that of the days immediately preceding and following it than it is to a randomly chosen day.

Givens *et al.* (2007) also found a statistical significant trend in assignment probability for individuals harvested at Barrow in the spring. To test for such a trend in the simulated datasets, we calculated the linear regression coefficient for assignment probability regressed on day of year for all spring Barrow whales.

Results

Simulation diagnostics

The population trajectories for the 500 replicates are given in Figure 3. At the nadir, the median abundance was 1197 with a range of 806 – 1608. 4% of the replicates ended with an abundance greater than 12,000 in 2006. Figure 4 shows the distribution of ages within each stage at the end of burn-in (A) and at the end of the simulation (B). The mean age of all reproductive individuals was 48 (95-percentile = 13 – 129) at the end of burn-in and 33 (95-percentile = 12 – 76) at the end of the simulation. At the end of the simulation, approximately 48% of the individuals were reproductive adults and the sex ratio was not significantly different from 50:50.

Genetic diversity of the empirical data, as measured by the number of alleles (haplotypes for mtDNA), heterozygosity, and θ_H, was similar to the distributions of these metrics from the matched simulated samples (Figure 5). Only measures of heterozygosity and θ_H for mtDNA were outside of the simulated distributions, a result of the skewed haplotypic frequency distribution in the empirical data.

Standard genetic analyses

In the empirical data, nine of the 33 loci were found to be out of Hardy-Weinberg equilibrium (HWE) with a combined *p*-value using Fisher's method of 2.3 x 10^{-6}. In the simulation, the median number of loci out of HWE was two, with a maximum of five (Figure 6A). When errors were added to the simulated data, the median number out of HWE increased to 3 with a maximum of 11 (Figure 6B). The *p*-value for the test with errors was 0.006.

There was a relatively uniform distribution of MCMC HWE *p*-values across loci without errors included (Figure 7A). The combined *p*-value using Fisher's method for the empirical data was 2.3 x 10^{-6}, which was less than the minimum value in the simulation of 0.0025. The 95-percentile of the simulated distribution was 0.022 – 0.989. When errors were introduced into the simulated data, the distributions of the MCMC HWE *p*-values and the Fisher's method *p*-values were highly skewed (Figure 7B). The median of Fisher's method *p*-values was 2.5 x 10^{-2}, with a 95-percentile of 3.14 x 10^{-5} – 5.1 x 10^{-1}. 61% of this distribution was ≤ 0.05. The empirical value was at the lower 1% of this distribution, making it inconsistent with the model.

In the analyses of empirical mtDNA data, the only comparisons found to be significant with standard permutation tests were the three χ^2-tests of cohorts born before 1949 and those born after 1979 (Table 3A). However when compared to the simulated data, these stratifications were found to be consistent with our model (simulation *p*-values > 0.15). The only stratification of the mtDNA data found to be inconsistent with our model was the comparison between fall and spring samples from St. Lawrence Island (*p* = 0.008). All other temporal, spatial, and cohort stratifications had simulation *p*-values > 0.15, indicating that they were consistent with our model.

Givens *et al.* 2007 found significant differentiation in the microsatellite data between samples from Barrow and SLI using a Fisher's exact test. Using standard permutation tests, none of the comparisons we examined were significant, although the Barrow v. SLI comparison had the lowest permutation *p*-value for both F_{st} and χ^2 tests (Table 3B). The Barrow v. SLI comparison was also the only stratification of the microsatellite data that was inconsistent with our model with a simulation *p*-value of 0.04. All other stratifications had simulation *p*-values > 0.09.

The introduction of errors into the simulated microsatellite data did not make a substantial change in the results of any of the F_{st} or χ^2-tests. While the *p*-value of some results changed slightly, there was not a consistent pattern in the direction of the change.

STRUCTURE

The variance between runs in assignment probability decreased rapidly as the run length was increased from 50,000 iterations to 200,000 iterations, and then remained relatively stable for run lengths up to 1,000,000 iterations (Figure 8). We therefore ran all of our analyses for 200,000 iterations, following a burn-in of 30,000 iterations.

The results of our STRUCTURE analyses of the empirical dataset were similar to those reported by Givens *et al.* (2007) despite the fact that we included all 33 microsatellite loci in our analyses and did not include samples from Canada or the Sea of Okhotsk (Figure 9). A smoothed fit of day-of-year versus group 1 assignment probability for fall Barrow samples shows a 'pulse' of group 1 ancestry between days 270 and 280, though the effect is not as pronounced as in Givens *et al.*'s analysis (Figure 10). We did find a significant trend in group 1 ancestry over the course of the spring migration at Barrow (Figure 11). However, unlike in Givens et al.'s analysis, this trend was not statistically significant (*p* = 0.111).

The results of the STRUCTURE analysis of the empirical dataset were consistent with those from the simulated datasets for all measures we examine (Table 4). The *p*-values for the various comparisons ranged from 0.112 to 0.402 without genotyping errors added to the simulated datasets, and from 0.126 to 0.462 with errors added.

Discussion

Consistency of empirical results with simulation

In most respects, the results of our analysis indicate that the empirical genetic data sampled from BCB bowhead whale are consistent with our model of a single, randomly-mating population that mimics their history of whaling and subsequent recovery. Of the 55 spatial, temporal, and cohort comparisons we examined (11 stratifications for 5 measures of genetic differentiation), only two, the mitochondrial F_{ST} between fall and spring St. Lawrence Island, and the microsatellite F_{ST} between Barrow and St. Lawrence Island, exhibited a significant difference between the simulated and empirical datasets. The results of the STRUCTURE analyses of the empirical dataset have had the greatest influence on the construction of stock structure hypotheses in the bowhead special assessment. However, our results show that the empirical STRUCTURE analyses are entirely consistent with a single population that is out of genetic equilibrium due to the effects of commercial whaling.

There were three analyses that suggest the empirical data are not consistent with our model. First, the number of loci out of HWE in the empirical data from the Barrow samples is significantly greater than in the simulated data. When errors were introduced in the simulated data, the difference between the two decreased dramatically but remained significant. This was the only analysis for which the introduction of genotyping errors into the simulated datasets had a substantial impact on the results. While the comparison between the empirical and simulated results was still significant even with errors introduced, our results highlight the sensitivity of HWE to genotyping errors. For instance, in the simulation without genotyping errors only 0.026 of the simulated datasets had 5 or more loci out of HWE. When genotyping errors were introduced, this probability jumped to 0.298.

The introduction of errors showed a greater effect on the distribution of the overall Fisher's method *p*-value for HWE. Without errors, this distribution was relatively uniform as would be expected under a standard null hypothesis. When errors were included, the distribution became highly skewed towards very small *p*-values. While this skew was not large enough to make the empirical finding of overall Hardy-Weinberg disequilibrium consistent with our model, the implication is that with even the relatively low error rate identified by Morin *et al.* (2007), there is a large probability (61% of the replicates had a $p \leq 0.05$) of falsely assessing widespread disequilibrium.

We used a very simplistic model for introducing genotyping errors to the simulated datasets. Though we did incorporate the observed allelic dropout rate, all errors were random with respect to the loci at which they occurred and the alleles and individuals that were affected. In reality, genotyping errors are often not random. Some loci may be more susceptible to errors than others, as are some samples. Allele length and frequency may also affect the likelihood of the allele being correctly scored. Stutter bands and slippage would result in the mis-scored allele being very close in length to the correct allele, rather than reflecting the overall allele frequency at the locus. Since we are unable quantify the various biases inherent in genotyping errors, we chose a simplistic model that only included allelic dropout and random errors. If we had incorporated other realistic biases into our simulated genotyping errors, we would likely have seen an even stronger impact on the expected distribution of the number of loci out of HWE.

The second analysis that indicated a lack of consistency between the empirical and simulated datasets was the F_{ST} test between fall and spring St. Lawrence Island for mitochondrial sequences. The magnitude of the observed mtDNA F_{ST} value (0.054) in this test results from the difference in the frequency of one haplotype (BH42) between spring and fall samples (6 in fall, 1 in spring). Givens *et al.* (2007) report a potential difference between seasons in the SLI samples in the microsatellite data. Because they do not see a difference between spring Savoonga (one of two villages in SLI) and Barrow, but significant difference between fall Savoonga and both spring and fall Barrow, the implication is that the fall Savoonga samples are outliers. Indeed, five of the six fall samples that possessed haplotype BH42 came from Savoonga. It should be noted that the sample sizes are relatively small in both fall and spring strata and therefore may not adequately represent the haplotypic distributions of the whales near these villages in these seasons.

The final analysis that was inconsistent with our model was the F_{ST} test between Barrow and St. Lawrence Island (SLI) for the microsatellite markers. It is possible that this result is being influenced either by the unusual distribution of the fall Savoonga samples mentioned above or by the loci that were found to be out of HWE in the Barrow samples. We partially examined the effect of the latter by running this analysis again after removing the six samples most influential on HWE, which were identified by Morin *et al.* (2007). The removal of these samples did not significantly change the lack of consistency between the empirical and simulated data either with or without errors introduced into the simulated data.

Simulation construction

Rather than the null hypothesis used in standard genetic analyses of a panmictic population at equilibrium, the simulation presented in this paper represents a null hypothesis based on a very specific model of a single population that is out of equilibrium due to its population history. One of the strengths of this simulation is that by matching the age and sex characteristics of the empirical samples where possible, this null hypothesis inherently incorporates any potential demographic biases in the sampling process.

Our model relies on several parameters controlling the population dynamics and genetic diversity. Where possible, we have used empirical data and parameter values from independent sources. If these were not available, parameters were iteratively tuned to ensure that other aspects of simulation either fit published results or matched the empirical data as closely as possible. It is important to note that this process does not ensure that our parameter values are accurate with respect to a "true" single population, as actual stock structure and biological/demographic parameters remain unresolved at this time.

An example is the procedure by which carrying capacity (*K*) was selected. With the value of the logistic growth shape parameter (*z*) set at the median posterior value from Brandon and Wade's (2007) backward projection model, we selected a value of *K* such that a majority of the replicates did not go extinct and passed through the abundance "gates". Under these constraints it can be seen that many of the population dynamics parameters, most notably *K, z,* and the population growth rate (*r* – not specified in the model, but resulting from the reproduction and survival matrices), will be closely correlated such that multiple combinations would work. While our goal was not to estimate these parameters, through iterative testing we determined that, given the historical catch record, there was a small range over which they could vary and still meet the extinction and abundance constraints. In our tests, choices of *K* outside of the range of approximately 11,900 - 12,400 would not produce useable replicates. This range is well within the 95% credibility interval for *K* (9,112 – 13,610) from the Brandon and Wade (2007) assessment model most similar to our simulation, which is expected given that this study used the same historical catch and abundance data.

This simulation makes the assumption that the carrying capacity of the population now is the same as it was prior to the onset of commercial whaling. This could be violated if there has been a substantial change in the ecosystem or the range of the population has either expanded or contracted. The 2001 abundance estimate of 10,545 suggests that the population is very close to the carrying capacity estimated in Brandon and Wade, making it unlikely that there has been a decrease in carrying capacity. Whether or not there has been a significant increase in carrying capacity will require future surveys.

A second result of the model constraints is that we were unable to directly control population abundance at the nadir or the range over which it varied. Because a population reduced to a very small size will be unable to contain the entire genetic diversity of its larger progenitor, this factor is likely to greatly affect the degree of genetic disequilibrium within the population. The smaller the nadir, the stronger the signal of a generational gene-shift (Ripley *et al.* 2006) is expected to be, in which the genotypes of individuals born before and after the nadir will appear to have come from two different distributions. Therefore, it is important to note that the results of study are conditional on the nadir being approximately 1100.

Strong GGS can also arise if the commercial kill was selective with respect to sex or age. Though there is no evidence that whalers were intentionally selective in their hunting, the fact that bowheads segregate by age and sex during migration may have resulted in selectivity on the basis of availability (Bockstoce, 1986). Evidence for some selectivity can be found in the in the fact that the average size of whales killed decreased between the beginning of the fishery and 1874, the only period for which such data are available (Bockstoce and Burns, 1993). If a similar kind of selectivity continued throughout the commercial hunt, the portion of the population that survived through the nadir would tend to represent younger cohorts. This would have the effect of temporarily increasing the rate of genetic drift leading to significantly different genotypic distributions. Sensitivity of the results to selectivity of the harvest, as well as a range of carrying capacities, size of the nadir, and non-random mating are examined in further detail using a simpler form of the simulation previously described by Ripley *et al.* (2006) and presented in Martien *et al.* (2007).

While this particular simulation models a single stock, a two-stock version would be similar in structure, but would require several important additions. In addition to all the complexities of the population dynamics where the stock identity of the historical catch is unknown, genetic simulations require establishing the pre-whaling genetic conditions. These initial conditions result from the relative abundances and degree of gene-flow between the two populations. To assess whether simulated cases matched the empirical cases, empirical samples and their simulated equivalents would need to be assigned to stock, introducing further uncertainty.

Acknowledgments

The authors would like to thank Allan Strand for the design of and continued support with *rmetasim* throughout this project, Ryan Huebinger and John Bickham for the microsatellite genotypes, and the SWFSC Population Identity Program's Molecular Genetics Lab for the mtDNA sequencing. Janet Lowther assisted by troubleshooting STRUCTURE results. Judy Zeh and June Morita were invaluable in compiling and interpreting age estimates and providing thoughtful insights on estimating age from morphometric characters. Funding for authors came from the National Marine Fisheries Service and the North Slope Borough.

References

Angliss, R.P., Rugh, D.J., Withrow, D.E. and Hobbs, R.C. 1995. Evaluations of aerial photogrammetric length measurements of the Bering-Chukchi-Beaufort Seas stock of bowhead whales (*Balaena mysticetus*). *Rep. int. Whal. Commn* 45:313-24.

Baddeley, A. and R. Turner. 2005. Spatstat: an R package for analyzing spatial point patterns. Journal of Statistical Software. 12:6, 1-42.

Bockstoce, J.R. 1986. Whales, Ice and Men. University of Washington Press. Seattle, WA. 400 pp.

Bockstoce, J.R. and J.J. Burns. 1993. Commercial whaling in the north Pacific sector. Pp. 563-577 in "The Bowhead Whale" (J.J. Burns, J.J. Montague, and C.J. Cowles, eds.). Society for Marine Mammalogy, Special Publication Number 2.

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.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. Classification and Regression Trees. Chapman and Hall, New York. 358pp.

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

Chessel, D., A.-B. Dufour, and J. Thioulouse. 2004. The ade4 package-I- One-table methods. *R News* 4:5-10.

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.

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/M07/AWMP submitted to the Annual Meeting of the Scientific Committee of the International Whaling Commission, May, 2007. 28pp.

Goudet, J., M. Raymond, T. De Meeüs, and F. Rousset. 1996. Testing differentiation in diploid populations. Genetics 144:1933-1940.

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.

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. 5pp.

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.

Martien, K.K., E. Archer, B.J. Ripley, and B.L. Taylor. 2007. The genetic consequences of non-equilibrial dynamics in bowhead whales. Paper SC/59/BRG16 submitted to the annual meeting of the Scientific Committee of the International Whaling Commission, May, 2007.

Morin, P.A, R.G. LeDuc, E. Archer, K. K. Martien, B.L. Taylor, R. Huebinger, J.W. Bickham. 2007. Estimated genotype error rates from bowhead microsatellite data. Paper SC/59/BRG15 submitted to the annual meeting of the Scientific Committee of the International Whaling Commission, May, 2007. 11pp.

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.

Ripley, B.J., K.K. Martien and B.L. Taylor. 2006. A simulation approach to understanding non-equilibrial 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.

Roff, D.A. and P. Bentzen. 1989. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Molecular Biology and Evolution 6:539-545.

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.

Suydam, R. S. and George, J. C. Subsistence harvest of bowhead whales (Balaina mysticetus) by Alaskan Eskimos, 1974-2003. SC/56/BRG12.

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.

Therneau, T.M. and B. Atkinson. 2006. rpart: Recursive Partitioning. R package version 3.1-34.

Table 1. Demographic parameters at carrying capacity (λ=1.00) and near zero population size (λ=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).

	$\lambda = 1.00$					$\lambda = 1.042$				
Stage	***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
F	50	0.978	0.011	0.967	0.011	50	0.985	0.004	0.981	0.004
M	50	0.978	0.011	0.967	0.011	50	0.985	0.004	0.981	0.004

Table 2. Probability of assignment to age bins for leaves of the CART tree in Figure 2. Bins are inclusive of the lower boundary.

	Age Bin									
Leaf	**< 3**	**3 - 5**	**5 - 10**	**10 - 18**	**18 - 26**	**26 - 37**	**37 - 50**	**50 - 60**	**60 - 90**	**≥ 90**
I	0.84	0.11	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
II	0.00	0.86	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00
III	0.21	0.21	0.36	0.21	0.00	0.00	0.00	0.00	0.00	0.00
IV	0.00	0.06	0.88	0.00	0.00	0.00	0.06	0.00	0.00	0.00
V	0.00	0.00	0.36	0.57	0.00	0.00	0.07	0.00	0.00	0.00
VI	0.00	0.00	0.33	0.00	0.50	0.17	0.00	0.00	0.00	0.00
VII	0.00	0.00	0.00	0.33	0.17	0.50	0.00	0.00	0.00	0.00
VIII	0.00	0.00	0.00	0.13	0.53	0.27	0.00	0.00	0.07	0.00
IX	0.00	0.00	0.00	0.00	0.00	0.53	0.29	0.18	0.00	0.00
X	0.00	0.00	0.00	0.00	0.00	0.14	0.43	0.29	0.00	0.14
XI	0.00	0.00	0.00	0.11	0.00	0.00	0.22	0.11	0.22	0.33
XII	0.00	0.00	0.00	0.00	0.00	0.11	0.11	0.11	0.67	0.00
XIII	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.17	0.17	0.50

Table 3. Results of comparisons of observed F_{st}, χ^2, and Φ_{st} values with distributions from 500 replicates of the simulation. Permutation *p*-values are from 1000 permutations. Simulation *p*-values are proportion of replicates with test statistics ≥ than empirical. Shaded boxes indicate *p*-values ≤ 0.05. The comparisons where the empirical are inconsistent with the simulated are the shaded boxes in the columns for "simulation p-value".

A) mtDNA

	F_{st}			χ^2		Φ_{st}		
Strata	**observed**	**permutation *p*-value**	**simulation *p*-value**	**permutation *p*-value**	**simulation *p*-value**	**observed**	**permutation *p*-value**	**simulation *p*-value**
Barrow (258) v SLI (52)	-0.002	0.697	0.772	0.774	0.802	-0.003	0.584	0.486
Barrow-F (133) v Barrow-S (125)	0.000	0.344	0.406	0.562	0.586	0.002	0.219	0.230
SLI-F (13) v SLI-S (11)	0.054	0.066	0.008	0.194	0.672	-0.014	0.503	0.492
before 1918 (8) v 1918-1949 (13)	-0.010	0.530	0.640	0.310	0.848	-0.035	0.703	0.598
before 1918 (8) v 1950-1979 (25)	-0.013	0.670	0.740	0.669	0.530	-0.026	0.672	0.610
before 1918 (8) v after 1979 (34)	0.001	0.366	0.410	0.031	0.236	0.001	0.339	0.360
before 1950 (21) v 1950-1979 (25)	-0.007	0.605	0.728	0.814	0.844	0.005	0.308	0.284
before 1950 (21) v after 1979 (34)	0.009	0.205	0.186	0.010	0.186	0.012	0.214	0.242
1918-1949 (13) v 1950-1979 (25)	-0.010	0.654	0.720	0.976	0.982	0.007	0.315	0.328
1918-1949 (13) v after 1979 (34)	0.010	0.259	0.186	0.038	0.448	0.011	0.268	0.308
1950-1979 (25) v after 1979 (34)	0.007	0.218	0.216	0.090	0.312	-0.015	0.762	0.658

B) microsatellites

	F_{st}				χ^2		
Strata	**observed**	**permutation *p*-value**	**simulation *p*-value (no errors)**	**simulation *p*-value (w/ errors)**	**permutation *p*-value**	**simulation *p*-value (no errors)**	**simulation *p*-value (w/ errors)**
Barrow (213) v SLI (25)	0.002	0.076	0.042	0.040	0.073*	0.126	0.138
Barrow-F (115) v Barrow-S (98)	0.001	0.124	0.098	0.092	0.196	0.374	0.394
SLI-F (14) v SLI-S (11)	-0.007	0.947	0.988	0.986	0.972	1.000	1.000
before 1918 (5) v 1918-1949 (9)	0.002	0.418	0.378	0.384	0.689	0.992	0.996
before 1918 (5) v 1950-1979 (16)	-0.011	0.912	0.976	0.962	0.971	1.000	1.000
before 1918 (5) v after 1979 (24)	0.002	0.406	0.356	0.348	0.346	0.956	0.952
before 1950 (14) v 1950-1979 (16)	-0.004	0.791	0.844	0.840	0.758	0.998	1.000
before 1950 (14) v after 1979 (24)	0.002	0.293	0.272	0.256	0.044	0.494	0.510
1918-1949 (9) v 1950-1979 (16)	0.001	0.404	0.368	0.390	0.348	0.954	0.968
1918-1949 (9) v after 1979 (24)	0.003	0.315	0.240	0.248	0.051	0.446	0.478
1950-1979 (16) v after 1979 (24)	-0.003	0.793	0.858	0.866	0.227	0.888	0.902

*Givens *et al.* (2007) found significant differentiation for this stratification using Fisher's exact test which has been shown to have more power than permutation-based F_{st} or χ^2 tests (Goudet et al. 1996).

Table 4. Summary of STRUCTURE results. For each statistic, the p-value represents the proportion of the simulated datasets for which the value of the statistic was greater than or equal to the value observed for the empirical dataset.

Summary statistic	**Observed value**	***p*-value (no errors)**	***p*-value (with errors)**
Variance across individuals in logit-transformed assignment probability	1.65	0.116	0.140
F_{ST} between groups	0.006	0.330	0.292
R^2 for regression of Barrow spring samples	0.026	0.136	0.126
Nearest-neighbor correlation coefficient	0.092	0.372	0.332
Area between 3-phase step function and horizontal line	4.47	0.402	0.462

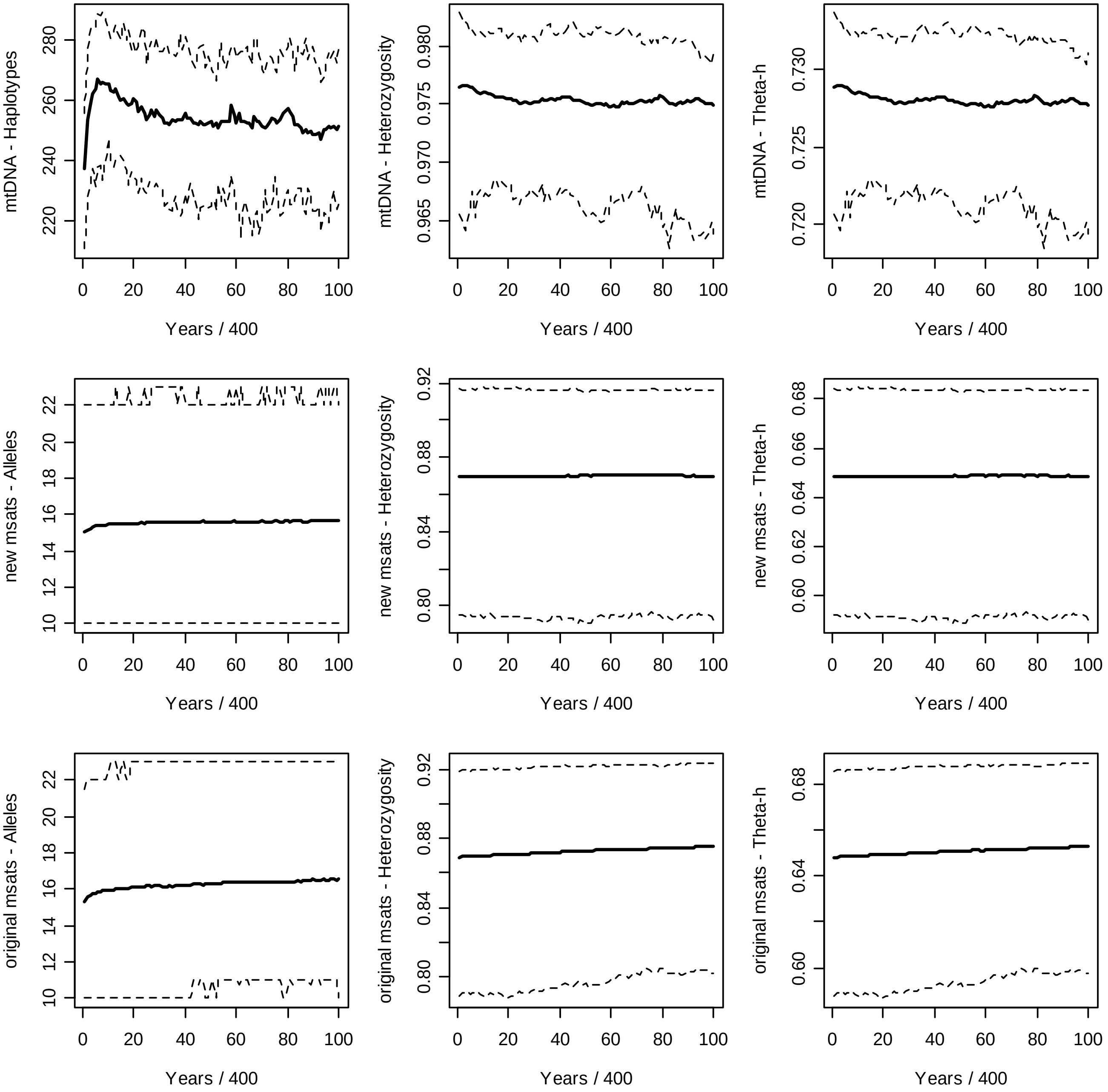


Figure 1. Mean values (solid lines) and 95-percentiles (dashed lines) for number of haplotypes (mtDNA) and alleles (microsatellites), heterozygosity, and Theta-h during 50 burn-ins.

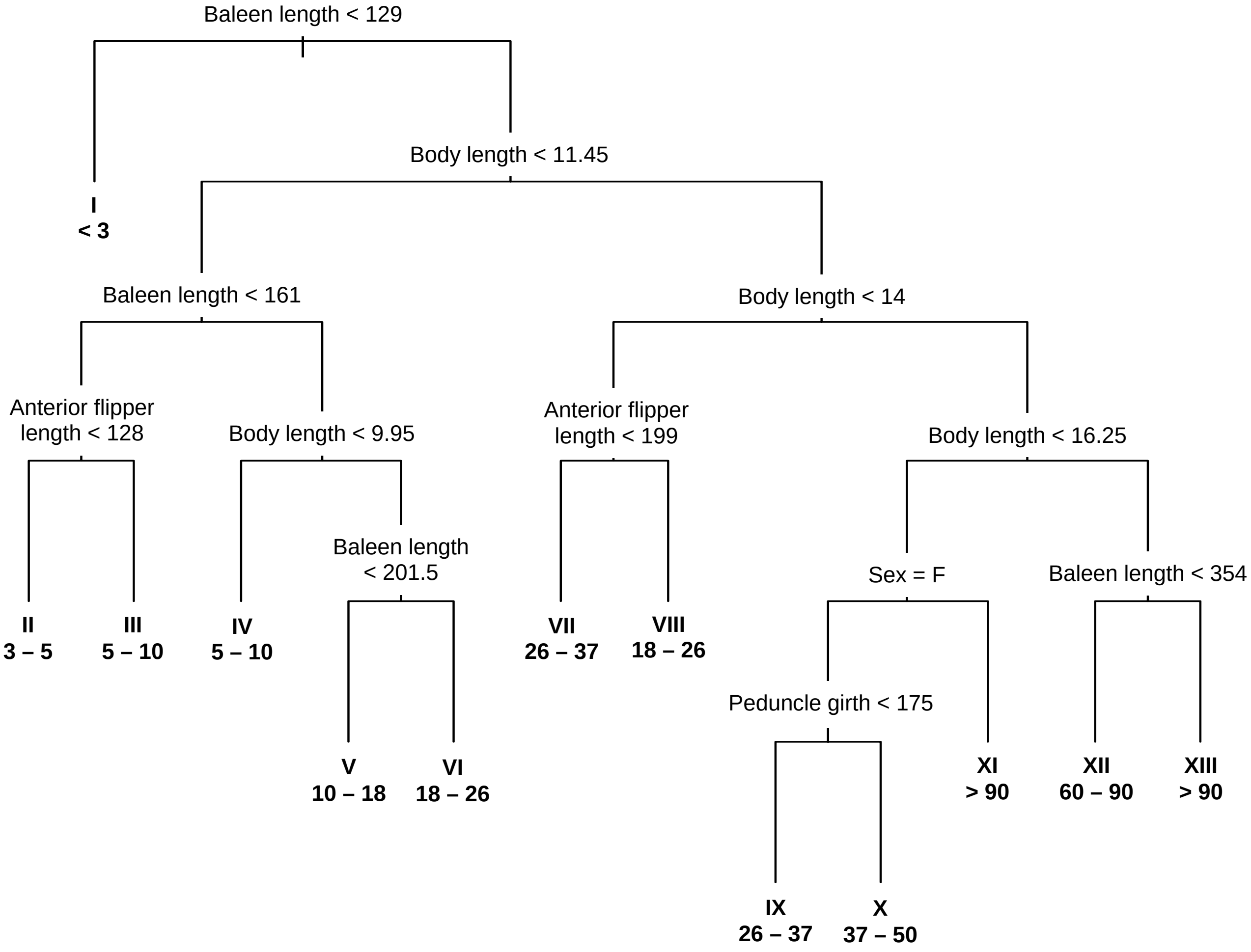


Figure 2. CART tree with primary splits used for age estimation. Cases meeting the criteria at a node are sent down the left. Roman numerals are leaf identifiers corresponding to rows in Table 2. Values below leaf identifiers are estimated age bin of leaf.

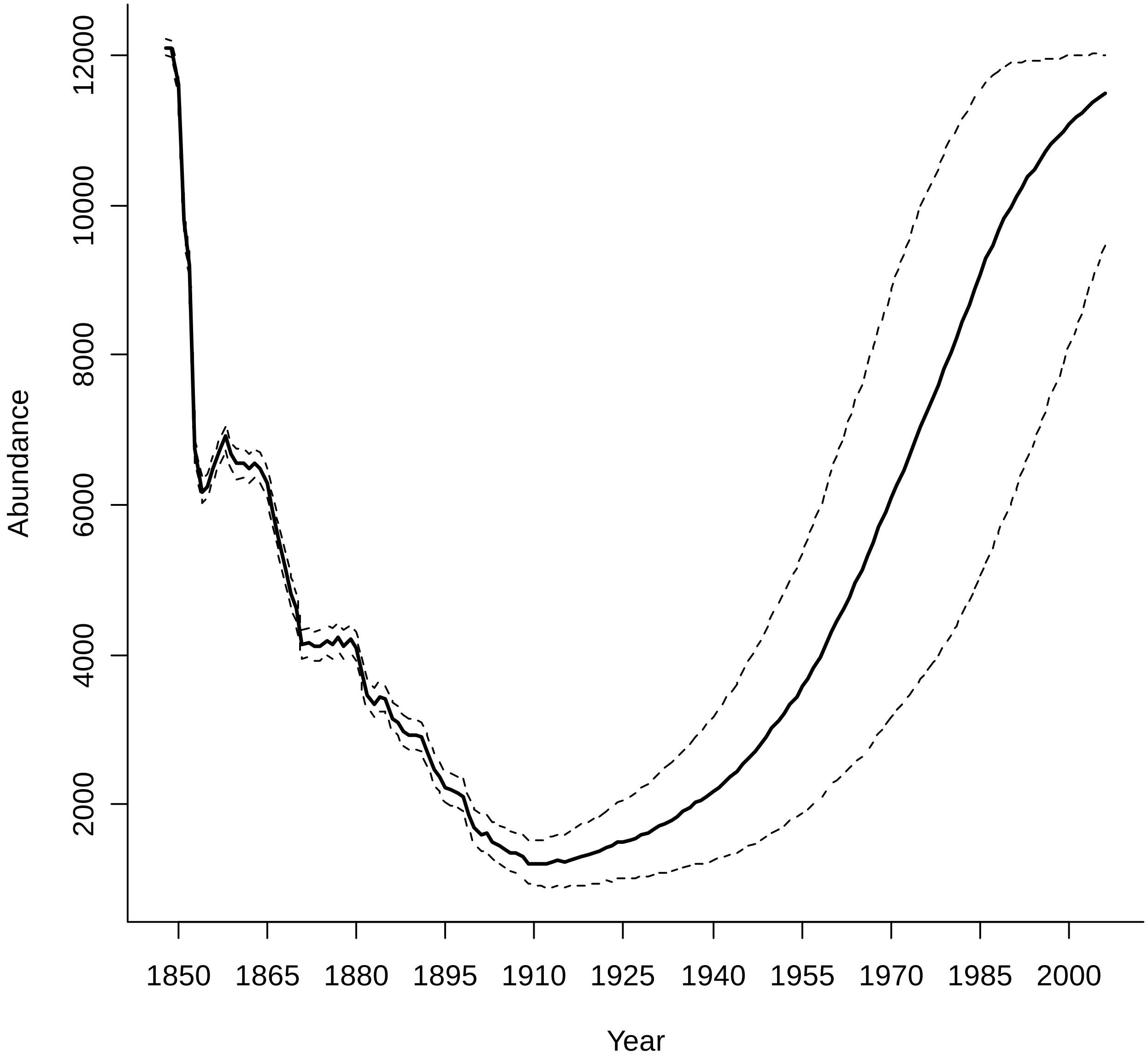
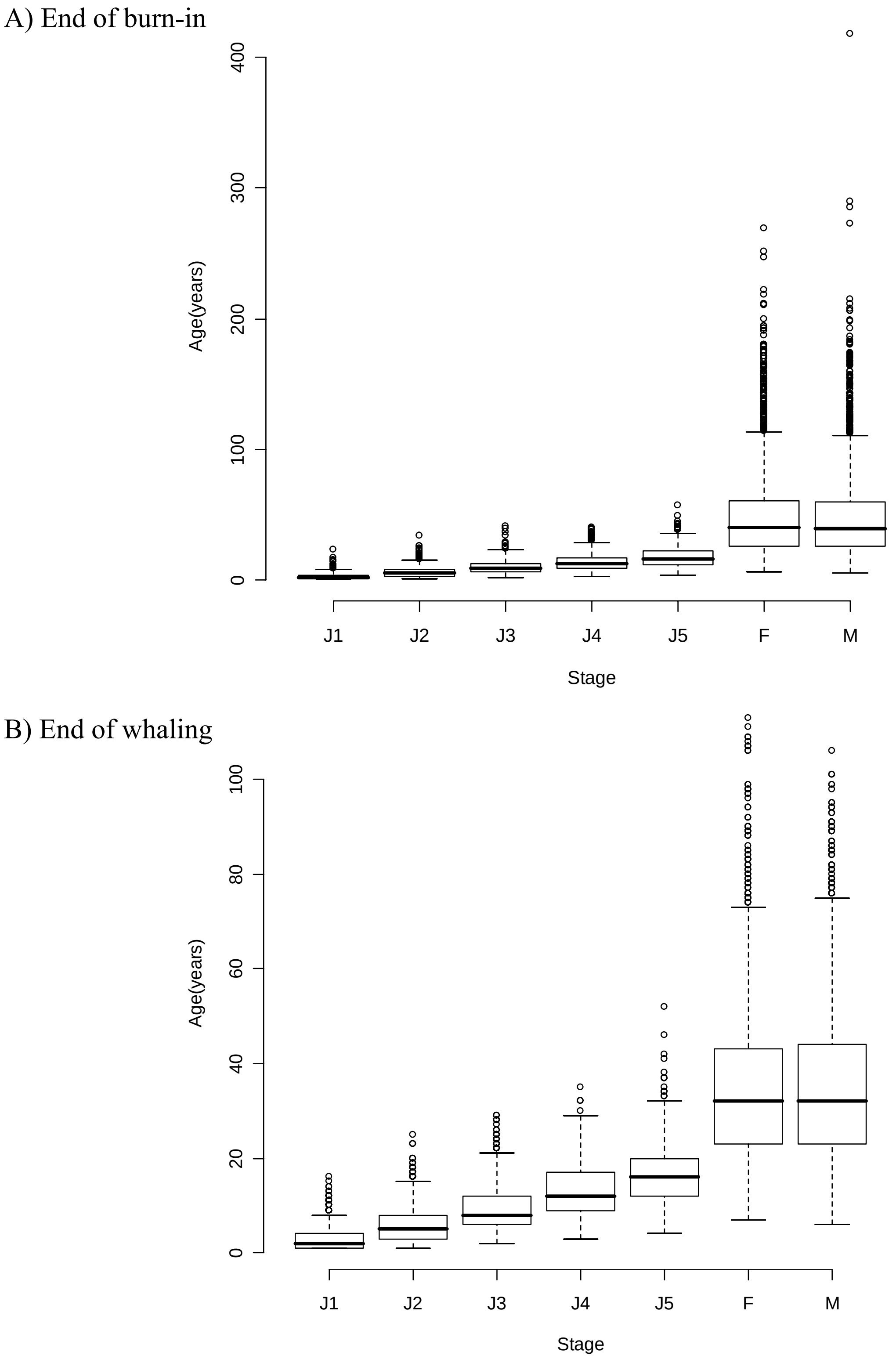


Figure 3. Median population abundance for 500 simulation replicates from 1848 to 2006. Dashed lines bound the 95-percentile of abundance in each year.



A) End of burn-in

B) End of whaling

Figure 4. Distribution of age within each demographic stage for all simulated individuals at the end of burn-in (A), and the end of whaling (B), for an example simulation replicate.

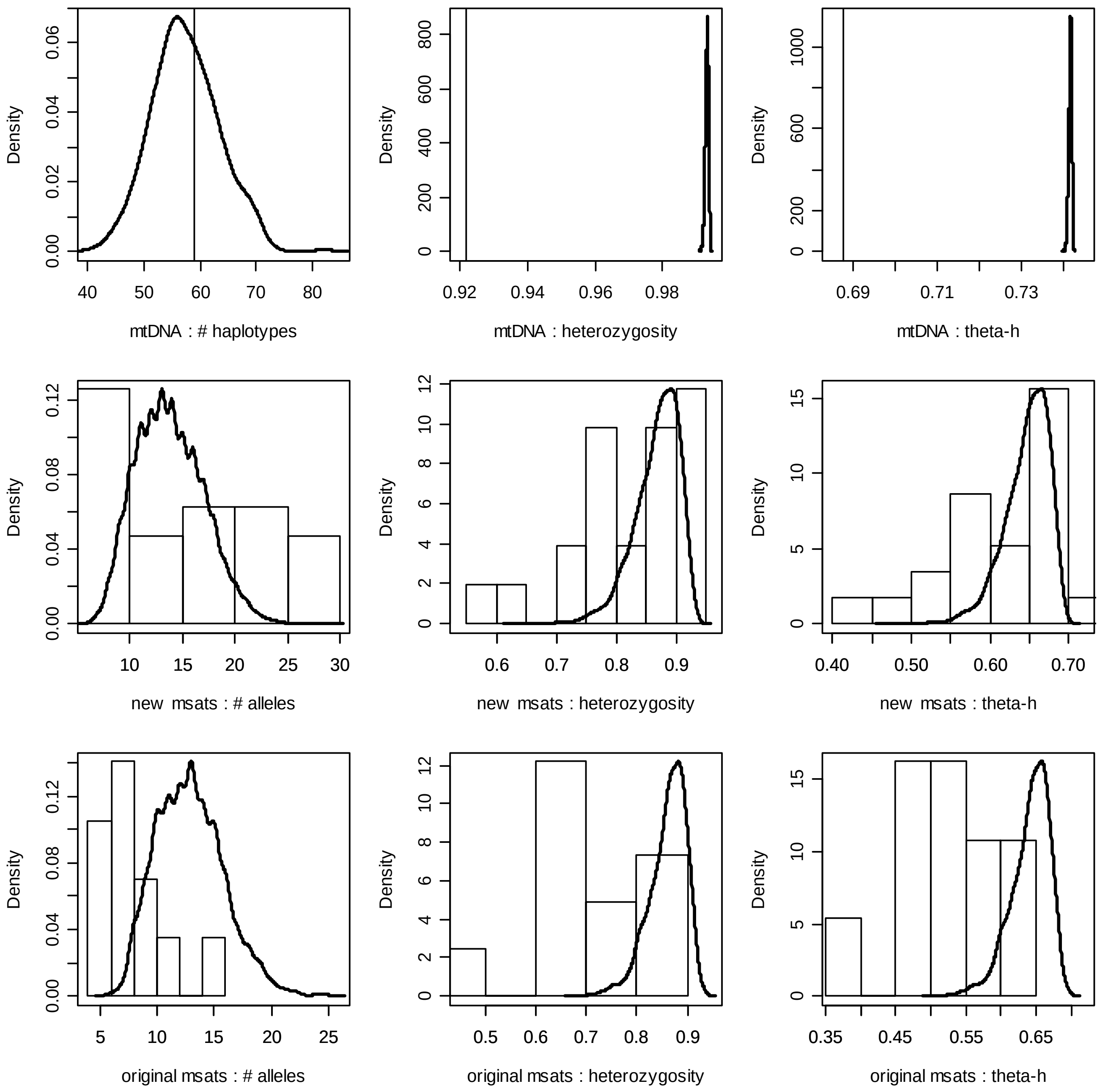
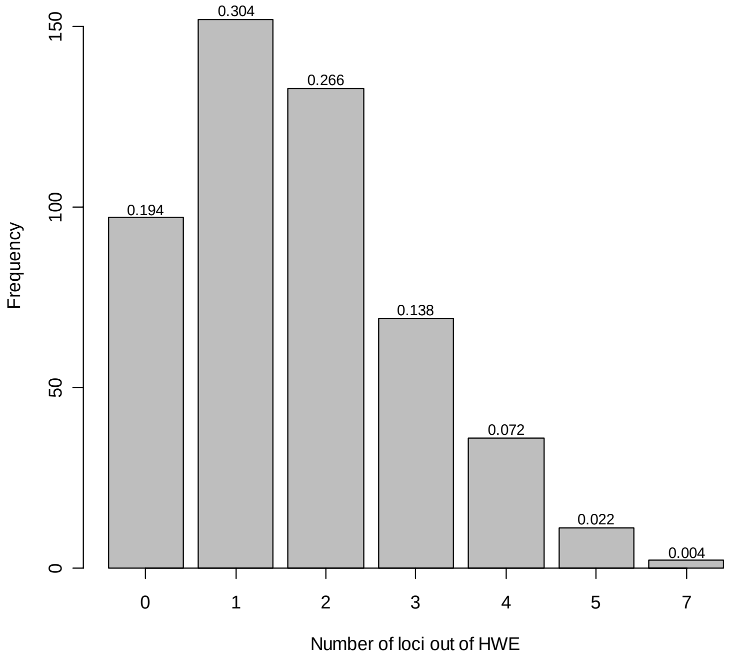


Figure 5. Distribution of single-locus measures of genetic diversity in the empirical data (histograms over the 33 loci), and 500 replicates of the simulation (bold lines). For mtDNA, the empirical data value is given by a single line.

A) Without errors



B) With errors

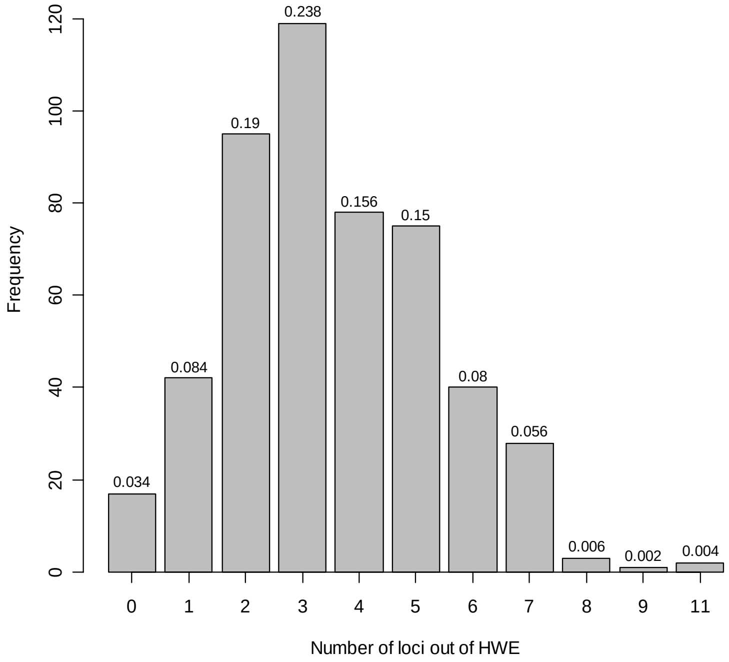


Figure 6. Distribution of the number of loci out of Hardy-Weinberg equilibrium in the 500 replicates without (A) and with (B) errors included. Numbers above the bars are the fraction of the total number of replicates represented by that bar.

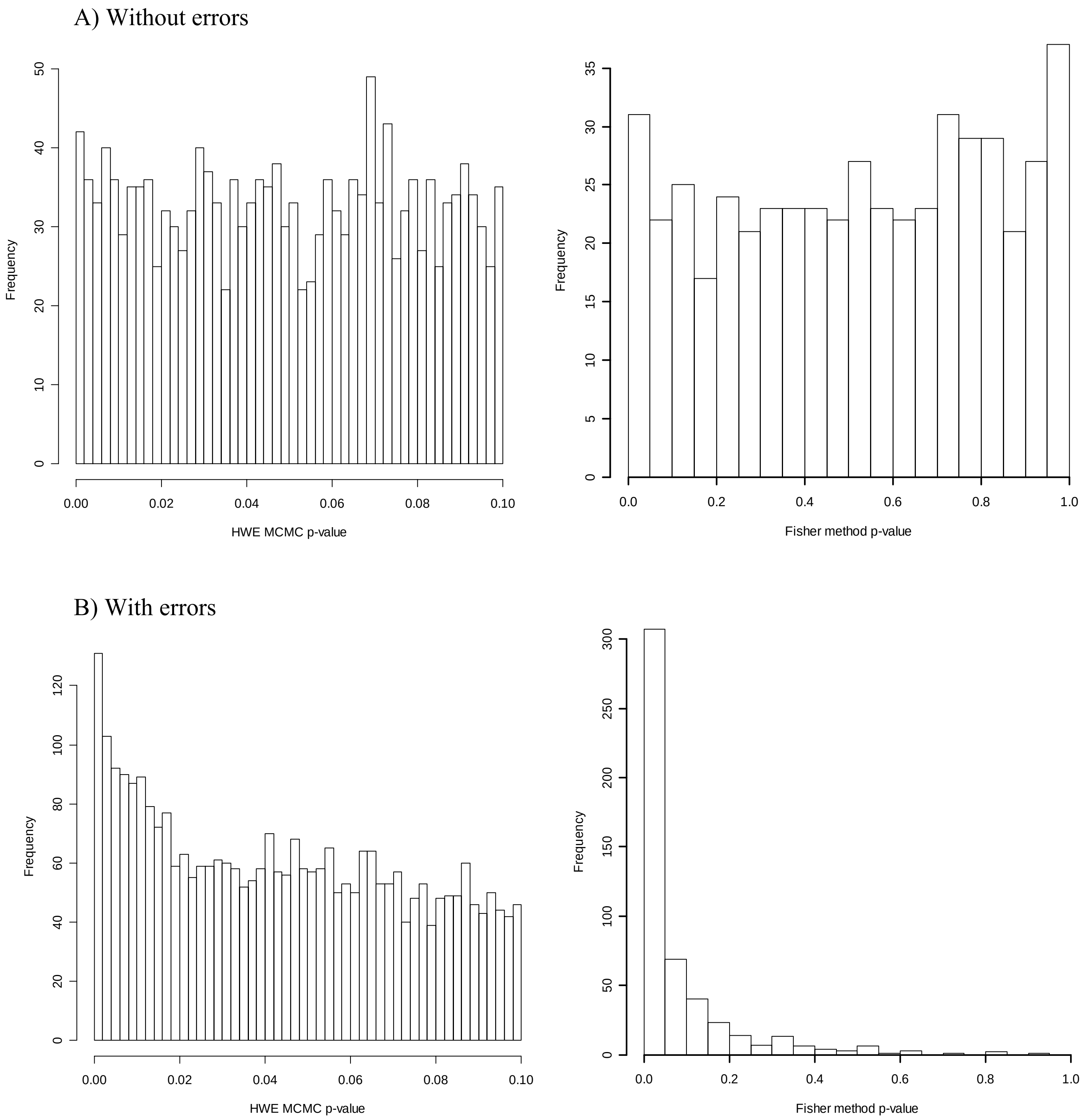


Figure 7. Distribution of *p*-values for Hardy-Weinberg equilibrium (HWE) without (A) and with (B) errors included. Figures on left are distributions of locus *p*-values from Genepop (truncated to values ≤ 0.1). Figures on right are distributions of overall *p*-value using Fisher's method.

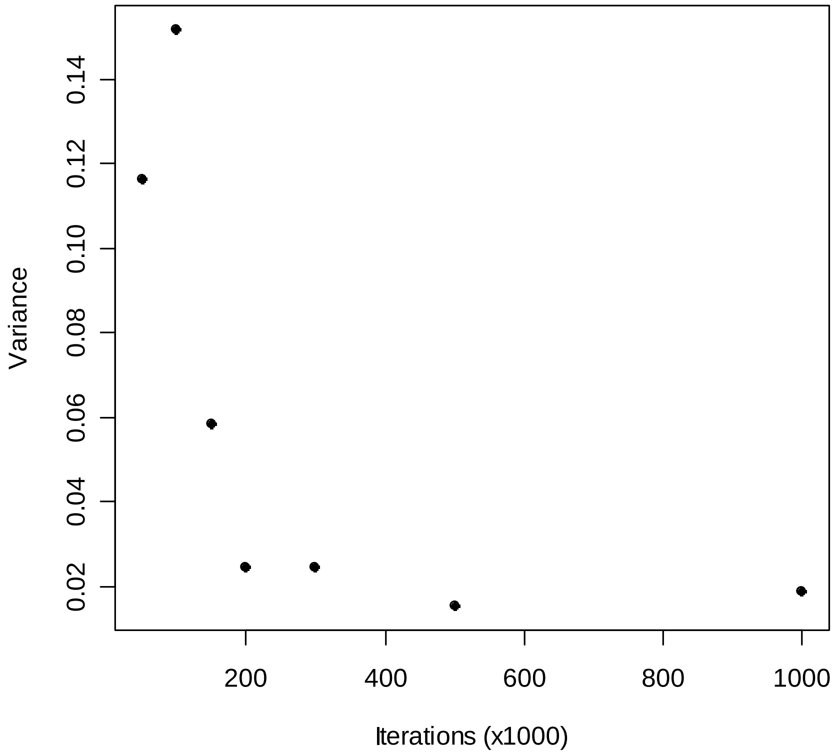


Figure 8. Variance in the logit-transformed assignment probability of individuals as a function of the number of iterations used in the STRUCTURE analyses. Each point was obtained by calculating the variance for each sample across ten replicate analyses, and then averaging across samples.

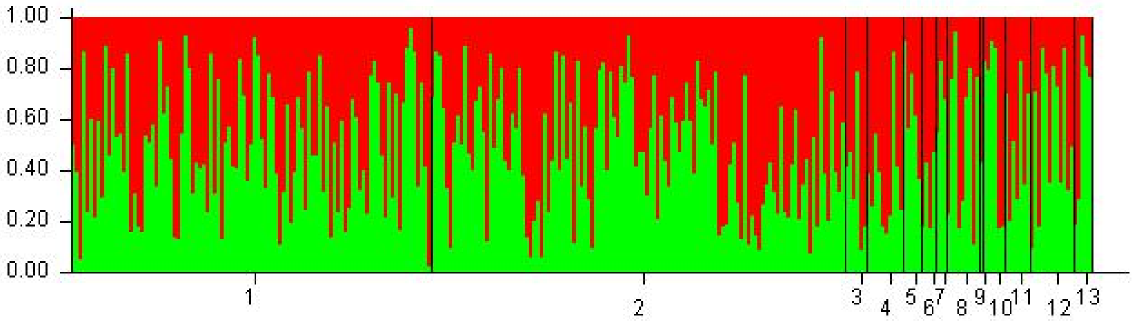


Figure 9. STRUCTURE results for k=2 for the empirical dataset. This analysis included all 33 loci, but excluded samples from Canada and the Sea of Okhotsk. Only samples from the 33 locus reference set (i.e., those with genotypes for at least 30 out of 33 loci) were included. Strata are numbered as in Givens et al. (2007): 1=spring Barrow, 2=fall Barrow, 3=spring Savoonga, 4=fall Savoonga, 5=spring Gambell, 6=fall Gambell, 7=spring Chukotka, 8=fall Chukotka, 9=Diomede, 10=Point Hope, 11=Wainwright, 12=Kaktovik, 13=Nuiqsut.

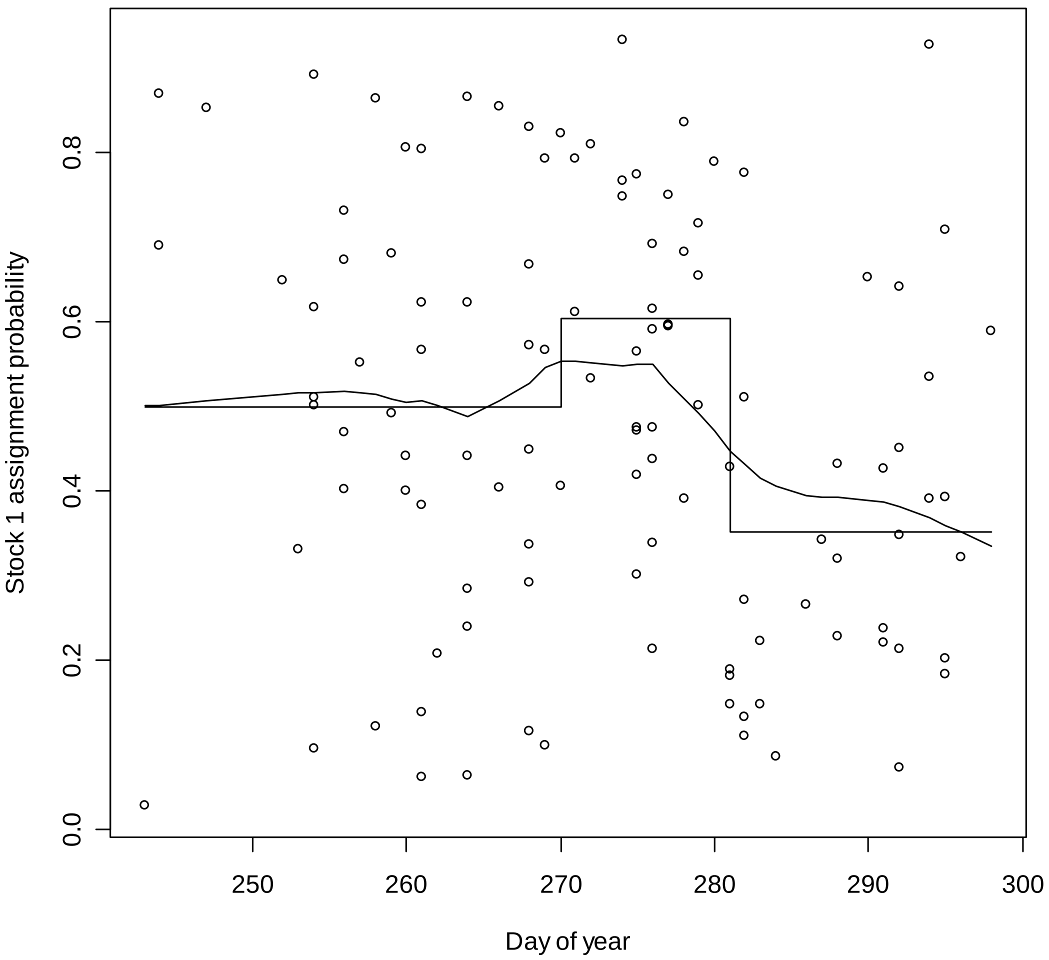


Figure 10. A three-phase step function fit to the plot of day of year versus red ancestry for the empirical data. A smooth fit is also overlaid on the plot for comparison to the smooth fit shown in Givens *et al.* (2007).

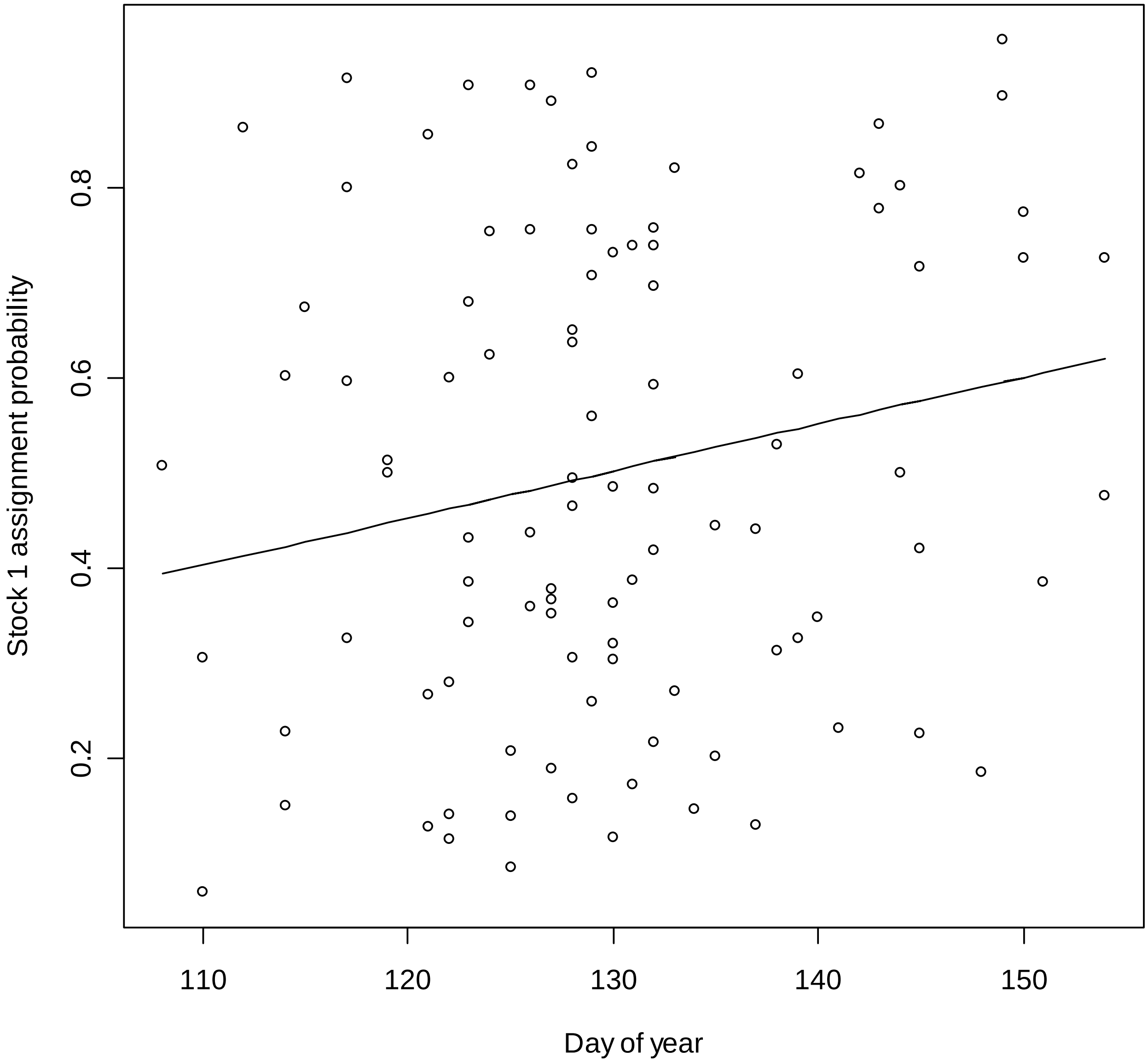


Figure 11. Back-transformed linear regression fit for logit (red ancestry) versus capture date for spring Barrow samples from the empirical dataset. The slope of the regression was not statistically significant (*p* = 0.111).