Assessing molecular substitution patterns in the mitochondrial control region compared to protein coding genes in two marine mammals.

John W. Bickham[1], Ryan M. Huebinger[1], Caleb D. Phillips[1], John C. Patton[1], Richard LeDuc[2], Lianne D. Postma[3], J. Craig George[4] and Robert Suydam[4]

[1]Center for the Environment and Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907, USA; [2]Southwest Fisheries Science Center, 8604 La Jolla Shores Dr., La Jolla, CA, 92037; [3] Fisheries and Oceans Canada, Central and Arctic Region, Winnipeg, Manitoba, Canada R3T 2N6; [4]North Slope Borough, Department of Wildlife Management, Barrow, AK 99723, USA

Abstract

This paper compares substitution patterns in three mitochondrial genes, control region, cytochrome b, and ND-1 from 245 bowhead whales representing three populations; the BCB stock, the eastern Canadian arctic stock and the Sea of Okhotsk stock. Substitution patterns for the hypervariable region I (HVRI) of the mitochondrial control region were compared to the more conservative protein coding genes in order to identify hypervariable sites which are sources of homoplasy and evolutionary "noise." Since HVR1 is one of the most frequently used genetic markers for population genetic and phylogeographic studies in mammals, a method to increase the resolution of this marker would increase our understanding of the population processes that drive genetic patterns. In particular, it will result in a better understanding of long-term effective population sizes. We present an evolutionary analysis of HVRI, cytochrome *b* and ND-1 sequences in order to quantify the occurrence of homoplasy in HVRI. The data from bowheads are compared to a larger dataset from Steller sea lions in which the estimated rate of substitution at HVRI is approximately 24 times the substitution rate at cytochrome *b* with an absolute rate of HVRI substitution estimated at 27.45% per million years. In contrast, bowheads exhibit a lower rate of mtDNA substitution but the depth of bowhead lineage leads to a considerable amount of recurrent substitution in protein coding genes; more than in the Steller sea lions. Analysis of historical population size in both species show considerable variation in female effective population size through evolutionary time.

Introduction

The hyper variable region (HVRI) of the control region of the mitochondrial DNA is one of the most frequently used genetic 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 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.

The purposes of this study were to investigate site specific mutation rates in bowhead whales, to estimate the degree to which variation in these rates causes homoplasy, and to present haplotype networks of one, two and three genes concatenated to illustrate the degree to which resolution can be achieved. Comparisons are made to Steller sea lions for which a more complete analyses has been achieved. Accurate estimation of mutation rate is key to the calculation of theta; $\Theta = 2N_{e(f)}\mu$ where $N_{e(f)}$ 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., in press), 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 Bayesian skyline plots to show the evolutionary trends of female effective population sizes for Steller sea lions and bowhead whales 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 (list citations from p.21). For Steller sea lions, Phillips et al. (in press) calculated 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 sequences (397 base pairs) were obtained from 245 bowhead whales. Cytochrome b sequences were obtained from the same 245 whales using the methods described in Phillips et al. (in press). ND1 sequences were obtained from the same 245 whales following Phillips (2009).

Describing rates and patterns of substitution at HVRI
Site-specific rates of substitution were calculated from composite haplotypes and relative rates of substation were calculated for each of the three genes. ModelTest 3.7 (Posada and Crandall, 1998) calculated 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 by the program 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.

The method of tree weighting and substitution mapping as developed by Phillips et al. (in press) is presented for Steller sea lions and a modification of this procedure was used for bowhead whales due to evidence of homoplasy in the protein coding genes which makes problematical their use in tree weighting. The method employs the construction of neighbor-joining trees that are constrained to the topology determined by cytochrome b (in the case of Steller sea lions) and the subsequent examination of recurrent substitutions at HVRI on the tree. This allows for the exact quantification of site-specific numbers of recurrent substitutions at HVRI and a subsequent independent estimate of relative rates of substitution of the these genes. Another method used to estimate gene specific substitution rates, that described by Alter and Palumbi (2009) was employed for Steller sea lions (Phillips et al., in press) and for bowhead whales. This method calculates the substitution rate per base as 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, Phillips et al. (in press) calculated the synonymous pairwise distance (Li et al., 1985) among nine otariid species for cytochrome b. The silent substitution rate was then calculated as half the slope of the 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.

Results and Discussion
Describing rates and patterns of substitution at HVRI
Figure 1 illustrates one method for the identification of homoplasies in HVRI using data from Steller sea lions (Phillips et al., in press). In this method multiple associations of HVRI haplotypes with cytochrome b haplotypes are identified as filled circles on the HVRI network within each of the large circles representing cytochrome b haplotypes.

From this, the specific number of homoplastic mutations can be identified. As seen in Figure 2 (bottom panel), 19 of the 41 variable sites had multiple mutations and two sites had 18 and 11 recurrent mutations, respectively. A comparison of this with the top panel shows imperfect agreement between site-specific mutation rates estimated by Γ rate category assignment and the tree weighting and substitution mapping method. In the top panel, 5 of the 41 variable sites are placed in the highest rate category which is estimated as 39.27 times higher than the average of the concatenated sequence.

We applied a modification of the tree weighting and substitution mapping method to the bowhead dataset. Figure 3 shows site specific rate estimates for bowheads using the Γ rate category assignment method as well as the estimated number of recurrent substitutions per site. This figure shows that there are considerably more recurrent substitutions estimated for the protein coding genes for bowheads than for Steller sea lions (Figure 2). Compared to Steller sea lions, bowhead whales have fewer recurrent substitutions in the control region and more in the coding regions. It seems likely that the relationship between relative gene substation rates and depth of the phylogeny contributes to the difference in patterns seen in these species. 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).

It is clear from the Steller sea lion dataset that the three methods to calculate site specific mutation rates and the relative rates of substitution of HVRI to cytochrome b are consistent. The relative rates of substitution of HVRI to cytochrome b calculated by the Γ rate category assignment 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. All of these are in reasonably good agreement. However, using the Γ rate category assignment method the relative rates of substitution of HVRI to cytochrome b for bowhead whales was 3.77 and the substitution mapping method yielded a rate of 4.24. The bowhead HVRI substitution rate was 2.8% per million years.

Figure 4 shows the effect of the addition of sequence to resolving haplotype networks. Not surprisingly, as you go from HVRI alone (left panels) to HVRI plus cytochrome b, to HVRI/cytochrome b plus ND1 the number of reticulations in the networks is reduced. However, this method alone does not resolve the fully concatenated 3-gene networks. But an examination of Figure 5 shows that by extending the sequence to include the recurrent mutation identified by the tree weighting and substitution mapping method in Steller sea lions as additional characters the network becomes fully resolved. Likewise, the same method applied to bowhead whales resulted in a fully resolved haplotype network. This “extended” network was applied in a nested clade analysis (Phillips, 2009) of Steller sea lions and 17 mutations defining significant associations of clades with geographic distribution were identified. These 17 clades are mapped onto a tracing of

atmospheric methane (Figure 6). It was found that most of these mutations map to times of low methane levels and in particular they fall in periods of glaciations. Interestingly, population inferences are substantially different for the past three glaciation events. For the most recent, there were no significant associations indicating a stable population. The second most recent glaciation was characterized by range expansion. The third most recent glaciation was characterized by allopatric fragmentation.

Thus, as one goes back farther in time in the evolutionary history of Steller sea lions the mutations associated with geographic distributions are the result of more severe population impact. The ostensible reason for the severity of impacts can be seen by an examination of the Bayesian skyline plot which map Θ through time (Figure 7, top panel). The estimated time to most recent common ancestor (TMRCA) for Steller sea lions mtDNA is approximately 360,000 years. Beginning at that point until approximately 130,000 years ago effective population size was relatively low with $\Theta < 1$. Approximately 130,000 years ago effective population size began to increase until approximately 20,000 years ago when it stabilized at $\Theta > 10$. The most severe population effects (allopatric fragmentation) occurred when effective population size was lowest, intermediate effects (range expansion) occurred when effective population size was increasing, and no effects were observed when effective population size was highest during the last glaciation. Thus, this analysis shows empirically that effective population size has a profound impact on buffering the effects of climate change on a species of marine mammal.

Figure 7 (bottom panel) also shows the Bayesian skyline plot for bowhead whales. The estimated TMRCA for this species 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 ($\Theta > 1$) but considerably larger than Steller sea lions. 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. Given the lower mutation rates of whales compared to other mammals including Steller sea lions, this would indicate an even greater relative value of Ne.

The results presented here have implications for understanding the basic evolutionary and population biology of bowhead whales and the eventual accurate calculations of Ne which these methods promise might impact conservation decisions of the IWC (Roman and Palumbi, 2003). It should be noted 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 yet they represent substantially different effective population sizes.

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) and Fisheries and Oceans Canada.

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.

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.

Phillips CD (2009) Systematics, molecular evolution, and phylogeography of Steller sea lions, *Eumetopias jubatus*. Doctoral Dissertation, Purdue University, West Lafayette, IN.

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

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.

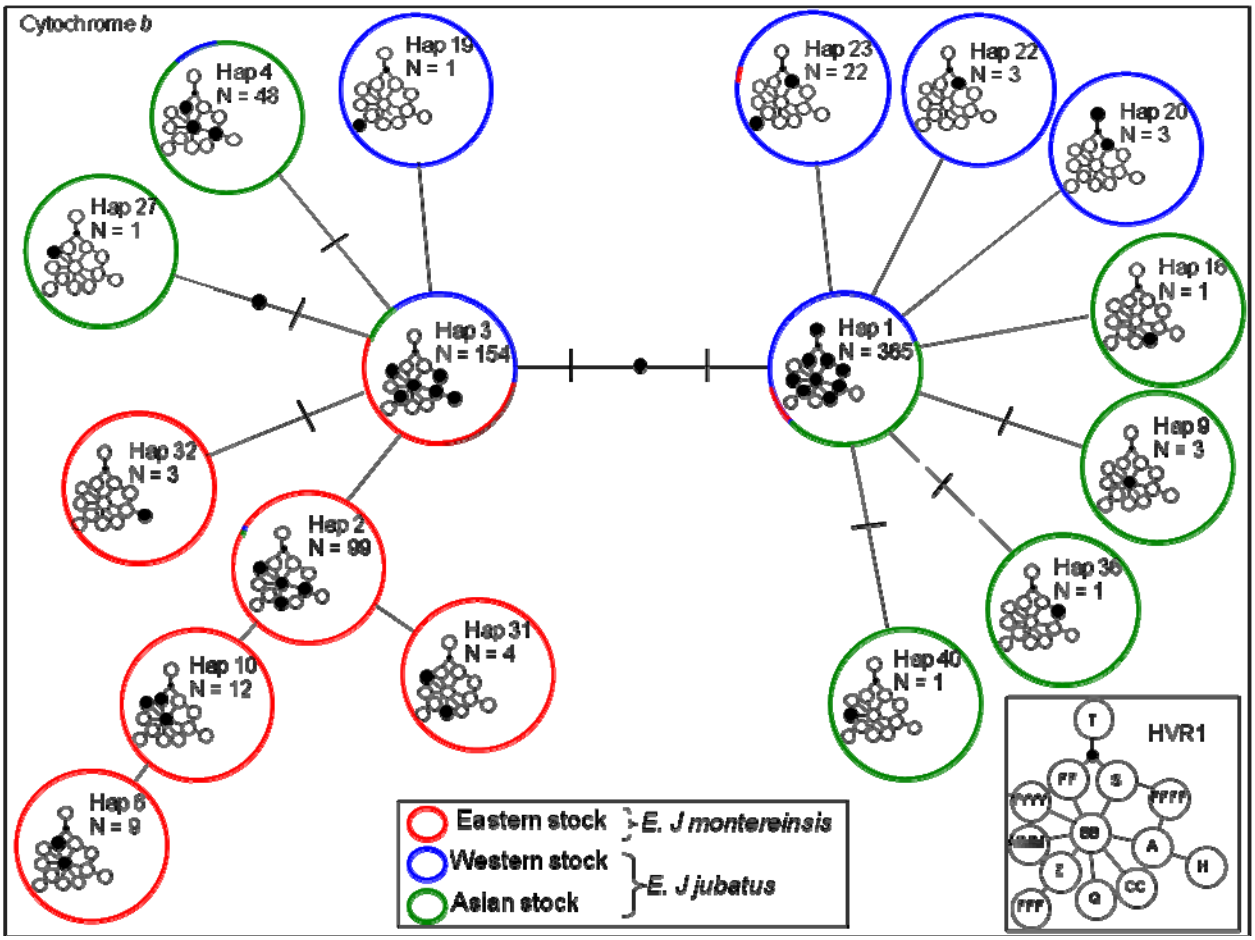


Figure 1. Steller sea lions cytochrome b haplotype network (large circles) and HVRI network (small network inside large circles). The filled circles of the HVRI network show which haplotypes are associated with that cytochrome b haplotype.

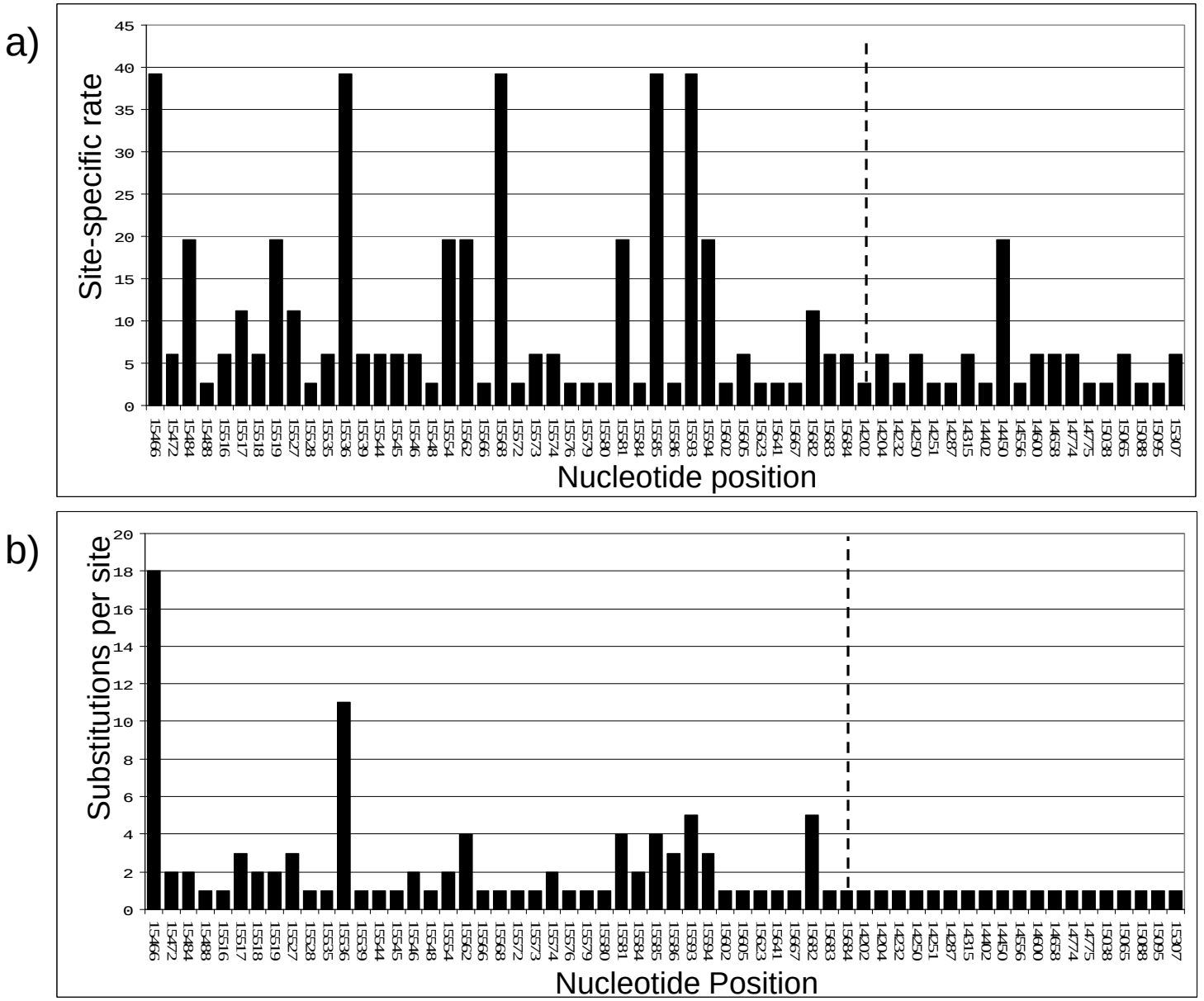


Figure 2. Estimated site-specific mutation rates (top) and calculated substitutions per site for variable HVRI (left of dashed line) and cytochrome b sites in Steller sea lions.

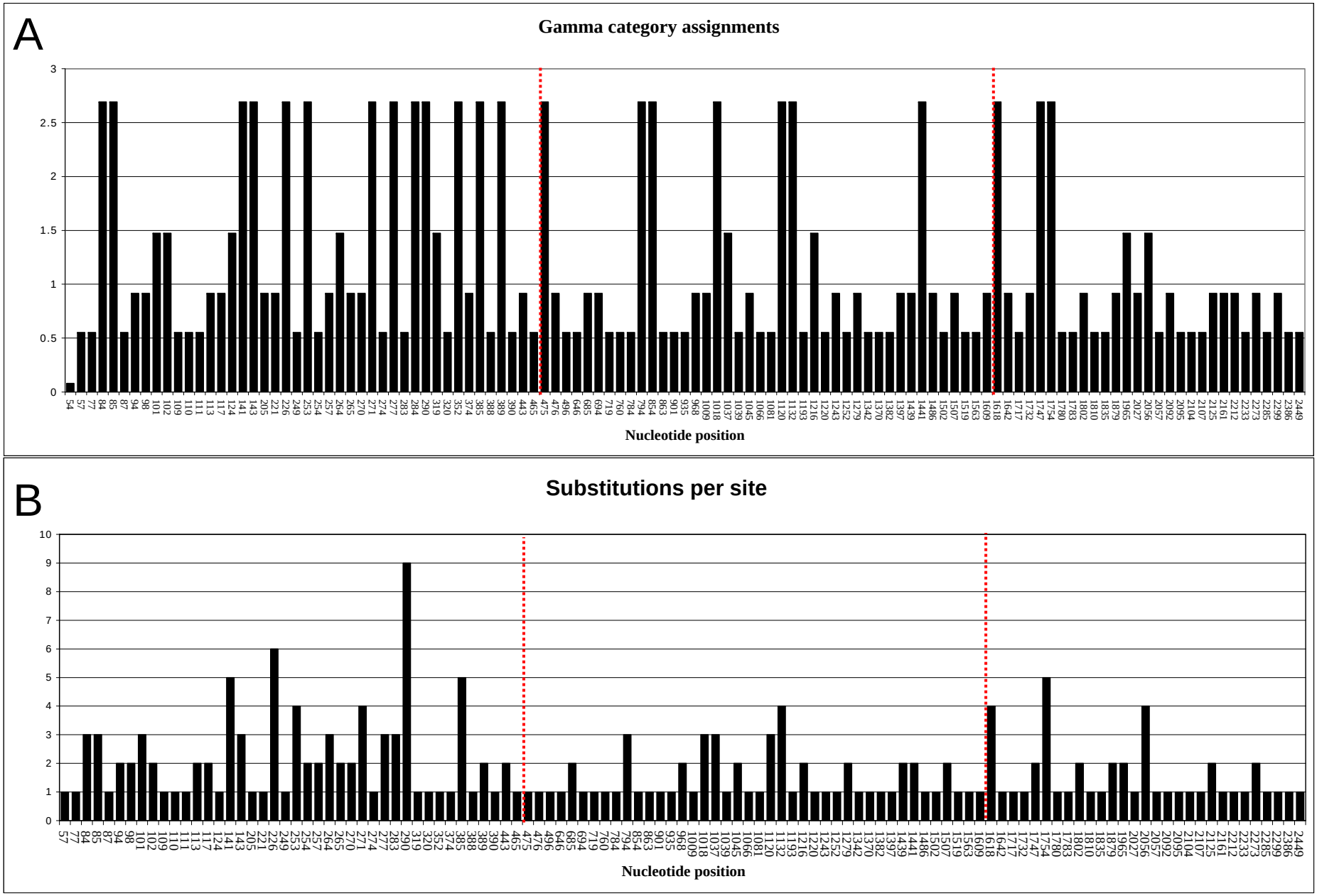


Figure 3. Estimated site-specific mutation rates (A) and calculated substitutions per site (B) for HVRI (left), cytochrome b (middle), and ND1 (right) variable sites for bowhead whales.

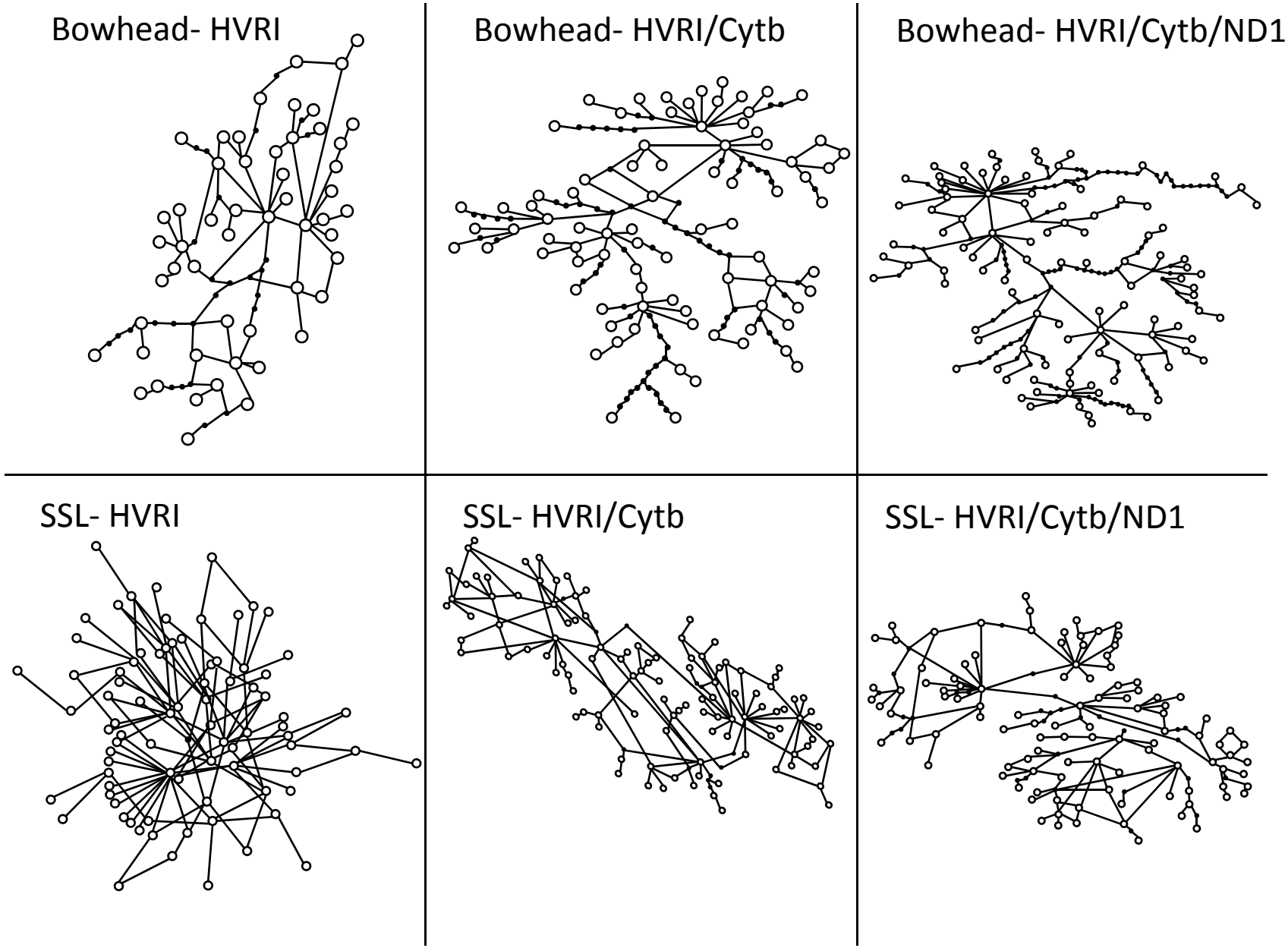


Figure 4. Haplotype networks for bowhead whales (top row) and Steller sea lions bottom row). Left panels show the HVRI networks. Middle panels show HVRI and cytochrome b concatenated sequence networks. Right panels show HVRI, cytochrome b, and ND1 concatenated sequences networks.

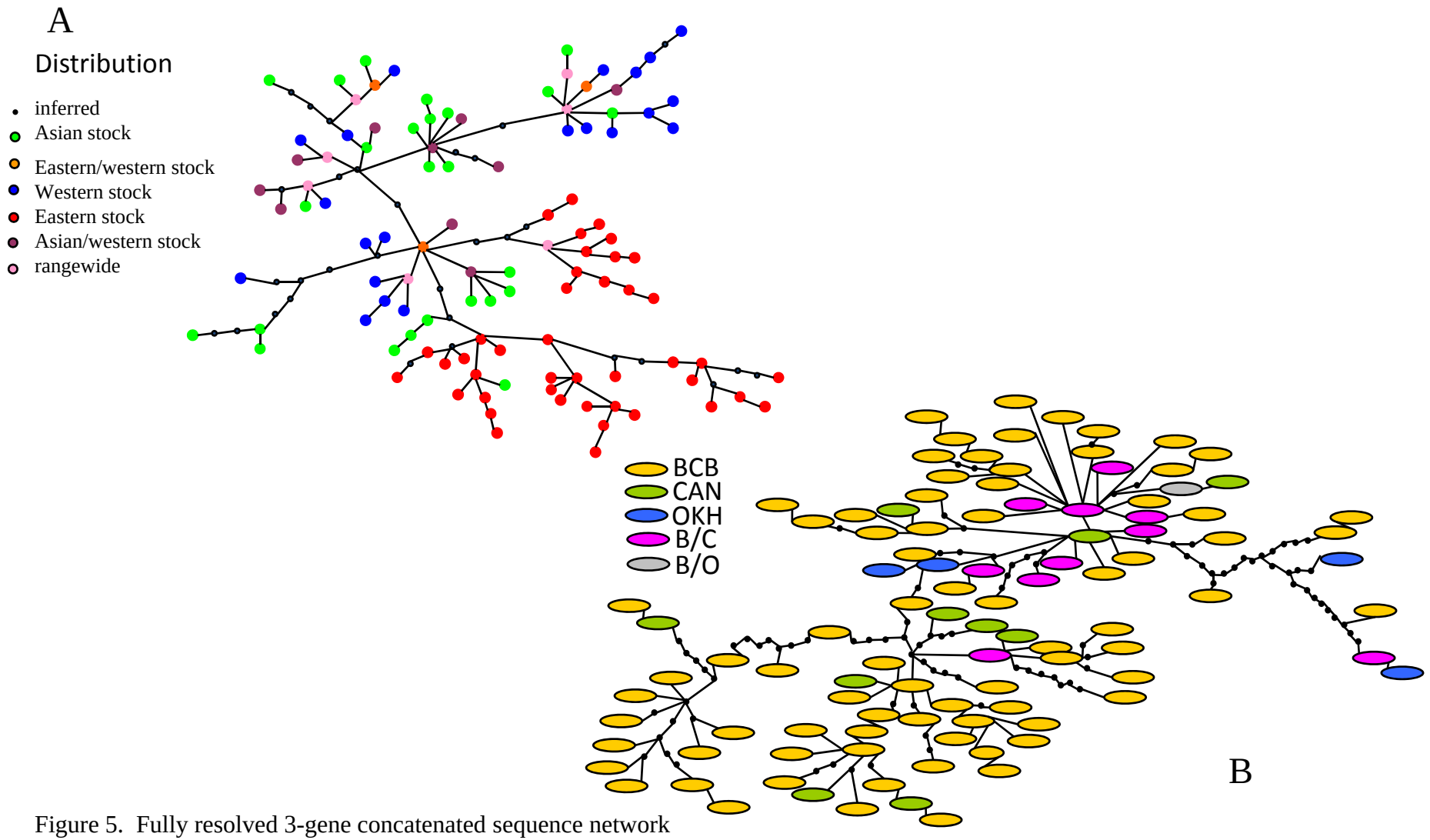


Figure 5. Fully resolved 3-gene concatenated sequence network resulting from the extension of the sequence to account for the multiple substitutions at hyper-variable sites in Steller sea lions (A) and bowhead whales (B).

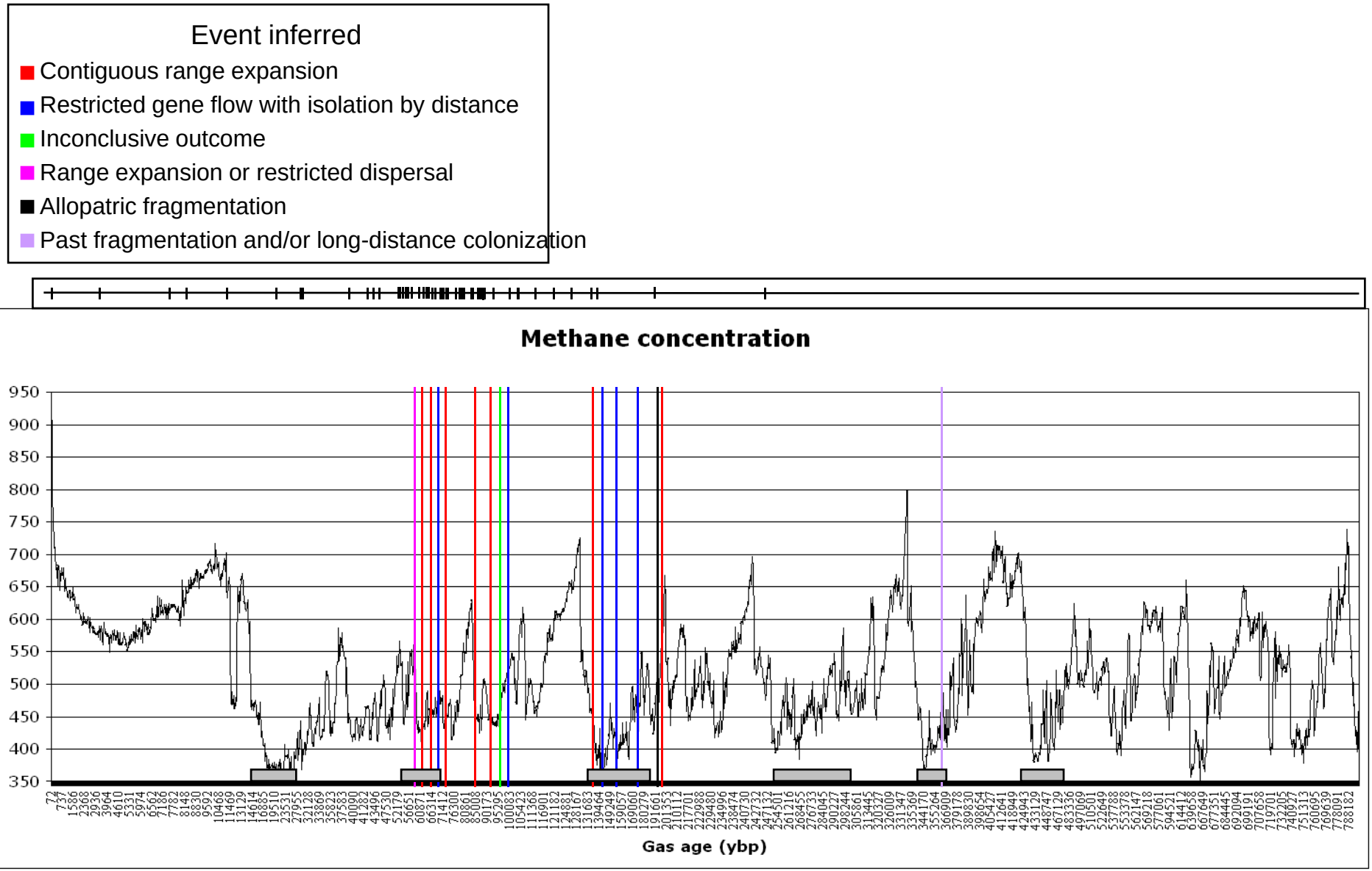
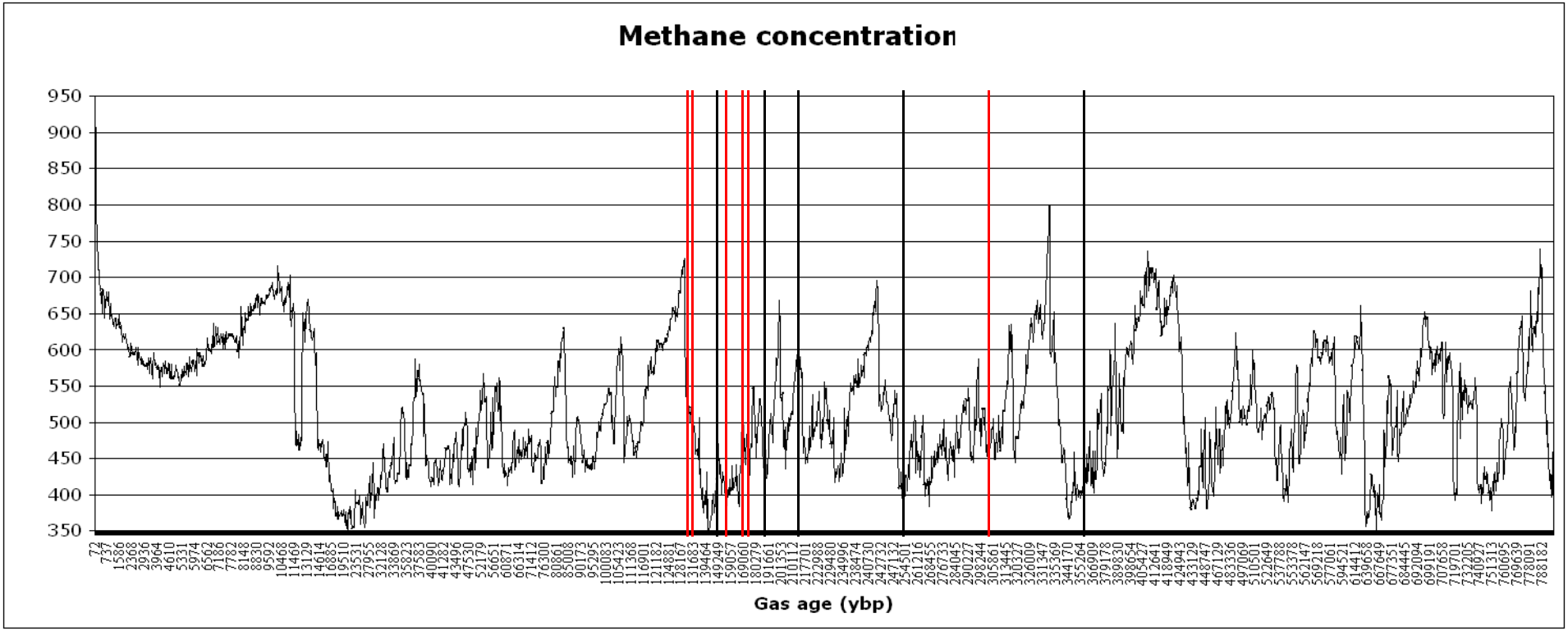


Figure 6. Statistically significant associations of Steller sea lion clades (colored vertical bars) with geographic distribution overlaid on a plot of atmospheric methane concentrations. Horizontal gray bars span the times of recognized glaciations and the tic marks show non-significant clades.

important nodes from unconstrained tree from dell computer



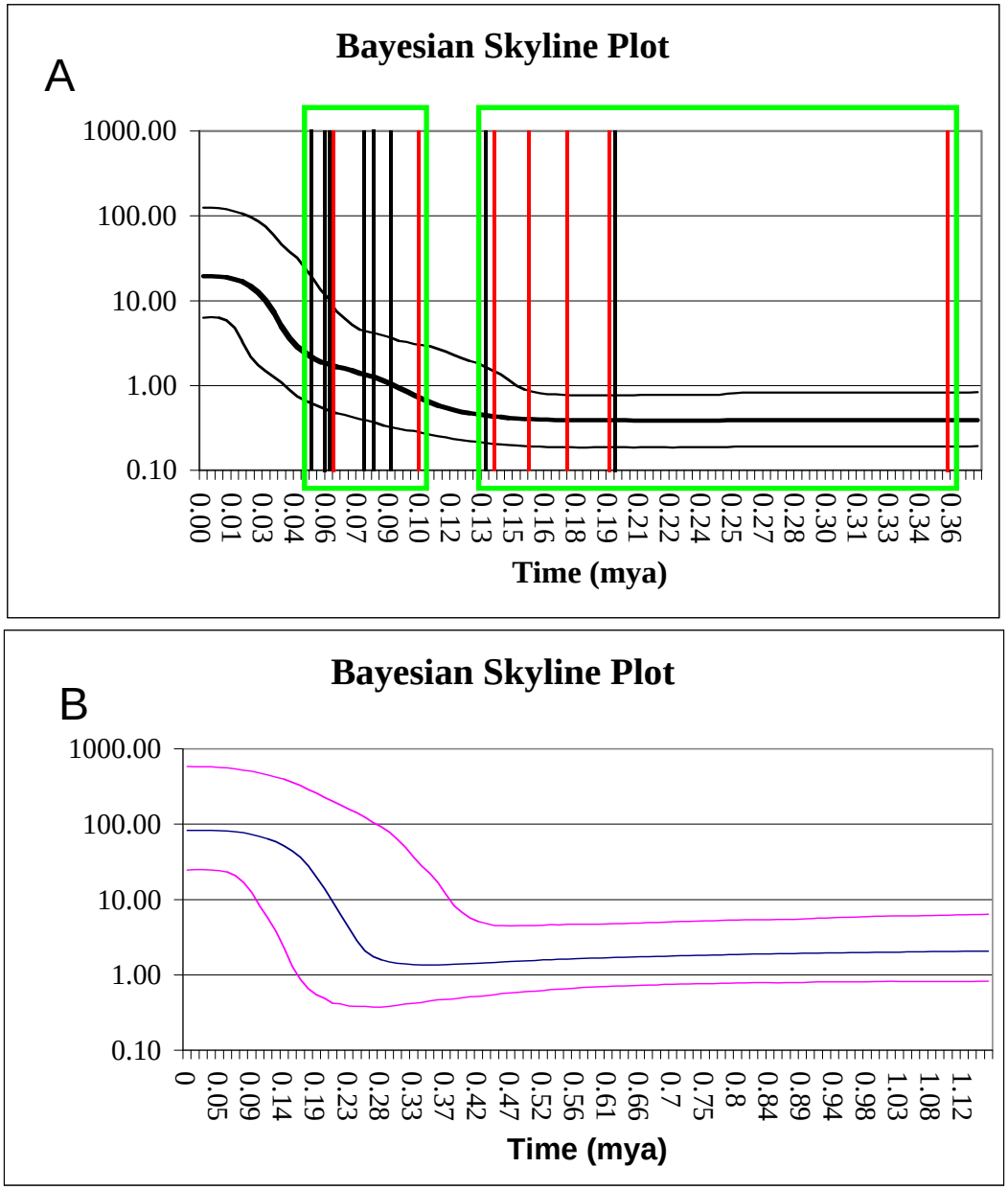


Figure 7. Bayesian skyline plots for Steller sea lions (top) and bowhead whales (bottom). The time estimates for the Steller sea lion clades that are statistically significantly associated with geographic distribution are shown.