A Brief and Preliminary Look at SNPs Data for some Bering-Chukchi-Beaufort Seas Bowhead Whales

Geof Givens[1] and Marisa Williams[1], Phillip A. Morin[2], Brittany Hancock[2], J. Craig George[3]

March 3, 2007

Abstract

We present preliminary analyses of 18 single nucleotide polymorphism markers (SNPs) for 106 bowhead whales from Barrow and St. Lawrence Island. We find no evidence for disequilibrium, population substructure, or genetic variation associated with temporal spacing of whales in the migration.

Introduction

We present here a preliminary examination of a small dataset of single nucleotide polymorphism markers (SNPs) obtained from samples from 126 bowhead whales (Balaena mysticetus) from the Bering-Chukchi-Beaufort Seas region. These data were developed by Morin et al. (2007), using the methods of Morin and McCarthy (2007). Primary analyses of these data are given by Morin et al. (2007); we offer only some preliminary secondary examinations here. The main goal of our paper is to provide a first look at one analysis not provided by those authors related to a temporal lag autocorrelation feature found in microsatellite analysis of fall Barrow samples (Jorde et al., 2007; Givens and Ozaksoy, 2006; Givens et al. 2007). We also confirm some basic findings of Morin et al. (2007) using a slightly different set of samples and SNP loci.

We analyzed data for 18 SNP loci from 106 bowhead whales. The samples were collected from three sites: Barrow and the villages of Savoonga and Gambell on St. Lawrence Island (SLI). The Barrow data were all collected during the spring migration period. The Savoonga and Gambell data came from whales sampled during both the fall and spring migrations, as well as from baleen samples and whale skulls obtained from villagers and discards from historical aboriginal hunts.

Some adjustments were made to the raw data prior to analysis. In the original dataset there were 126 whales typed on 19 loci. First, 16 individuals were deleted from the dataset because they were found to be duplicates of whales already in the dataset. Second, two fetuses from Barrow and 2 fetuses from Savoonga were deleted because data from their mothers were already in the dataset. Third, one locus (Bmys402M56) was deleted from the data set because it was sex-linked. Table 1 shows the numbers of whales analyzed using the 18 remaining SNPs loci.

---

[1] Dept. of Statistics, Colorado State University, Fort Collins CO 80523.
[2] Southwest Fisheries Science Center, 8604 La Jolla Shores Dr., La Jolla CA 92037.
[3] North Slope Borough Dept. of Wildlife Management, PO Box 69, Barrow AK 99723.

Table 1: Number of whales analyzed.

	Fall migration	Spring migration	Baleen	Skull	Total
Barrow	0	56	0	0	56
Savoonga	10	6	3	0	19
Gambell	3	4	2	19	31
Total	13	66	5	19	106

Among the remaining SNPs loci, one locus was virtually monomorphic, and two had minor allele frequencies less than 5%. The rest had minor allele frequencies between 10% and 47%. The highest number of missing genotypes for an individual was 3, and the largest number of missing genotypes for a locus was 10 (8.1%), with the next highest only 4 (3.3%).

Methods

Tests for Hardy-Weinberg disequilibrium and linkage disequilibrium were conducted for the Barrow and the combined Barrow/SLI data using GENEPOP (Raymond and Rousset, 2004). The Hardy-Weinberg test relies on the method of Rousset and Raymond (1995); we used 1000 batches of 1000 iterations each with a burn-in of 10000. Inferences across all loci were carried out by pooling p-values using Fisher's method (Fisher, 1948). Linkage disequilibrium was also examined using GENEPOP. Comparisons of allele frequencies were made using GENEPOP using the method of Guo and Thompson (1992) and the same Monte Carlo simulation parameters as above. Again Fisher's method was used to pool results across loci. Estimates and 95% confidence bounds for F_{st} were computed in FSTAT (Goudet, 1995) using the approach of Weir and Cockerham (1984). Confidence intervals for the estimate of F_{st} were done by bootstrapping over loci with 500 permutations. Finally, the analysis approach developed by Givens and Ozaksoy (2006) was adjusted for SNPs data and used to examine questions of temporal pulses in the spring Barrow migration.

Results

We found no significant evidence of disequilibrium using the Hardy-Weinberg test (p=0.75 for Barrow samples alone, p=0.21 for all data pooled). Locus Bmys1R248 was not included in this analysis because of insufficient variability.

Testing for linkage disequilibrium using only the Barrow samples, we found only 6/153 (3.9%) of the locus pairs to appear linked at the nominal p-value of 0.05. Pooling the Barrow and SLI data, only 9/153 (5.9%) of the locus pairs met this threshold. Since one would expect about 5% of locus pairs to present false positive evidence of linkage, our findings do not provide convincing evidence of linkage.

Table 2 shows the results of our tests for allele frequency differences. We compared Barrow to SLI, along with various comparisons among strata defined by village and season. For the tests not including Barrow, loci Bmys1R248 and Bmys48S269 were not included in the analysis because these loci were not variable within SLI samples. Clearly there is no significant difference in allele frequencies for the different sub-groups.

Table 2 also shows F_{st} estimates and 95% confidence intervals for the same comparisons. No evidence of differences between strata is seen.

Table 2: P-values for tests for allele frequency differences, and F_{st} estimates and confidence intervals.

	Barrow vs. SLI	Barrow vs. Savoonga	Barrow vs. Gambell	SLI spring vs. SLI fall	Savoonga spring vs. Savoonga fall
Allele freq. difference p-value	0.954	0.998	0.719	0.977	0.978
F_{st} estimate	-0.004	-0.009	0.002	-0.014	-0.014
95% confidence bounds	(-0.008, -0.000)	(-0.013, -0.005)	(-0.006, 0.010)	(-0.085, 0.057)	(-0.085, 0.057)

Finally, we applied the method developed by Givens and Ozaksoy (2006) to determine whether any substructure could be detected and attributed to time lag between migrating whales for the Barrow samples. Because of the small sample sizes, the default approach of limiting comparisons to whale pairs with both individuals caught in the same year produces unreliable results. The fit is unreliable due to overfitting. Furthermore, the stratified resampling within catch year has only limited effectiveness at producing a suitable null distribution. Therefore, the data were analyzed using two resampling strategies: both with and without stratification by year.

The results are shown in Figures 1 and 2. In these plots, X is the difference in capture time between two paired whales. If there is no relationship between X and allele match probability, then the fitted black curve should be flat and contained within the dotted 95% null bands. Furthermore, if there is no evidence of unexplained excess allele matching, the fitted black curve should lie near the flat black reference line.

The result of the stratified analysis (Figure 1) shows a wild fit most likely caused by spline overfitting. (For example, the number of whales per stratum is as low as 1.) We believe these results should be discarded until more data become available. Without stratification (Figure 2), the fit is nearly flat and stays within the confidence bounds. This is suggestive of no evidence for temporal pulsing, but power is extremely low and further weakened by the non-stratified approach of allowing comparisons across years.

Conclusion

Our preliminary examinations found no evidence of spatial or temporal stock structure. The lack of evidence of population substructure may reflect reality, or it may be the result of low statistical power. As more SNPs data become available (both more markers and more whales), we may be better able to provide independent confirmation or refutation of some of the findings from microsatellites and mtDNA that have thus far been the main genetic basis for inference about bowhead whale stock structure.

Acknowledgments

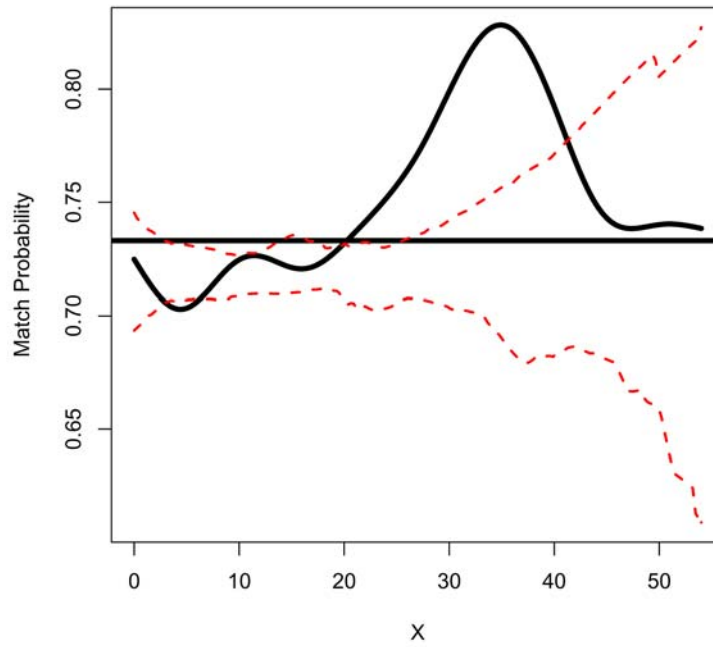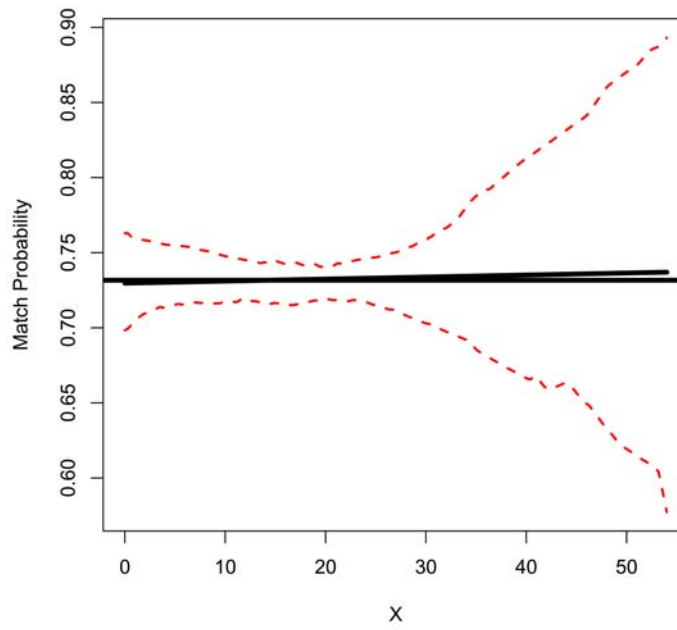**Figure 1: Analysis with time lag covariate and stratified permutation approach.**

Figure 2: Analysis with time lag covariate and non-stratified permutation approach.

References

Fisher, R. A. (1948) "Combining independent tests of significance", *American Statistician*, vol. 2, issue 5, page 30.

Givens, G. H., Ozaksoy, I. (2006) Population structure and covariate analysis based on pairwise microsatellite allele matching frequencies. Paper SC/58/SD3 presented to the Scientific Committee of the International Whaling Commission, May 2006

Givens, G. H., Huebinger, R. M., Bickham, J. W., George, J. C., Suydam, R. (2007) Patterns of genetic differentiation in Bowhead whales (Balaena mysticetus) from the Western Arctic. Paper SC/M07/AWMP1 presented to the IWC Scientific Committee Standing Working Group on the Development of an AWMP, March 2007

Gou, S. W., Thompson, E. A. (1992) Performing the exact test of Hardy-Weinberg proportions for multiple alleles. *Biometrics*, **48**, 361-372.

Goudet, J. (1995). FSTAT version 1.2: a computer program to calculate Fstatistics. *Journal of Heredity*, **86**(6): 485-486.

Jorde, P. E., Schweder, T., Stenseth, N. C. (2007) "Detecting genetic structure in migrating bowhead whales off the coast of Barrow, Alaska", *Molecular Ecology*: to appear

Morin, P. A., Hancock, B., and George, J.C. (2007) Development and application of single nucleotide polymorphisms (SNPs) for bowhead population structure analysis. Paper SC/59/XXXX presented to the Scientific Committee of the International Whaling Commission, May 2007.

Morin, P.A. and McCarthy, M. (2007) Highly accurate SNP genotyping from historical and low-quality samples. *Molecular Ecology Notes*, in review.

Raymond, M., Rousset, F. (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Heredity*, **86**:248-249

Rousset, F., Raymond, M. (1995) Testing heterozygote excess and deficiency. *Genetics*, **140**, 1413-1419.

Weir, B. S., Cockerham, C. C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.