SC/66a/BRG/12

Continued development of MTDNA and SNP databases for bowhead whales

Amy B. Baird, Alesha R. Rimmelin, Christina Q. Tran, John C. George, Robert S. Suydam, and John W. Bickham

INTERNATIONAL WHALING COMMISSION

CONTINUED DEVELOPMENT OF MTDNA AND SNP DATABASES FOR BOWHEAD WHALES

Amy B. Baird[1], Alesha R. Rimmelin[1], Christina Q. Tran[1], John C. George[2], Robert S. Suydam[2], and John W. Bickham[3]

[1]Department of Natural Sciences, University of Houston – Downtown, 1 Main St., Houston, TX 77002, USA; [2]North Slope Borough, Department of Wildlife Management, Barrow, AK 99723, USA; [3]Battelle Memorial Institute, Houston, TX, USA

ABSTRACT

This paper details the current status of ongoing genetic studies of bowhead whales. Our research has two main objectives: building a mitochondrial DNA (mtDNA) database and the development of a single nucleotide polymorphism (SNP) panel for bowheads. Here we report the current status of the mtDNA database, as well as the pros and cons of several methods of SNP genotyping we tested on bowheads. Our mtDNA database currently includes sequences from 711 whales (447 complete cytochrome-b sequences, 427 complete ND1 sequences, and 638 HVR1 sequences). Of these, 345 whales have been sequenced for all three loci. We report on the methods and criteria used to select a panel of 96 bowhead SNPs, which currently are being analyzed.

INTRODUCTION

Genetic data have played an important role in bowhead whale conservation, management and in establishing an aboriginal quota through the IWC to meet nutritional and cultural needs of Inuit people in Alaska, USA, and Chukotka, Russia. The long-term acquisition of mitochondrial DNA (mtDNA) data has been integral in studies examining stock structure and also provides a long term perspective on population size estimates as a means of estimating effective population size, N_e. Initially, studies on bowhead mtDNA focused on a single locus, the control region (e.g. Rooney et al. 2001, LeDuc et al. 2005). Recent studies have expanded the mtDNA sequencing to two additional loci: ND1 and cyt-b (Phillips et al. 2012). The expansion of the mtDNA database has increased the number of variable, informative sites, which has allowed more sophisticated analyses to be conducted, with more accurate results obtained. This, in turn, has shed more light on the evolutionary history and population structure of bowhead whales. Specifically, Phillips et al. (2012) used the 3 mtDNA loci to examine historical demography of bowheads to investigate the impact of commercial whaling and concluded that no genetic signature of recent population depletion was observed. The same data were able to show evidence, however, for historical population expansion and subsequent decline due most likely to Pleistocene climate change. The analyses using data from the 3 mtDNA loci provided a much more accurate picture of historical relative population size estimates than similar studies using only the mtDNA control region (Rooney et al. 1999 and Rooney et al. 2001).

Moreover, nuclear data are an important source of genetic information to answer questions that mtDNA alone cannot address, such as genetic interchange between populations, and more accurate estimates of relatedness between individuals, and population subdivision. Previous studies of bowheads have utilized different nuclear markers to address these questions. In the past, microsatellites were used with good success (LeDuc et al. 2005, Givens et al. 2010); however, these markers can be problematic in several ways. Microsatellites are considered to have a high error rate and are not always replicable between labs. More recently, single nucleotide polymorphisms (SNPs) have been used on bowheads with good success. When directly compared to microsatellite data in bowheads, SNPs were shown to be successful at answering questions about population size and structure (Morin et al. 2012). Morin et al. (2012) concluded that SNPs are the preferred method for genotyping bowheads because of their ability to be replicated across labs, their ability to be used on old or degraded samples, and their cost effectiveness.

This project has two objectives. The first is to continue to build a mtDNA database that includes sequences from three genes. We continue to sequence the commonly used mtDNA control region, as well as two protein coding genes (cytochrome-b and ND1). The data presented here build upon mtDNA sequence data presented in SC/64/AWMP9 (Bickham et al. 2012) and SC/65b/BRG13 (Baird et al. 2014).

The second focus of this project is on the continued development of a SNP panel. Previous studies have reported on the identification of these variable loci for bowheads, including SC/65a/BRG22 and SC/65b/BRG13. The most recent of these papers described the 155 SNP loci that have been identified for bowheads. Here, the goal was to select a panel of 96 of those previously identified loci and determine the most cost-effective method to analyze them for large sample sizes of whales.

The combination of mtDNA and SNP data, along with the genetic resources database provided by the bowhead genome project (Seim et al., 2014; Keane et al., 2015), will allow us to continue addressing issues related to bowhead stock structure (such as estimating potential genetic interchange between BCB and Canadian or Sea of Okhotsk populations), revisit issues of historical demography and molecular ecology, and continue to monitor the genetic health of the BCB bowhead population.

METHODS

New mtDNA sequence data acquisition. – Bowhead whale DNA was extracted, amplified, and sequenced following the methods of LeDuc et al. (2008) and Phillips et al. (2011). The hyper variable region-1 (HVR1) region of the mtDNA control region was sequenced (397bp) along with the complete cytochrome-b (cytb) gene (1140bp) and the complete ND1 gene (957bp). Individual sequences were compared to existing haplotypes from previous studies deposited on GenBank and haplotype codes were assigned.

Single nucleotide polymorphism (SNP) selection. – We examined all 155 SNP loci that were reported in previous papers to determine the best loci to use in a SNP panel. Our evaluation of each locus included noting position of the SNP in the genome; proximity of a given SNP to any other known SNPs; proximity of the SNP to repeated genetic elements that might make amplification difficult; and whether the SNP was a base substitution or an insertion/deletion.

Because some of our overall goals with the bowhead genetics program are to estimate current and historical effective population size and examine whether there is genetic interchange between stocks, a priority was to find SNP markers that would inform these questions. In combination with our mtDNA data, some of the most informative markers for these questions will be SNPs from the sex chromosomes. Therefore, we chose all usable variable loci on the X chromosome (12 loci) and Y chromosome (1 locus).

Next, we wanted to minimize the linkage between any of the autosomal SNPs that we chose, and we therefore prioritized SNPs that were not reported from the same amplicons in the papers that originally described the SNP loci.

Selection of SNP method of analysis.—We tested three different methods of SNP genotyping and considered a fourth method without testing it. The methods tried included TaqMan genotyping, High Resolution Melting analysis and the Fluidigm SNPtype assay. We considered the Illumina MiSeq method, but it was ruled out before we attempted to use it (see below). To select which method to use, we considered several factors: cost per sample; ease of data interpretation; reliability upon replication; and labor intensity.

TaqMan genotyping was performed using an Eppendorf real time thermal cycler using 6 loci described by Morin et al. (2010).

The High Resolution Melting analysis was performed using the same thermal cycler on 4 loci.

Fluidigm SNPtype assay primers were designed for 96 SNPs using a combination of loci described in Morin et al. (2010), SC/65b/BRG13 (Baird et al. 2014), and SC/65a/BRG22 (Bickham et al. 2013).

RESULTS

Mitochondrial DNA database. – To date, we have samples representing 711 whales. Sequences in the database include data from 447 whales with complete cyt-b sequences, 427 whales with complete ND1 sequences, and 638 whales with complete HVR1 sequences. Of these, 345 whales have been sequenced for all three loci. The majority of our samples come from the BCB stock, but we also have representatives from the Eastern Canadian Arctic and Sea of Okhotsk.

Selection of SNP method of analysis.—We initially considered the Illumina MiSeq method of SNP genotyping (http://www.illumina.com/). This method is a type of next-generation sequencing that can produce 2x300 bp paired-end reads in one run, allowing for the collection of

a large amount of data in a single run. For our purposes, we would initially need to design primers for each SNP locus that would be compatible with the MiSeq platform. Depending on the service company, it would be possible to design primers for up to 90 SNPs that could all be run on a single run for 96 samples. The benefits of this method include the ability to run a large number of samples/SNPs at a time and that many service companies have the ability to use this platform. The major downside of this method was the cost. For the development of ~90 primers there would be a one-time cost of $15,000. Then one would need to amplify each sample for each primer pair, resulting in significant costs for consumables. Finally, each miSeq run of 90 primers x 96 samples costs approximately ~$3500 (depending on the service company).

We then considered the Life Technologies TaqMan genotyping method http://www.lifetechnologies.com/us/en/home/life-science/pcr/real-time-pcr/real-time-pcr-assays/snp-genotyping-taqman-assays.html). This method is a real-time PCR-based method that works by using allele-specific primers each tagged with a unique fluorescence. Upon annealing and extension of an allele-specific primer, it (or they, in the case of a heterozygote) will fluoresce. We initially tested the method on 6 loci described in Morin et al. 2010. The advantages of this method for our purposes included the ability to run each locus individually, so it was not a large cost burden up front. We could also do all of the work in-house because we own a real-time PCR machine. However, we had several reservations about the feasibility of this method for our purposes. Although the up-front costs per locus were not burdensome (each custom primer pair cost $242.50 which was enough product to run 1500 samples), when summed for the number of loci needed to produce a usable panel of SNPs for a large number of samples, as well as costs associated with real-time PCR enzyme master mixes, the costs for this method would quickly accumulate. More worrisome for us was that we could not reliably replicate our data with this method. We are not completely sure of the cause of this, but we suspect it had more to do with the real-time PCR machine we were using (Eppendorf) rather than the TaqMan technology itself. Because we were trying to detect single base-pair changes, the machine needs to be very sensitive to those changes in fluorescence. We believe that our machine was unable to accurately distinguish between different bases at a given SNP in many instances.

Next, we tested another real-time PCR-based method: high resolution melting analysis. We used the SensiFAST HRM Kit from Bioline for our test runs (http://www.bioline.com/us/sensifast-hrm-kit.html). This method works by amplifying the SNP target (and surrounding sequence), then slowly melting the amplicon. Amplicons with different sequences will melt at slightly different temperatures. The kit cost is quite modest (~$100), and has enough to run 200 reactions, or a larger kit for $1050 (for 2000 reactions). The kit can then be used with normal primers, which are significantly cheaper than fluorescent primers. The major benefits of this method included a much lower cost than the methods we tested, and it could be done in-house with our real-time PCR machine. The downside was that, again, we could not reliably replicate the results. We again suspect that the problem may have had more to do with our PCR machine than with the chemistry/technology of the method.

Finally, after testing the other methods and consulting with colleagues who have also searched for a cost-effective and reliable method for SNP genotyping, we settled on the Fluidigm SNPtype assay. This method is based on allele-specific PCR SNP detection. It requires a Fluidigm platform to run/read the assay, requiring samples to be run at a service facility. The method is quite cost effective, fast, and results are reproducible (according to colleagues who have tested it extensively). The method can handle multiplexing SNP loci in groups of 96 or 192 for 96 samples at a time. Each set of 96 loci costs just under $5000 and includes enough product to run about 14,000 samples, so this cost is a one-time expense. Then, each run of 96 samples (for all 96 loci) costs approximately $1300 (depending on the service company). There is essentially no labor involved if they are run by a service company, other than extracting and quantifying the DNA.

Selection of SNP loci.—After having settled on the Fluidigm method, the most cost effective approach was to design primers for 96 loci of the 155 SNPs identified for bowheads. The first step was primer design (described below) and quality control.

Fluidigm's primer design software required having a minimum of 60bp sequence both upstream and downstream of the desired SNP. Indels could not be greater than 10bp. A desired SNP could also not be within 30bp of another variable locus. Using these criteria, along with our own criteria for our desired SNP loci (see Methods), we designed primers using the Fluidigm software.

First, we acquired all available sequence data surrounding the SNPs of interest. These data were acquired from GenBank (for loci described in Morin et al. 2010) or from positional information reported in previous papers and aligned to the bowhead whale genome (Keane et al. 2015; http://www.bowhead-whale.org/). Several of the 155 available SNPs were described from single amplicons, which placed some SNPs close to one another (sometimes less than 30bp apart). To avoid linkage between autosomal SNPs, we first only selected one SNP per amplicon. The flanking sequence was then submitted to Fluidigm for quality control.

A small number of our initially submitted loci were rejected by Fluidigm based on quality control concerns. These included a target close to an area of "repeats and/or difficult genomic diversity" and indels being too complicated. To then get to the target of 96 loci, this required us to select some SNPs that would have otherwise not met our criteria, particularly regarding the concern over linkage. We ended up selecting 9 autosomal loci that were located in proximity to other SNPs in the panel. In total, we selected 13 X- and Y-chromosome SNPs and 83 autosomal SNPs.

DISCUSSION

The continued expansion of a mitochondrial DNA database for bowhead whales has been, and will continue to be, a valuable tool in analyzing population size and structure for this species. The inclusion of two protein-coding genes has proven to add considerable value to the mtDNA

database, rather than including only the HVR1 region of the control region (Bickham et al. 2012). We are continuing to expand this database, with the eventual goal of having all samples sequenced for all three loci.

Having selected a cost-effective and reliable method for SNP analysis will also result in valuable data to inform our analyses of population and evolutionary genetics of BCB bowheads, as well as examining the potential for genetic interchange between BCB and other populations. These data will be important for having a firm understanding for what stock boundaries are and whether there is any gene flow between them. This is clearly an important issue for management and conservation in this species.

The 96-locus SNP panel we designed has utilized the best data and ascertainment schemes we have available to date. The 42 SNPs identified by Morin et al. (2010) were obtained through a random locus approach and a targeted gene approach using bowhead samples from eastern Russia, St. Lawrence Island, Barrow, Alaska and eastern Canada. The sex chromosome SNPs from SC/65a/BRG22 (Bickham et al. 2013) were obtained using a targeted gene approach (exon-primed intron crossing) from Alaskan whales. The remaining autosomal SNPs described in SC/65b/BRG13 (Baird et al. 2014) came from examining variable sites in the bowhead transcriptome, which was based on multiple individuals of Alaskan whales. Therefore, our subset of 96 SNPs contains loci identified by a variety of methods and from a comparison of whales from multiple populations.

SNP ascertainment schemes can significantly affect conclusions regarding population history (McTavish and Hillis 2015). This has long been known to be true of other sources of population genetics data, including microsatellites (Barbara et al., 2007; Putman and Carbone, 2014). Biased selection of SNP loci can affect estimates of F_{ST}, population admixture, and population histories. Ascertainment schemes can be potentially biased toward samples from a geographic region (which would overestimate polymorphisms in that region) or toward loci polymorphic in multiple groups. Simulated data presented in McTavish and Hillis (2015) demonstrated that ascertainment schemes inflating geographic polymorphisms led to decreased differentiation among populations using F_{ST}. When the ascertainment bias favors shared polymorphisms among multiple groups, population differentiation was obscured.

Keeping in mind the above caveats regarding ascertainment bias, we sought to include SNPs identified from multiple populations whenever possible. A total of 33 of the 96 SNPs in our panel come from the Morin et al. (2010) study which used whales from multiple populations for SNP identification. The remaining SNPs come from studies which used only individuals from the BCB population. To improve this potential bias toward BCB polymorphisms, we plan to develop a second 96-SNP panel that will be derived from bowhead genome/transcriptome data containing data from both Alaskan and Greenland whales (Keane et al. 2015; http://www.bowhead-whale.org/).

ACKNOWLEDGMENTS

We thank the Alaska Eskimo Whaling Commission (AEWC) and the 11 villages of the Whaling Captains' Associations for their confidence, guidance, and support of our research. We gratefully acknowledge funding provided by the North Slope Borough Department of Wildlife Management and the many people who helped collect the tissue samples for this work. We thank Donald Stoeckel and Andrew DeWoody for sharing their experience with different methods of SNP genotyping.

LITERATURE CITED

Baird, A. B., R. S. Suydam, J. C. George, and J. W. Bickham. 2014. Update on mtDNA and SNP database for bowhead whales. Paper SC/65b/BRG13 submitted to the International Whaling Commission Scientific Committee.

Bickham, J. W., R. M. Huebinger, C. D. Philips, J. C. Patton, L. D. Postma, J. C. George, and R. S. Suydam. 2012. Assessing molecular substitution patterns in mitochondrial control region compared to protein coding genes in bowhead whales: update of SC/63/BRG13. Paper SC/64/AWMP9 submitted to the International Whaling Commission Scientific Committee.

Bickham, J. W., G. W. Stuart, H. K. Downing, J. C. Patton, J. C. George, and R. S. Suydam. 2013. Comparison of methods for molecular assessment of sex chromosome polymorphisms and levels of genetic divergence in the bowhead whale. Paper SC/65a/BRG22 submitted to the International Whaling Commission Scientific Committee.

Barbara T, Palmer-Silva C, Paggi GM, Bered F, Fay MF, et al. (2007) Cross-species transfer of nuclear microsatellite markers: potential and limitations. Mol Ecol 16: 3759–3767.

Givens, G. H., R. M. Huebinger, J. C. Patton, L. D. Postma, M. Lindsay, R. S. Suydam, J. C. George, C. W. Matson, and J. W. Bickham. 2010. Population genetics of bowhead whales (*Balaena mysticetus*) in the western Arctic. Arctic 63: 1-12.

Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand D, Marques PI, Michalak P, Kang L, Bhak J, Yim HS, Grishin NV, Nielsen NH, Heide-Jorgensen MP, Oziolor EM, Matson CW, Church GM, Stuart GW, Patton JC, George JC, Suydam R, Larsen K, Lopez-Otin C, O'Connell MJ, Bickham JW, Thomsen B, de Magalhaes JP (2015) Insights into the evolution of longevity from the bowhead whale genome. Cell Reports. 10, 112-120.

LeDuc RG, Martien KK, Morin PA, Hedrick N, Robertson KM, Taylor BL, Mugue NS, Borodin RG, Zelenina DA, Litovka D, George JC (2008) Mitochondrial genetic variation in

bowhead whales in the western Arctic. *Journal of Cetacean Research and Management* 10, 93–97

LeDuc, R. G., A. E. Dizon, A. M. Burdin, S. A. Blokhin, J. C. George, and R. L. Brownell, Jr. 2005. Genetic analysis (mtDNA and microsatellites) of Okhotsk and Bering/Chukchi/Beaufort Seas populations of bowhead whales. Journal of Cetacean Resource Management 7: 107-111.

McTavish, E. J., and D. M. Hillis. 2015. How do SNP ascertainment schemes and population demographics affect inferences about population history? BMC Genomics 16:266.

Morin, P. A., V. L. Pease, B. L. Hancock, K. M. Robertson, C. W. Antolik. 2010. Characterization of 42 single nucleotide polymorphism (SNP) markers for the bowhead whale (*Balaena mysticetus*) for use in discriminating populations. Marine Mammal Science 26: 716-732.

Morin, P. A., F. I. Archer, V. L. Pease, B. L. Hancock-Hanser, K. M. Robertson, R. M. Huebinger, K. K. Martien, J. W. Bickham, J. C. George, L. D. Postma, B. L. Taylor. 2012. Empirical comparison of single nucleotide polymorphisms and microsatellites for population and demographic analyses of bowhead whales. Endangered Species Research 19: 129-147.

Phillips, C. D., T. S. Gelatt, J. C. Patton, and J. W. Bickham. 2011. Phylogeography of Steller sea lions: relationships among climate change, effective population size, and genetic diversity. Journal of Mammalogy 92:1091-1104.

Phillips, C. D., J. I. Hoffman, J. C. George, R. S. Suydam, R. M. Huebinger, J. C. Patton, and J. W. Bickham. 2012. Molecular insights into the historic demography of bowhead whales: understanding the evolutionary basis of contemporary management practices. Ecology and Evolution 1-20.

Putman, A. I. and I. Carbone. 2014. Challenges in analysis and interpretation of microsatellite data for population genetic studies. Ecology and Evolution 4: 4399-4428.

Rooney, A. P., R. L. Honeycutt, S. K. Davis, J. N. Derr. 1999. Evaluating a putative bottleneck in a population of bowhead whales from patterns of microsatellite diversity and genetic disequilibria. Journal of Molecular Evolution 49: 682-690.

Rooney, A. P., R. L. Honeycutt, and J. N. Derr. 2001. Historical population size change of bowhead whales inferred from DNA sequence polymorphism data. Evolution 55: 1678-1685.

Seim, I., S. Ma, X. Zhou, M. V. Gerashchenko, S.-G. Lee, R. Suydam, J. C. George, John W. Bickham, and V. N. Gladyshev. 2014. The transcriptome of the bowhead whale *Balaena mysticetus* reveals adaptations of the longest-lived mammal. Aging 6:879-899.