Assessing molecular substitution patterns in the mitochondrial control region compared to protein coding genes in bowhead whales.

John W. Bickham¹, Ryan M. Huebinger², Caleb D. Phillips³, John C. Patton¹, John C. George⁴ and Robert S. Suydam⁴

¹Center for the Environment and Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907, USA; ²Department of Surgery, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA; ³Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409; ⁴North Slope Borough, Department of Wildlife Management, Barrow, AK 99723, USA

Abstract

This paper reports new data on substitution patterns in three mitochondrial genes, control region, cytochrome b, and ND-1 from 296 bowhead whales representing three populations; the Bering-Chukchi-Beaufort Seas (BCB) stock, the eastern Canadian arctic (Canada) stock and the Sea of Okhotsk (Okhotsk) stock. We used a previously described method to identify recurrent substitutions in the hypervariable region I (HVRI) of the control region and in two protein coding genes. The sequence matrix was extended to include the multiple substitutions at sites as indicated through character mapping of the neighbor-joining phylogeny. This results in a fully resolved haplotype network and allows for improved phylogenetic and phylogeographic analysis. An estimation of mtDNA mutation rate in bowheads (2.8% per million years) is reduced relative to most other whales. But, bowheads have maintained a relatively high female effective population size and the estimated time to most recent common ancestor of the mtDNA is 1.16 million years. Calculations of Fst and migration estimates were performed for the three stocks of bowhead whales. Canada and the BCB did not have a statistically significant Fst and the estimated number of migrants, 46 per generation, is consistent with previous studies. Okhotsk had a significant Fst with both BCB and Canada. The Fst between Okhotsk and BCB reported here for the 3-gene sequence is about double a previous report based on control region alone. Migration estimates are low between Okhotsk and both BCB (3.41 migrants per generation) and Canada (2.87). Tests for neutrality differed between HVRI and the protein coding genes with both of the latter showing evidence of selection or a rapid population expansion. This result was likely due to a lack of resolution in HVRI due to recurrent substitions.

Introduction

This paper is an update of SC/61/BRG7 including additional analyses and the addition of new sequence data.

The hypervariable region (HVRI) of the control region of the mitochondrial DNA is one of the most frequently used markers in studies of mammalian population genetics and evolutionary genetics. In studies of cetaceans, it is frequently used to address questions of stock structure, gene flow, genetic diversity, effective population size, evolutionary history and phylogeography.

The underlying reason that makes HVRI such an important genetic marker for population studies, its high mutation rate which gives rise to high haplotype diversity, also gives rise to its major limitation, homoplasy. It is widely known that site-specific rates of mutation are highly variable in HVRI and that recurrent mutations result in homoplasy that obfuscate accurate calculations of mutation rates and the recovery of accurate tree or network topologies. Nevertheless, this problem is generally ignored and, in most studies HVRI haplotypes are considered as characters identical by descent rather than state.

In contrast to HVRI, the mtDNA protein coding genes such as cytochrome b and nicotinamide adenine dinucleotide dehydrogenase 1 (ND1), which are also widely used in studies of mammalian evolution and population genetics, have much slower evolutionary rates, lower haplotype diversity, and typically lower levels of homoplasy. Because these genes are linked with HVRI on the non-recombining mtDNA molecule, and thus inherited as a single locus, parallel comparisons of HVRI to the protein coding genes offer an opportunity to investigate how the substitution processes of HVR1 influence patterns of haplotype diversity and distribution. Moreover, the combination of rapidly evolving HVRI with more conservative protein coding genes plus the identification of homoplasies can lead to highly resolved networks or trees from which significant correlations of maternal lineages with geographic patterns can be obtained (Phillips et al., 2009, in press).

The goals of this study were to investigate site specific mutation rates in bowhead whales (Balaena mysticetus), to estimate the degree to which variation in these rates causes homoplasy, to present haplotype networks of one, two and three genes concatenated to illustrate the degree to which resolution can be achieved, to compare diversity estimates and test for neutrality among the different genes, and to re-evaluate migration rates among the BCB, eastern Canadian Arctic, and Sea of Okhotsk populations. Accurate estimation of mutation rate is key to the calculation of theta; $\Theta = 2Ne(f)\mu$ where Ne(f) is the effective female population size and μ is the mutation rate. Thus, for estimating long-term effective population size, a key element to designing many conservation programs, accurate mutation rate estimates are needed. Recently, three different approaches to the problem have been developed that are potentially applicable to bowhead whales. One is the use of Γ rate category assignment to estimate site-specific nucleotide substitution rates, a second is a tree weighting and substitution mapping procedure (Phillips et al., 2009), and the third method uses the substitution rate at a linked and more slowly evolving locus for calibration (Alter and Palumbi, 2009). In all cases we compare HVRI sequences with cytochrome b and ND1 sequences for the same individuals. Estimates of Θ are presented as a Bayesian skyline plot to show the evolutionary trends of female effective population sizes from time to most recent common ancestor to the present.

Only a few studies have attempted to calculate site-specific mutation rates at HVRI and to identify recurrent substitutions. In humans, both phylogenetic and familial based estimates of substitution rates were calculated (Parsons et al. 1997; Sigurðardóttir et al. 2000; Heyer et al 2001; Howell et al 2003; Henn et al 2009). For Steller sea lions, Phillips et al. (2009) calculated

SC/63/BRG13

site specific rates using the Γ rate category assignment method as well as presenting a new method based on tree weighting and substitution mapping. Alter and Palumbi (2009) developed the method using the substitution rate at a more slowly evolving locus (in this case cytochrome b) to estimate rates at HVRI in three species of cetaceans (gray, humpback, and Antarctic minke whales).

Methods

Sequence data acquisition

HVRI (397 base pairs), cytochrome b (1140 base pairs), and ND1 (957 base pairs) sequences were obtained from 296 bowhead whales using methods described in LeDuc et al. (2008) and Phillips et al. (in press).

Describing rates and patterns of substitution at HVRI

Site-specific rates of substitution were calculated from composite haplotypes and relative rates of substitution were calculated for each of the three genes. ModelTest 3.7 (Posada and Crandall, 1998) was used to calculate the best model of substitution and optimal number of discrete Γ rate categories using the Akaike Information Criterion and hierarchical likelihood ratio tests. Maximum likelihood tree construction was performed using TREE-PUZZLE (Schmidt et al., 2002) which uses the quartet puzzling algorithm employing 50,000 quartet puzzling steps. A python script was used for the extraction of Γ rate category assignments from the output of TREE-PUZZLE. Relative rates of substitution for the three genes was calculated from the ratio of the average rate for each gene because individual site rates within the genes are expressed as rates relative to the average rate of the combined genes.

A modified method of tree weighting and substitution mapping as developed by Phillips et al. (2009) was applied to the bowhead whale data set. However, evidence of homoplasy in the protein coding genes (not just in HVRI) was observed, which makes problematical their use in tree weighting. Therefore, coding regions were not weighted in this instance subsequent to tree building and substitution mapping. The sequence matrix was extended to express the occurrence of multiple substitutions at sites as indicated through character mapping of the resulting neighbor-joining phylogeny. This approach allowed for quantification of site-specific numbers of recurrent substitutions at all genes and a subsequent independent estimate of relative rates of substitution of the three genes. An additional method was used to estimate gene specific substitution rates (Alter and Palumbi, 2009) was employed. This method calculates the substitution rate per base at HVRI as $x/((1/2) \cdot w \cdot n)$. In this equation, x is the mean number of pairwise differences at HVRI for individuals identical at cytochrome b, w is the estimated waiting time until the next substitution at cytochrome b, n is the number of base pairs in the HVRI sequences used. To accomplish this, we calculated the synonymous pairwise distance (Li et al., 1985) among ten baleen whale species species for cytochrome b as done by Alter and Palumbi (2009). The silent substitution rate was then calculated as half the slope of the

SC/63/BRG13

relationship of the regression of synonymous pairwise distances against divergence times. Using that rate, w was calculated as $(\mu \cdot n)$ -1 where μ is substitutions per base per year and n is the number of fourfold degenerate sites plus one-third the number of twofold degenerate sites.

Divergence dating and Bayesian phylogeny estimation

The program BEAST was used to simultaneously estimate divergence time and phylogeny. The divergence time of 5.38 million years between bowhead whales and right whales (*Eubalaena*) was used to estimate divergence rates (McGowan et al., 2009). Initial divergence dating within the Balaena-Eubalaena lineage was performed using the two coding regions (cytb and ND1) with each partition receiving its own model of DNA evolution determined through model testing. In this analysis all codon positions were included. Program operations followed the author guidelines in all cases. Dates estimated for the time to most recent common ancestor and the next most basal set of divergences within the bowhead lineage were recorded and retained as node date priors. A second BEAST analysis was then performed to estimate dates of the more terminal nodes using only HVRI data. From the final tree file the maximum clade credibility tree was obtained using TreeAnnotator (part of the BEAST package). Next, a Bayesian skyline plot (BSP) which allowed 5 discrete changes in population size was constructed using standard MCMC sampling procedures to estimate posterior distributions of theta ($\theta = N_{ef}\tau$, where N_{ef} = female effective population size, and τ = generation length) through time. This analysis uses a flexible demographic model directly from the sample of gene sequences (Drummond et al. 2005). We allowed five demographic changes to identify historic demographic changes while not over-parameterizing the analysis. A bowhead whale generation time of 52 years was assumed for this estimation (Taylor et al., 2007).

Measures of neutrality

We analyzed measures of neutrality using Fu and Li's D* and F* statistics (Fu and Li, 1993) in the program DnaSP, Ver.5.10.01. Significant test statistics (P \leq 0.05) indicate non-neutral evolution at a locus, from which can be inferred selection at the locus or a linked locus, or a rapid population expansion. Tests were performed on HVRI (n = 55), cytochrome b (n = 56), ND1 (n = 58), and concatenated cytochrome b/ND1 sequences (n = 58).

Diversity estimates

DnaSP was used to calculate two estimates of diversity; gene diversity and nucleotide diversity. Test were performed with genes and sample sizes as described under Measures of neutrality.

Calculations of inter-stock migration rates

Concatenated sequences of the three mitochondrial regions including the entire cytochrome b gene, the entire ND1, and HVR1 were used to calculate migration rates between stocks of Bowhead whales. Sample sizes for stocks were as follows: BCB = 250, Canada = 39 and

Okhotsk = 7. Migration estimates (*M*) between stocks were calculated in Arlequin version 3.5 by assuming the two populations compared were of size *N* and drawn from a large pool of populations which exchange a fraction *m* of migrants each generation. The mutation rate μ is considered negligible as compared to the migration rate *m*, and following the simple relationship at equilibrium between migration and drift where,

 $F_{st} = 1/2M + 1$

Therefore, M, which is the absolute number of migrants exchanged per generation between the two populations, can be estimated by,

 $M = 1 - F_{st} / 2F_{st}$

Results and Discussion

Describing rates and patterns of substitution at HVRI

We applied two methods to estimate site specific mutation rates for three mitochondrial genes. Figure 1A shows site specific rate estimates for bowheads using the Γ rate category assignment method and Figure 1B shows the estimated number of recurrent substitutions per site based on a modification of the tree weighting and substitution mapping method (Phillips et al., 2009). This figure shows that considerably more recurrent substitutions occur in the control region than in the coding regions; 52 recurrent substitutions in HVRI region, and 35 in the two protein coding genes combined. In comparison, Phillips et al. (2009) observed no recurrent substitutions in cytochrome b (the only protein coding gene they studied) in Steller sea lions. It seems likely that the relationship between relative gene substitution rates and depth of the phylogeny contributes to the difference in patterns seen in these two species. That is, because of a much shorter evolutionary time frame in the Steller sea lion mtDNA phylogeny, insufficient time has elapsed for recurrent substitutions to become fixed in the more slowly evolving protein coding genes. Whereas in the more rapidly evolving Steller sea lion HVRI, 60 recurrent substitutions were observed.

We also applied the method of Alter and Palumbi (2009) to the bowhead dataset. Using this method we calculated the HVRI substitution rate for bowheads to be 2.8% per million years. This compares to the values reported by Alter and Palumbi (2009) of 5.4%, 5.2% and 5.0% per million years for gray, humpback and minke whales, respectively. Those values are considerably higher than the silent substitution rate for cytochrome b (1% per million years).

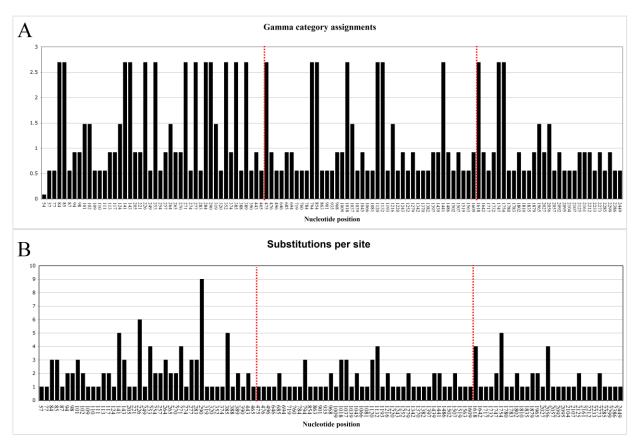


Figure 1. Estimated site-specific mutation rates (A) and calculated substitutions per site (B) for HVRI (left), cytochrome b (middle), and ND1 (right) variable sites for bowhead whales.

It is clear from the results of this study and that of Phillips et al. (2009) that the three methods to calculate site specific mutation rates and the relative rates of substitution of HVRI to protein coding genes are consistent within a species but vary considerably between Steller sea lions and bowhead whales, and between bowheads and other whales. The relative rates of substitution of HVRI to cytochrome b calculated by the Γ rate category assignment for Steller sea lions (Phillips et al, 2009) was 25.46 and the tree weighting and substitution mapping method yielded a value of 23.52. The Alter and Palumbi (2009) method yielded an estimated rate of 27.45% per million years for sea lion HVRI. All of these are in reasonably good agreement but differ markedly with values calculated for Bowhead whales. Using the Γ rate category assignment method the relative rates of substitution mapping method yielded a rate of 4.24. The rate of evolution of HVRI in sea lions is approximately 10 times the rate in bowheads and the rates for gray, humpback and minke whales are about double that of bowheads.

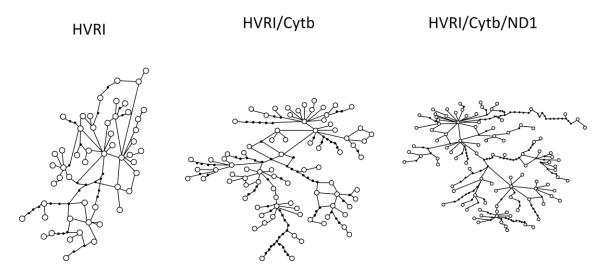


Figure 2. Haplotype networks for bowhead whales. HVRI is the control region alone, HVRI/Cytb is the control region plus cytochrome b and HVRI/Cytb/ND1 are all three gene sequences concatenated.

Accurate calculations of evolutionary rates and resolutions of haplotype networks provide increased resolution for estimators of population genetic parameters. Figure 2 shows the effect of the simple addition of sequence to resolving haplotype networks. Not surprisingly, as you go from HVRI alone (left panel) to HVRI plus cytochrome b, to HVRI plus cytochrome b plus ND1 the number of reticulations in the networks is reduced. However, this method alone does not fully resolve the concatenated 3-gene network. But an examination of Figure 3 shows that by extending the sequence to include the recurrent mutations identified by the tree weighting and substitution mapping method in Phillips et al. (2009) as additional characters the network becomes fully resolved.

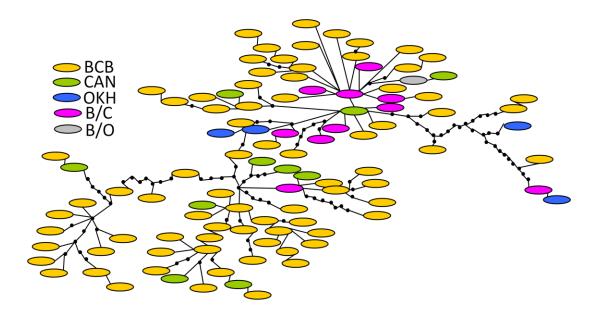


Figure 3. Fully resolved 3-gene concatenated sequence network resulting from the extension of the sequence to account for the multiple substitutions at hyper-variable sites.

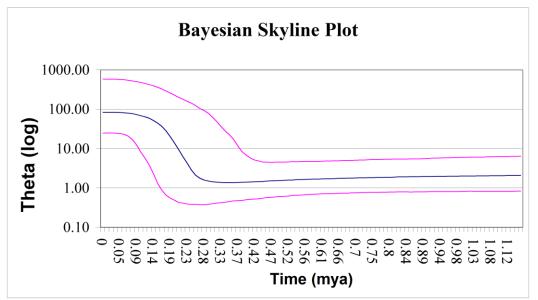


Figure 4. Bayesian skyline plot for bowhead whales. Theta is proportional to female effective population size.

We used the concatenated 3 gene network to investigate evolutionary history of the bowhead populations. Figure 4 shows a Bayesian skyline plot, which maps Θ through time, for bowhead whales. The estimated TMRCA is 1.16 million years indicating a stable population for a very long period of time. Beginning 1.16 million years ago population size was relatively small ($\Theta >$ 1). Approximately 280,000 years ago population size began to increase until approximately

90,000 years ago when it stabilized at $\Theta < 100$. It is remarkable that Θ for bowhead whales is estimated to be nearly an order of magnitude larger than for Steller sea lions (Phillips et al., in press).

Gene flow and population structure

Table 1.—Matrix of per-generation migration estimates (above the diagonal) and Fst values (below the diagonal; P values in parentheses) among 3 populations of bowhead whales.

	Canada	BCB	Okhotsk
Canada	-	46.57	2.87
BCB	0.01 (0.07)	-	3.41
Okhotsk	0.15 (0.001)	0.13 (0.02)	-

Table 1 presents the pairwise estimates of Fst, which is a measure of population differentiation. It can be seen that the Okhotsk population is significantly distinct from both the Canadian and BCB populations whereas Canada and BCB do not differ significantly. LeDuc et al. (2005) measured Fst between Okhotsk and BCB using HVRI sequences as Fst = 0.062 (P = 0.026). Our measure of Fst is about double their estimate and could be due either to sample size differences between the studies (we studied 250 BCB and 7 Okhotsk, they studied 25 Okhotsk and 29 BCB) or to the greater resolution provided by the 3-gene sequence used in this study (or a combination of both). Another study (Borge et al. 2007) compared control region sequences between historical samples from bones at archaeological sites from Svalbard with BCB HVRI sequences reported in Rooney et al. (2001). They reported Fst = 0.013 (P < 0.0001) but they downplayed the meaning of this difference due to the high diversity values in both populations and the fact that the Svalbard samples represent a sequence through immense time. It is interesting to note, nonetheless, that the Okhotsk population seems to be the most distinct of the populations and this could well be due to the severe depletion of the stock by whaling. The Svalbvard population was even more severally impacted but the samples studied by Borge et al. (2007) are thought to represent pre-whaling conditions.

Migration estimates of course reflect the estimated levels of population differentiation. The Okhotsk population exchanges only 2-3 individuals per generation with the BCB and Canadian populations which is consistent with its geographic isolation. This small level of estimated migration could be due to homoplasy in the sequences, related to the non-equilibrium nature of the existing populations, or in fact some small number of female migrants could be exchanged. We cannot rule out any of these possibilities at the present time but we note that the management implications of such a low level of exchange are not likely to be great. This is because the estimate represents a 2-way estimate of migration and thus half is likely due to recruitment into

the Okhotsk population. Only about 1.5 immigrants from the Okhotsk population into the BCB population would be expected per generation. Taylor et al (2007) estimated the generation time in Bowhead whales to be 52 years, and this, combined with the large size of the BCB population (ca. 12,000) would translate into a very slight risk of mortality from the hunt to an immigrant whale.

Tests for neutrality and diversity estimates

Fu and Li's (1993) test statistics were both negative for HVRI sequences indicating this locus is evolving in a neutral fashion and there is no evidence of population expansion or selection. However, cytochrome b was significant for F* but not for D*, and ND1 was significant for both indicating these loci are evolving under selection or they show the indication of a past rapid population expansion. The two genes concatenated (cytochrome/ND1) are significant for both test statistics. The fact that HVRI shows a different pattern of selection, or population demography, from the protein coding genes indicates caution should be used in interpreting population genetic patterns when HVRI is used alone. The most likely explanation for the difference is that when recurrent substitutions are not accounted for, HVRI loses resolution enough to misinterpret evolutionary patterns and demographic history.

Haplotype diversity estimates for the HVRI, cytochrome b, ND1 and the concatenated cytochrome b/ND1 is as follows: Gene diversity 0.94, 0.94, 0.79, 0.98; Nucleotide diversity 0.012, 0.004, 0.002, 0.003. ND1 is shown to have relatively low gene diversity compared to the other genes and HVRI has higher nucleotide diversity compared to the protein coding genes.

Conclusions

The results presented here have implications for understanding the basic evolutionary and population biology of bowhead whales. Recently, calculations of long-term effective population size were used to estimate the census number of whales in the Atlantic populations of the northern right whale (*Eubalaena glacialis*), humpback whale (*Megaptera novaeangliae*), fin whale (*Balaenoptera physalus*), and minke whale (*Balaenoptera acutorostrata*) (Roman and Palumbi, 2003). The pre-whaling population estimates based on genetic diversity for all species was considerably higher than conventional estimates. Because the Revised Management Plan of the IWC does not allow catches on stocks below 54% of the estimated carrying capacity, some aboriginal or future commercial hunts would be prevented. It should be noted however that these methods are relatively new and not yet widely in use and that great uncertainty characterizes all of these calculations. Therefore, caution is warranted as we are a long way from having a complete understanding of the biological basis of the unique molecular evolutionary patterns of cetaceans. For example, if we wish for whale stocks to recover to pre-whaling levels, does this mean $\Theta < 1$; $\Theta = 10$, or $\Theta \approx 100$? All of these values are technically pre-whaling (Fig. 4) yet they represent substantially different effective population sizes.

The use of protein coding genes in addition to HVRI has value beyond just providing additional sequence information. By allowing the identification and incorporation of recurrent substitutions into the data matrix resolution is increased to allow more fully resolved haplotype networks and to more accurately estimate evolutionary and population parameters.

Acknowledgments

We thank the Alaska Eskimo Whaling Commission (AEWC) and the Barrow Whaling Captains' Association for their confidence, guidance and support of our research. We gratefully acknowledge funding provided by the North Slope Borough Department of Wildlife Management and National Oceanic and Atmospheric Administration (through the AEWC).

Literature Cited

Alter SE, Palumbi SR (2009) Comparing evolutionary patterns of variability in the mitochondrial control region and cytochrome *b* in three species of baleen whales. *Journal of Molecular Evolution*, 68, 97-111.

Borge T, Bachmann L, Bjornstad G, Wiig Ø (2007) Genetic variation in Holocene bowhead whales from Svalbard. *Molecular Ecology*, 16, 2223-2235.

Drummong AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22, 1185-1192.

Fu Y-X, W.H. Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133, 693–709.

Henn, BM, Gignoux CR, Feldman MW, Mountain JL (2009) Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. *Molecular Biology and Evolution*, 26, 217-230.

Heyer E, Zietkiewicz E, Rochowski A, et al. (2001) Phylogenetic and familial estimates of mitochondrial substitutions rates: study of control region mutations in deep-rooting pedigrees. *American Journal of Human Genetics*, 69, 1113-1126.

Howell N, Bogolin Smejkal C, Mackey, DA, et al. (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *American Journal of Human Genetics*, 72, 659-670.

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

Li WH, Chu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, 2, 150-174.

McGowen MR, Michelle Spaulding M, Gatesy J (2009) Divergence date estimation and a comprehensive molecular tree of extant cetaceans. *Molecular Phylogenetics and Evolution*, 53, 891–906.

Parsons TJ, Muniec DS, Sullivan K, et al. (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nature*, 15, 363-368.

Phillips CD, Trujillo RG, Gelatt TS, Smolen MJ, Matson CW, Honeycutt RL, Patton JC Bickham JW (2009) Assessing substitution patterns, rates and homoplasy at HVRI of Steller sea lions, *Eumetopias jubatus*. *Molecular Ecology*, 18, 3379-3393.

Phillips CD, Gelatt TS, Patton JC, Bickham JW (in press) Phylogeography of Steller sea lions: how climate change and effective population size determine patterns of genetic diversity. Journal of Mammalogy.

Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics*, 14, 817-818.

Roman J, Palumbi SR (2003) Whales before whaling in the North Atlantic. *Science*, 301, 508-510.

Schmidt HA, Strimmer K, Vingron M, Haeseler A (2002) Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18, 502-504.

Sigurðardóttir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P (2000) The mutation rate in the human mtDNA control region. *American Journal of Human Genetics*, 66, 1599-1609.

Taylor, B., LeDuc, R., George, J.C., Suydam, R., Moore, S., and Rugh, D. 2007. Synthesis of lines of evidence for population structure for bowhead whales in the Bering-Chukchi-Beaufort region. Paper SC/59/BRG35 presented to the IWC Scientific Committee, May 2007. Available from The International Whaling Commission, The Red House, 135 Station Road, Impington, Cambridge, Cambridgeshire CB24 9NP, United Kingdom.