AN ABSTRACT OF THE DISSERTATION OF

Alana Alexander for the degree of Doctor of Philosophy in Wildlife Science presented on September 15, 2014.

Title: The Influence of Social Structure and Molecular Evolution on Genetic Diversity in the Sperm Whale (*Physeter macrocephalus*)

Abstract approved:

_____

C. Scott Baker

The sperm whale (*Physeter macrocephalus*) shows some of the most derived characteristics of any mammal: a large body size, large brain, complex social organization and a capacity for deep foraging dives that few other marine mammals can match. Despite a history of exploitation that removed hundreds of thousands of individuals, the sperm whale population remains relatively abundant in comparison with other large whale species. Given this abundance, and the sperm whale's matrifocal social organization, it is surprising that previous research found that mitochondrial DNA control region (mtDNA CR) diversity in sperm whales is extremely low and population structure is relatively limited within oceans. This dissertation addresses several questions spanning evolutionary and ecological time scales, including whether the low levels of mtDNA CR diversity and differentiation seen in the sperm whale have been limited by sample size and geographic scope in previous studies; how sperm whale genetic diversity is partitioned at several hierarchical levels; and hypotheses explaining the low mtDNA CR diversity. To achieve this, I generated DNA profiles representing 557 individuals from

circum-equatorial regions, strandings around the coasts of New Zealand, Samoa and Oregon, and biopsy samples from the Gulf of Mexico.

DNA genotypes constructed from these samples (mtDNA CR, sex, 13 microsatellite loci), and mtDNA information from 1,167 previously published samples, indicated a high degree of mtDNA CR differentiation within the previously un-sampled Indian Ocean (F_{ST} 0.314, $p < 0.001$). The level of differentiation seen was similar to that found with the marginal seas of the Atlantic i.e. the Gulf of Mexico and the Mediterranean (F_{ST} 0.469, $p < 0.001$). In contrast, levels of mtDNA differentiation seen in the Pacific were much lower (F_{ST} 0.061, $p < 0.001$). Microsatellite differentiation was much less marked for all three oceans, consistent with tests indicating male-biased dispersal and gene flow. In addition to regional differentiation, significant differentiation was seen among social groups. However, the magnitude of this differentiation differed by ocean. Hierarchical mtDNA analyses showed that in the Pacific, 'social group' explained more variance than geographic region. In contrast, in the Indian Ocean, regions explained more variance than social group. In the Atlantic, the number of social groups within regions was too limited to make conclusions. Social group was the only level that explained significant variation in microsatellite allele frequencies in any ocean. Increased relatedness within social groups does not appear to explain the microsatellite differentiation. Instead, the likely explanation is different breeding males consorting with different female-dominated social groups. mtDNA differentiation seen among social groups appears to be driven by 'lenient matrilineality', where 38% of groups were strictly matrilineal, and a further 25% of groups were comprised of more than one matriline, but fewer matrilines than expected by

chance. However, the levels of matrilineality are too low to be consistent with the previously proposed hypothesis of a selective sweep linked with maternal cultural innovations as an explanation for low mtDNA diversity in the sperm whale.

To examine alternative hypotheses for low mtDNA diversity, next-generation sequencing (454 and Illumina) was used to sequence mitogenomes for 17 Pacific Ocean sperm whale samples, with other cetacean mitogenomes compiled from the literature. Using these mitogenomes, no evidence of slow substitution rates were found in the mtDNA CR or protein-coding genes of the mitogenome that could explain the low diversity. In addition, the mtDNA CR had the highest diversity of the entire mitogenome and showed genealogical patterns concordant with the rest of the mitogenome. This discounts mtDNA CR-specific constraints as the cause of low mtDNA CR diversity. To investigate the remaining hypotheses of a selective sweep, population bottleneck or expansion, 8 nuclear loci (~12,000 bp) were sequenced for 22 sperm whales (Pacific and Gulf of Mexico), and compared to 10 New Zealand pygmy sperm whales (*Kogia breviceps*). The results were inconsistent with a selective sweep and showed instead low diversity across both mtDNA and nuclear DNA, in comparison to the higher levels of genetic diversity in the pygmy sperm whale and other cetacean species. Demographic reconstructions showed the sperm whale to have had a stable, but small, population size for much of historical time. This suggests a recent population expansion is responsible for the low mtDNA (and nuclear DNA) diversity in the sperm whale. The inferred timing of the expansion corresponds with expansions in squid species (the primary prey of the sperm whale), and explains shared mtDNA haplotypes between oceans. Since this expansion, the marked philopatry

shown by female sperm whales at various hierarchical levels ranging from social groups (e.g. lenient matrilineality) to broader geographic scales, has led to maternally-mediated genetic drift driving striking differences in mtDNA haplotype frequencies between social groups, regions, and oceans.

The Influence of Social Structure and Molecular Evolution on Genetic Diversity in the Sperm Whale (*Physeter macrocephalus*)

by
Alana Alexander

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented September 15, 2014
Commencement June 2015

Doctor of Philosophy dissertation of Alana Alexander presented on September 15, 2014.

APPROVED:

Major Professor, representing Wildlife Science

Head of the Department of Fisheries and Wildlife

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request

Alana Alexander, Author

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Scott Baker. His guidance, lively discussions, and scholarly input have really shaped my understanding of the scientific method, and I'm very grateful for his time and dedication to myself and all his students. I'd also like to extend my sincere thanks to my committee members: Kathleen O'Malley, Aaron Liston, Roger Payne, and Gary Merrill for their guidance and input into my research.

Some say it takes a village to raise a child... this is a similar situation to how a dissertation gets finished! I'd like to thank my co-authors and collaborators for their input and editorial advice. Also, there are not enough thanks in the world for my amazing lab mates in the Cetacean Conservation and Genomics Laboratory: I'm very lucky to have had a supportive and engaging group of friends to go through graduate school with. I'd also like to thank the other students, staff, and faculty at the Hatfield Marine Science Center (HMSC) and Department of Fisheries and Wildlife (FW) for their support and friendship. I'd also like to acknowledge the financial support I have received from HMSC and FW both in research costs and through teaching assistantships.

Special thanks to queen of the lab, Debbie Steel, for providing sound advice and writing me up in the laboratory 'book of idiotry' when I deserved it! In addition, special thanks to my family and friends, especially my whanau back home in NZ: Mum, Dad, Becky, and Shaun. I owe a debt to my wonderful husband, Ryan, for putting up with me not moving off the couch for weeks at a time while I was writing up, and just for being an amazing

guy who could make me feel better after even the most disastrous day(s) in the lab/computer coding/on the 18th draft of a paper.

I also need to thank all the people involved in collecting the samples used in this study: the crew of the *Odyssey* and the Ocean Alliance, Rochelle Constantine and the New Zealand Department of Conservation, Bruce Mate, Jim Rice and the OSU Marine Mammal Institute. Finally, I would like to thank the Fulbright program for the opportunity to experience another culture and country during my doctoral research. I am forever indebted for the different perspective it has given me.

"You will never be completely at home again, because part of your heart will always be elsewhere. That is the price you pay for the richness of loving and knowing people in more than one place."
– Miriam Adeney

CONTRIBUTION OF AUTHORS

C. Scott Baker was involved in the development and editing of all chapters.

Debbie Steel was involved in processing of samples and discussions related to Chapter 2, 3, and 5. In addition, she was involved in discussions and editing of the publication that is Chapter 4.

Kendra Hoekzema was involved in the processing of samples, discussions and editing related to Chapter 2 and 5. In addition, she was involved in discussions and editing of the publication that is Chapter 4.

Daniel Engelhaupt, Iain Kerr, and Roger Payne were instrumental in the collection of samples pertaining to Chapter 2 and 3, and discussions relating to these chapters.

Sarah Mesnick provided previously published data used in Chapter 2, and discussion and edits relating to this chapter.

Renee Albertson contributed to discussions and editing relating to Chapter 3.

Beth Slikas, Colm Carraher, Matthew Parks, and Richard Cronn were involved in the processing of samples, discussions and editing of the publication that is Chapter 4.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figure | Page

LIST OF TABLES

LIST OF TABLES (Continued)

Table | Page

LIST OF APPENDIX FIGURES

LIST OF APPENDIX FIGURES (Continued)

Figure Page

LIST OF APPENDIX TABLES

LIST OF APPENDIX TABLES (Continued)

Table Page

1. General introduction

1.1. Evolutionary history

The sperm whale (*Physeter macrocephalus*) is a member of one of the earliest branching lineages in Cetacea (whales and dolphins), and is within the oldest lineage in Odontoceti (toothed whales). This lineage, the superfamily Physeteroidea (comprised of the sperm whale family Physeteridae, and dwarf/pygmy sperm whale family, Kogiidae), is thought to have diverged from the remaining odontocete species approximately 23 million years ago, based on morphological evidence (Fordyce and Barnes 1994). Kogiidae, comprised of the sister species to the sperm whale, the dwarf (*Kogia sima*) and pygmy (*K. breviceps*) sperm whales, are thought to have diverged from the ancestral sperm whale lineage over 10 million years ago (Fordyce and Barnes 1994). The dwarf and pygmy sperm whales are vastly smaller and show different social structure in comparison with the sperm whale.

The lack of close living relatives, and mixture of derived and ancestral character states also common to Mysticeti (baleen whales), has led to some controversy about the placement of the sperm whale on the cetacean phylogeny. Some evidence suggests that the sperm whale could be a basal mysticete, rather than odontocete (Milinkovitch 1995, Milinkovitch et al. 1993, Nikaido et al. 2001). However, more recent studies show firm support for sperm whales and odontocetes sharing a monophyletic origin (May-Collado and Agnarsson 2006, O'Leary and Gatesy 2008). Given the possibility of fast substitution rates in sperm whale mitochondrial DNA (Lyrholm et al. 1996), the analyses aligning

sperm whales with the mysticetes (which were based on mitochondrial DNA) might have been affected by long-branch attraction (Felsenstein 1978).

1.2. Morphology, physiology, foraging, and acoustics

The long independent evolutionary trajectory of the sperm whale is likely responsible for it being one of the most morphologically and physiologically derived of all mammals. As described in '*Moby-Dick'* (Melville 1851) the sperm whale has the largest body size of all extant odontocetes (Whitehead 2003). However, this large body size is not shared equally by both sexes. Sperm whales are the most sexually dimorphic of all cetaceans with mature males measuring longer than 15 m and weighing up to 45,000 kg, while females tend to be less than 11 m long and weigh only about 15,000 kg when full grown (Best 1979, Whitehead 2003). Despite size differences, both sexes of the sperm whale routinely make some of the deepest and longest dives of any marine mammal: over 73 min long and up to 2,035 m deep (Watkins et al. 1993, Watwood et al. 2006) to feed on squid, including the infamous giant (*Architeuthis* spp.) and colossal (*Mesonychoteuthis hamiltoni*) squids (Best 1979, Clarke et al. 1993, Pascoe et al. 1990).

Along with large size, additional physiological adaptations make these foraging dives possible including: increases in muscle myoglobin levels in comparison with terrestrial mammals (Noren and Williams 2000); a greater utilization of blood and muscle storage, rather than lung storage of oxygen (Kooyman and Ponganis 1998); collapsing of lungs at shallow depths (Kooyman and Ponganis 1998) that reduces the chance of nitrogen embolisms upon resurfacing (Boyd 1997); and muscle tissue able to tolerate short bursts

of highly active swimming, or long periods of anaerobic respiration (Castellini and Somero 1981). An additional adaptation is the sperm whale's nose, which covers at least 1/3 of the animal's body length (Whitehead 2003). The nose contains the spermaceti organ that is thought to function in the production of sound used in long-range biosonar to locate prey (Madsen et al. 2002, Watwood et al. 2006). The echolocation clicks produced by the sperm whale are some of the loudest sounds produced by any animal (Møhl et al. 2000), and could function in allowing sperm whales to home in on potential prey items as soon as they begin their foraging dives.

An interesting feature of the sperm whale's well-developed acoustic apparatus, is that it also seems to function in communication, as well as prey-detection (Madsen et al. 2002). Sperm whales produce stereotypical short series of clicks, known as 'codas' (Watkins and Schevill 1977). Codas appear to have a social function, as they are produced by sperm whales when individuals or groups meet (Watkins and Schevill 1977), and also appear to maintain social cohesion within social groups of females (Weilgart and Whitehead 1993). In contrast, adult males very rarely produce codas (Whitehead 2003).

1.3. Distribution, history of exploitation, and abundance

Along with the marked differences by sex in body size and coda production, the sperm whale also shows pronounced differences in the distribution of sexes in the world's oceans. As a whole the species has a global distribution, inhabiting all oceans of the world (Gosho et al. 1984). However, pods comprised of females and juveniles are restricted to the lower latitudes and equatorial regions, while adolescent and mature

males reside in both the equatorial regions and cold temperate/polar regions above 40° latitude (Figure 1.1, Best 1979, Drouot et al. 2004b, Whitehead 2003). The sexual segregation seen in sperm whales means that age- and sex-classes were disproportionately affected by different periods of whaling.

Female social groups were the focus of whaling on tropical 'grounds' by 18th - 19th century Yankee (open-boat) whalers, who exploited the tendency of group members to 'stand by' when one of their number was mortally injured by harpoon (Best 1979, Whitehead 2003). The products of this hunt were predominantly spermaceti oil, used as lamp fuel, and rendered blubber to grease the wheels of the Industrial Revolution (Whitehead 2003). However, as abundance of sperm whales began to fall and petroleum became widespread as a low-cost lubricant and fuel alternative, open-boat hunting of sperm whales began to decline around the late 19th century (Whitehead 2002; 2003).

The onset of modern whaling in the mid-1920s did not initially impact sperm whale populations, despite the new ability for factory-ship-oriented whaling operations to target the larger, and therefore more economical, bull whales in polar waters (Best 1979). It was not until the larger baleen whales, such as the blue (*Balaenoptera musculus*) and fin whale (*B. physalus*) had been depleted, that whalers turned their attention back to the sperm whale (Whitehead 2002; 2003). Following this, an unprecedented number of sperm whales were killed – approximately half a million – before a full moratorium was placed on the sperm whale hunt in 1988 (Whitehead 2002; 2003). It is estimated that prior to commercial whaling, the global population of sperm whales numbered around

1,110,000 individuals (Whitehead 2002). In 1999, ten years after the cessation of widespread commercial whaling, the global sperm whale population was estimated to still be at less than 32% of its pre-exploitation abundance (Whitehead 2002).

Although whaling had a marked effect on sperm whale abundance, it also informed some of the first discussions about possible stock structure in this species. Differences by area in the size distribution of whales, prey species present in stomachs and biochemical markers, as well as the recovery of '*Discovery* marks' in areas close to where individuals were originally marked, suggested restricted dispersal of individuals (Allen 1980, Best 1979, Brown 1967, Clarke 1972, International Whaling Commission 1980, Kasuya and Miyashita 1988). In addition, areas of local depletion, apparently due to whaling, argue for potentially restricted patterns of dispersal in this species (Whitehead et al. 1997). However, other evidence, such as the lack of clear discontinuity in catches between areas, the presence of 'foreign' prey species in stomachs, and the recovery of *Discovery* marks and harpoons from animals 1,000s of kilometers from where they were initially marked, argues for a lack of stock structure in the sperm whale (Aguilar 1985, Clarke 1972, Ivashin and Rovnin 1967, Kasuya and Miyashita 1988, Steiner et al. 2012). The discrepancies of these findings - even within the same methodology - could be driven in part by more extensive mixing of males in comparison with females, and potential changes in mixing of whales by season (Gosho et al. 1984, Jaquet 1996, Kasuya and Miyashita 1988).

1.4. Social organization

Whaling also provided scientists with some of the first direct observations of the social structure of this species. This has since been studied in detail using non-lethal methods. Female social groups are thought to contain long-term social units, consisting of females and their juvenile offspring (Christal et al. 1998, Coakes and Whitehead 2004, Dufault and Whitehead 1998, Richard et al. 1996a). Although social units vary in size between different geographic areas (Richard et al. 1996a, Whitehead et al. 2012), in most areas it appears long-term female social units consist of approximately 12 adults (Best 1979, Whitehead et al. 1991). These units will often combine with others to form temporary social groups of approximately 23 individuals that can remain associated for several weeks before separating back into component units (Best 1979, Whitehead et al. 1991). Among social groups, there are group-specific dialects in terms of the particular acoustic codas used (Weilgart and Whitehead 1997). These differences in coda dialects between groups appear to persist over years, and are also influenced by geographic variation in the types of codas used (Weilgart and Whitehead 1997).

In contrast, males generally disperse from their natal social units at an age of 3-15 years (Best 1979, Richard et al. 1996a, Whitehead 2003). Adolescent males form loose, size-assorted ‘bachelor groups’ (Letteval et al. 2002), but as they age they become increasingly solitary (Whitehead 2003). As males attain larger body sizes, they increase their latitudinal range into polar waters (Allen 1980, Best 1979, Whitehead 2003). It is thought that only the largest males have the diving capacity to hunt for prey at these higher latitudes (Best 1979, Whitehead 2003). After reaching sexual maturity, sometime

after 25-27 years old, males begin to breed (Best 1979), but do not remain permanently associated with any given female social group (Richard et al. 1996a, Whitehead 1993; 1994).

Male dispersal and close-knit sociality of female units in the sperm whale have led researchers to draw parallels between this species and elephants (Family Elephantidae) (Best 1979, Whitehead 2003). In elephants, female social units are matrilineal, consisting of a matriarch and her offspring (Vidya and Sukumar 2005), a social organization that some researchers believed was also present in sperm whales (Whitehead et al. 1991). However, this assumption has not been borne out by molecular ecology studies, which showed a mixture of matrilines present in social groups based on the maternally-inherited mitochondrial DNA (mtDNA), as well as a mixed matrilines within the component social units of social groups (Christal et al. 1998). These results are inconsistent with either social units or groups as purely matrilineal social structures (Christal et al. 1998, Mesnick et al. 1999, Richard et al. 1996a). However the lack of matrilineality has been suggested to be an artifact of limited geographic sampling or ongoing social disruption resulting from whaling (Whitehead 1999).

Both sexes of sperm whales are known to strand en masse. These ‘mass strandings’ involve as many as 72 apparently healthy individuals, and almost inevitably result in death of the individuals involved (Evans et al. 2002, Rice et al. 1986, Stephenson 1975, van Helden 2009). Although there has been recent concern that mass strandings in some cetacean species could be linked to the use of military and commercial sonar (Jepson et

al. 2003, Moore and Early 2004), mass strandings of some cetacean species, including the sperm whale, have occurred throughout human history. Although there is no universal consensus on the cause of sperm whale mass strandings, social cohesion could be one contributing factor (Whitehead 2003). This has led to the development of models investigating the impact of social groups being removed from local populations on levels of genetic diversity (Alexander 2006, Siemann 1994).

1.5. Patterns of genetic differentiation in the sperm whale

With the cessation of whaling, other methods of defining geographic structure in the sperm whale have become increasingly prevalent, including the use of genetic markers (Drouot et al. 2004a, Engelhaupt et al. 2009, Lyrholm and Gyllensten 1998, Lyrholm et al. 1999, Mesnick et al. 2011, Pinela et al. 2009, Rendell et al. 2012, Whitehead et al. 1998). In the sperm whale, variation among the maternally-inherited mtDNA control region (mtDNA CR) suggests some restriction in maternal gene flow between ocean basins (Lyrholm and Gyllensten 1998). This is expected given female home ranges are relatively limited in comparison to oceanic scales (Dufault et al. 1999, Jaquet et al. 2003, Whitehead 2001). Nuclear differentiation (based on microsatellite genotypes), in contrast, shows no differentiation between ocean basins. This has been interpreted as supporting the hypothesis of male dispersal and relative female philopatry in the sperm whale (Lyrholm et al. 1999).

Given evidence for limited dispersal of females based on photo identification (Whitehead 2001), mtDNA CR differentiation is also expected *within* ocean basins. However, only a

limited number of studies have detected significant genetic differentiation within oceans (Drouot et al. 2004a, Engelhaupt et al. 2009, Lyrholm and Gyllensten 1998, Mesnick et al. 2011). Differentiation appears to be more marked in the Atlantic, due to the presence of the isolated marginal seas of the Gulf of Mexico and the Mediterranean, than in the Pacific (Drouot et al. 2004a, Engelhaupt et al. 2009, Lyrholm and Gyllensten 1998, Mesnick et al. 2011, Rendell et al. 2012, Whitehead et al. 1998). Although regional differentiation within the Pacific is weak, strong mtDNA and nuclear differentiation has been observed among social groups (Lyrholm and Gyllensten 1998, Rendell et al. 2012). Similar analyses have also shown strong differentiation among social groups within the Atlantic (Engelhaupt et al. 2009). Regional genetic differentiation and differentiation by social group has not been examined within the Indian Ocean.

A potential explanation for the differing levels of genetic structure in the Pacific and Atlantic Oceans relies on 'acoustic clans'. Acoustic clans are comprised of social groups that use similar vocal coda repertoires. Previous research has suggested that female gene flow might occur only within such clans (Rendell et al. 2012, Whitehead 1999). In the Atlantic, acoustic clans are allopatrically distributed (Whitehead et al. 2012) so gene flow within clans would also result in geographically based genetic structure. However, in the Pacific, acoustic clans are sympatrically distributed (Whitehead et al. 2012). This could lead to gene flow across broad geographic scales, reducing the amount of geographically based genetic structure. Differences in acoustic clan structure between oceans could be driven by differences in oceanography, whaling, predation and/or culture (Whitehead et al. 2012). However, there are potential issues linking acoustic clans to genetic

differentiation including limited sample sizes, variation in coda production between individuals within a social unit not being accounted for, and failure to find significant proportions of genetic variance explained by acoustic clan (Rendell et al. 2012, Rendell and Whitehead 2003, Whitehead et al. 1998).

Overall, the processes shaping patterns of genetic differentiation in the sperm whale appear to involve female philopatry and male-biased dispersal, which could be influenced by resource availability (e.g. oceanography). In addition, codas and acoustic clans appear to have the potential to shape patterns of geographically-based genetic structure. The patterns of acoustic structure might also have been subject to disruption due to whaling, which differed in intensity by ocean basin (Whitehead et al. 2012). Finally, underlying geographic structure in the sperm whale are strong patterns of genetic differentiation among female social groups. Although these groups are not strictly matrilineal, the long-term social bonds and strong patterns of genetic differentiation between groups suggests the influence of matrifocal social structure.

1.6. Low mtDNA diversity in the sperm whale

Initial surveys of mtDNA differentiation in the sperm whale reported a surprising pattern. Sperm whales show strikingly low levels of mtDNA CR diversity in comparison with other cetaceans, even those such as the humpback whale (*Megaptera novaeangliae*), Antarctic blue whale, and southern right whale (*Eubalaena australis*) that were subjected to far greater whaling pressure than the sperm whale (Table 1.1; Lyrholm et al. 1996;

Whitehead 1998). Several hypotheses to explain the low mtDNA CR diversity in sperm whales have been proposed, including:

(1) Lineage extinctions relating to the mass stranding of matrilineal groups (Alexander 2006, Siemann 1994) or due to geographic variation in ocean productivity (Tiedemann and Milinkovitch 1999);

(2) Selective sweeps of mtDNA linked with beneficial maternal cultural innovations in matrilineal social groups (Whitehead 1998; 2005);

(3) Constraints on substitutions at certain sites in the mtDNA CR, based on Lyrholm et al.'s (1996) observation of non-random mtDNA CR substitutions confined to "hot spots";

(4) Slow rates of molecular evolution in the sperm whale mtDNA (Lyrholm et al. 1996, Whitehead 1998);

(5) An 'exploitation bottleneck' due to the extensive and systematic hunting of sperm whales. However, previous authors have discounted the possibility of hunting pressure causing a loss of mtDNA CR diversity due to high global sperm whale abundance (Whitehead 1998; 2002);

(6) A population expansion/population bottleneck (Lyrholm and Gyllensten 1998, Lyrholm et al. 1996) resulting from forces that pre-date whaling;

(7) Selective sweeps associated with beneficial substitutions in mtDNA protein-coding genes (Janik 2001).

1.7. 'Voyage of the *Odyssey*'

The analysis of genetic structure and diversity in the sperm whale has been limited by the (a) number and (b) geographic location of samples, due to the difficulty of collecting biopsy samples in deep offshore waters (Table 1.2). In this dissertation, I have the opportunity to use a dataset that addresses both of these factors that limited the scope of previous studies. The 'Voyage of the *Odyssey*' (Figure 1.1; http://www.pbs.org/odyssey/voyage/index.html) was a five-year expedition sponsored by the Ocean Alliance (http://www.whale.org/the-mission/about-us/). During the voyage, 900 non-lethal sperm whale biopsies were collected to investigate baseline oceanic contaminant levels. The sperm whale was chosen as an indicator species due to its position high on the trophic chain, and therefore potential to bioaccumulate contaminants. As part of this study, DNA extracted from these samples was provided to the Cetacean Conservation and Genomics Laboratory at Oregon State University to perform genetic analyses.

As well as scope and number of samples, the 'Voyage of the *Odyssey*' was also unique in that it targeted equatorial waters associated with historical sperm whaling 'grounds', occupied primarily by female social groups (Figure 1.1). These areas have been under-sampled in previous research, particularly in the western Pacific (Table 1.2). In addition, the *Odyssey* samples represent the first collection of samples from the subtropical/tropical areas of the Indian Ocean for population genetic analyses (*cf.* Table 1.2). This dataset was also complemented by samples from New Zealand, Samoa and Oregon stranding events (Figure 1.1). These included individually stranded whales, and examples of mass

strandings of both male and female social groups. In addition, biopsy samples from the Gulf of Mexico were included in analyses of nuclear and mitogenomic diversity (Chapter 5).

1.8. Dissertation scope and structure

Here I address a number of hypotheses proposed to describe patterns of diversity and differentiation in the sperm whale, by incorporating the sperm whale samples from equatorial waters and strandings (Figure 1.1) with other published surveys. I use the combined dataset to investigate specific questions relating to the influence of social structure and molecular evolution on genetic diversity in the sperm whale, as detailed in the following chapters:

*Chapter 2: Female philopatry drives high levels of regional structure in the globally distributed sperm whale (*Physeter macrocephalus*).* This chapter is intended for publication with Debbie Steel, Kendra Hoekzema, Sarah Mesnick, Daniel Engelhaupt, Iain Kerr, Roger Payne, and C. Scott Baker.

Chapter 2 examines patterns of genetic differentiation and diversity in the sperm whale, focusing largely on previously under-sampled areas, including equatorial regions and the Indian Ocean. Given the more 'clustered' sampling scheme of the *Odyssey*, I hypothesized fine-scale mtDNA structure would be detected in the sperm whale, and that evidence of sex-biased dispersal uncovered by previous studies would be confirmed. I also combined this dataset with previously published mtDNA data to re-evaluate levels of

mtDNA diversity in the sperm whale, with an increased number of samples and broader geographic coverage.

Chapter 3: Lenient matrilineality in groups of sperm whales from equatorial waters of the Pacific, Indian, and Atlantic Oceans. This chapter is intended for publication with Debbie Steel, Renee Albertson, Daniel Engelhaupt, Iain Kerr, Roger Payne, and C. Scott Baker.

Although high levels of regional structure were found using the *Odyssey* dataset in Chapter 2, previous research has shown that the social group is likely to be even more important in partitioning genetic diversity than geography. What is less clear is the mechanism driving the inter-group differentiation seen. In Chapter 3, I derive the 'standardized matrilineality index', a statistic for evaluating whether social groups from specific regions are more matrilineal than expected by chance i.e. whether groups have fewer mtDNA CR haplotypes than expected. I confirm previous studies showing matrilineality is not 'strict' but suggest instead that 'lenient matrilineality' can explain the levels of mtDNA differentiation seen between social groups. In addition, the observed levels of matrilineality are used to assess the 'cultural innovation' selective-sweep hypothesis as an explanation for low mtDNA diversity in the sperm whale.

Chapter 4 is reformatted from the publication: *Alexander A, Steel D, Slikas B, Hoekzema K, Carraher C, Parks M, Cronn R, Baker CS (2013) Low diversity in the mitogenome of sperm whales revealed by next-generation sequencing. Genome Biology and Evolution 5:113-129.*

To further investigate hypotheses for the low mtDNA diversity in the sperm whale, Chapter 4 involved the next-generation sequencing of 17 Pacific Ocean sperm whale mitogenomes to investigate substitution rates and control region constraints. Sperm whales were found not to have slow mitogenomic substitution rates in comparison with other cetaceans, discounting slow evolutionary rates as the mechanism responsible for low mtDNA diversity. In addition, the control region showed higher levels of mtDNA diversity, and similar patterns of evolution, in comparison to other regions of the mitogenome. This discounts control region constraints as being responsible for the low mtDNA CR diversity seen in the sperm whale.

Chapter 5: A bottleneck for Moby Dick*? Low mtDNA and nuclear diversity in the sperm whale.* This chapter is intended for publication with Kendra Hoekzema, Debbie Steel, and C. Scott Baker.

After discounting slow evolutionary rates and control region constraints, I investigated whether a population bottleneck, expansion or selective sweep were the most likely causes of the low mtDNA diversity observed in the sperm whale. Chapter 5 involved the deep sequencing of ~12,000 bp of nuclear DNA and mitogenomes from the sperm whale (n = 22), extending geographic coverage to the Gulf of Mexico. In addition, to act as a contrast, these markers were also sequenced for the pygmy sperm whale (n = 10). Patterns of diversity were contrasted between the two species and other cetacean species using estimates from the literature. Extended Bayesian skyline plots were also used to reconstruct demographic changes in the sperm whale and pygmy sperm whale. Based on these reconstructions, and low nuclear and mitogenomic diversity levels in the sperm

whale, I consider a population expansion the most likely explanation for low overall levels of genetic diversity in this species.

Chapter 6 synthesizes the results from the data chapters to develop the 'big bang/maternal drift' hypothesis. Following the recent 'big bang' expansion of sperm whales, marked female philopatry at levels ranging from the social group through broader geographic scales has led to maternally-mediated genetic drift. This has led to large differences in mtDNA haplotype frequencies between social groups, regions, and oceans, despite the low diversity seen in this species. In addition, I discuss the general implications of my doctoral research, as well as potential future research directions.

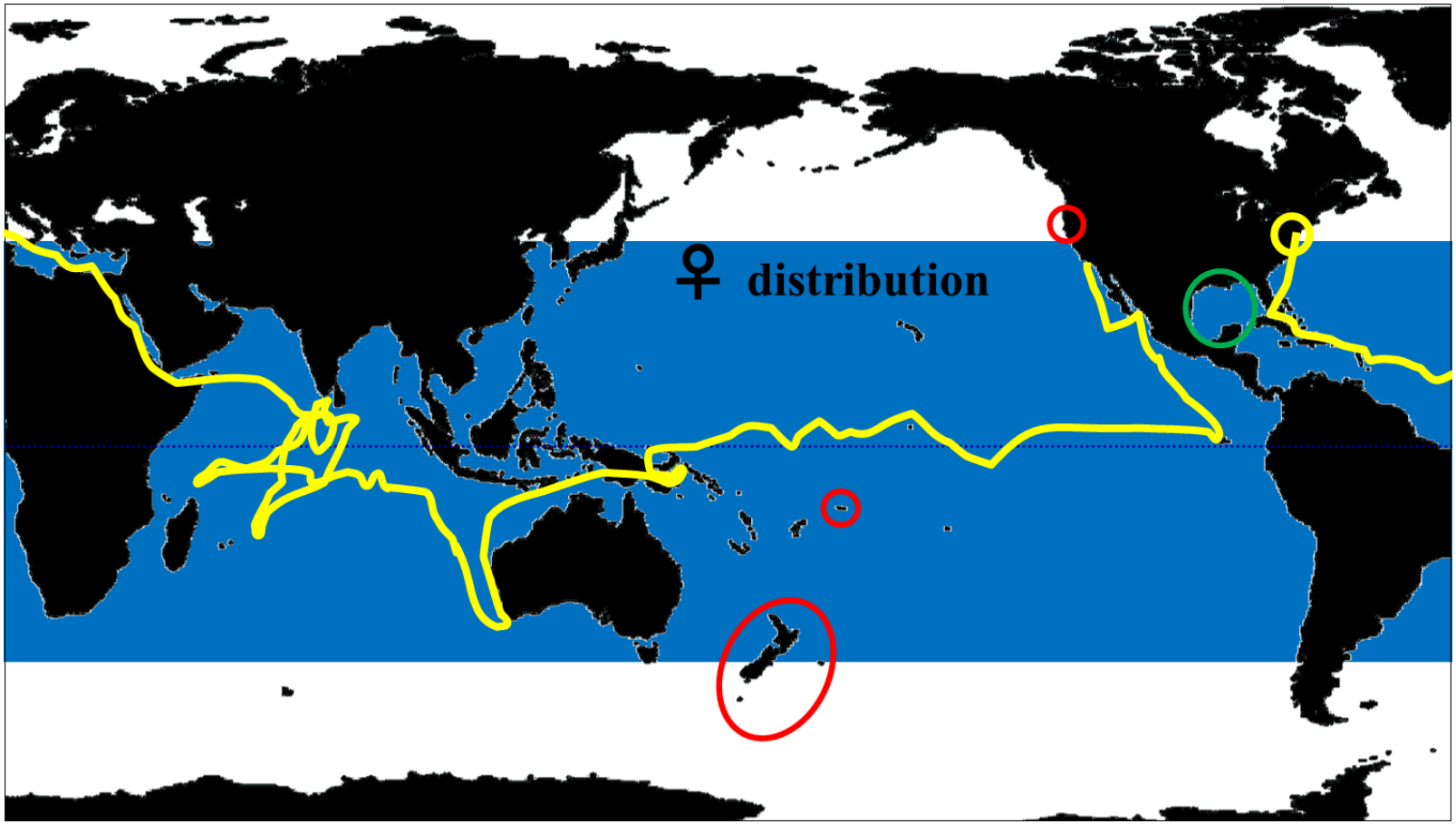


Figure 1.1: The location of samples used in genetic analyses in this dissertation, with the main area of female distribution shown in blue (40°N to 40°S). Track of the *Odyssey* shown in yellow (n = 895) samples, stranding sample locations shown in red (New Zealand n = 89, Samoa n = 1, Oregon n = 3) and Gulf of Mexico biopsy sample location shown in light green (n = 10).

Table 1.1: A comparison of mtDNA CR diversity across various cetacean species, ordered by nucleotide diversity. Sample size is designated by '*n*'.

Species	Population size	*n*	Number of haplotypes	Nucleotide diversity, π (%)	Length (bp)	References
Sperm whale (*Physeter macrocephalus*)	> 350,000 (Worldwide)	231	16	0.39	330	(Lyrholm and Gyllensten 1998, Whitehead 2002)
Hector's dolphin (*Cephalorhynchus hectori*)	3,000 – 4,000 (New Zealand South Island East Coast)	84	12	0.30	360	(Pichler and Baker 2000, Pichler et al. 1998)
Beluga (*Delphinapterus leucas*)	> 10,000 (North America)	628	39	1.00	234	(Brown Gladden et al. 1997)
Blue whale (*Balaenoptera musculus)*	~2,300 (Antarctica)	183	52	1.40	410	(Sremba et al. 2012)
Short-beaked common dolphin (*Delphinus delphis*)	> 180,000 (North Pacific)	18	17	1.80	404	(Rosel et al. 1994)
Bottlenose dolphin (*Tursiops truncatus*)	> 100,000 (Northwest Pacific)	17	12	1.90	386	(Wang et al. 1999)
Humpback whale (*Megaptera novaeangliae*)	> 11,000 (Worldwide)	90	37	2.57	463	(Baker et al. 1993)
Southern right whale (*Eubalaena australis*)	7,500 (Worldwide)	136	34	2.71	275	(Patenaude et al. 2007, Reilly et al. 2013)

Table 1.2: A summary of the geographic areas represented by previous 400 bp mtDNA CR analyses on sperm whales. Studies are included that explicitly reported the number of individuals sampled/samples collected (*n*). Within each ocean, regions are ordered from low latitude (areas likely to be dominated by female social groups) to high latitude (likely to be male dominated). Sample size per sex is given where available. For Mesnick et al. (2011), F includes juveniles, presumably of both sexes. M includes adult males, subadult males, and unknowns. Source: 1. Biopsy sample, 2. Sloughed skin, 3. Stranding/dead at sea, 4. Fishery bycatch, ? Tissue archive (source not given). For Lyrholm and Gyllensten (1998) and Mesnick et al. (2011), sample source for total dataset is presented for each region, as sample source not localized to region.

Geographic area	*n*	F	M	Source	Reference
Pacific					
Galapagos	38	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Galapagos, SW Pacific and Chile/Peru	194	--	--	2	Rendell et al. (2012)
Ecuador	56	44	12	2	Richard et al. (1996a)
Palmyra	4	4	0	1,2,3,4	Mesnick et al. (2011)
Western Baja	10	7	3	1,2,3,4	Mesnick et al. (2011)
South of Fiji	20	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Eastern tropical North Pacific	143	114	29	1,2,3,4	Mesnick et al. (2011)
Hawai'i	28	24	4	1,2,3,4	Mesnick et al. (2011)
Californian Current	52	32	20	1,2,3,4	Mesnick et al. (2011)
Off Japanese Coast (24-34° N)	29	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Western North Pacific(24-34° N)	34	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Central North Pacific	11	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Eastern North Pacific	31	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Southeast of Australia	14	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Aleutians	11	0	11	1,2,3,4	Mesnick et al. (2011)
Gulf of Alaska	33	1	32	1,2,3,4	Mesnick et al. (2011)
Kuril Islands	6	0	6	1,2,3,4	Mesnick et al. (2011)
Total	**714**	**226**	**117**		
Atlantic					
Dominican Republic	1	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Gulf of Mexico	76	55	21	1, 2	Ortega-Ortiz et al. (2012)
	153	107	42	1, 2	Engelhaupt et al. (2009)
Florida	1	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Western North Atlantic	84	38	20	1, 2	Engelhaupt et al. (2009)
Mediterranean	44	13	26	1, 2	Engelhaupt et al. (2009)
Azores	13	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Denmark	15	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Scotland	20	0	20	3	Engelhaupt et al. (2009)
Norway	8	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Sweden	1	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Iceland	8	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Total	**424**	**213**	**129**		
Indian Ocean					
Southern Indian Ocean	7	--	--	1, 2, ?	Lyrholm and Gyllensten (1998)
Total	**7**	--	--		

2. Female philopatry drives high levels of regional structure in the globally distributed sperm whale (*Physeter macrocephalus*)

2.1. Abstract

Sperm whales are globally distributed with a complex age-sex stratification: maturing males disperse from their natal groups to forage in cold temperate/polar waters. In contrast, females along with dependent young remain together in long-term, social groups in warm temperate and tropical waters. To date, few studies have surveyed genetic structure among equatorial regions, and none have included the Indian Ocean. Here, we assembled a worldwide database of mitochondrial DNA (mtDNA) control region haplotypes (394 bp) from 1,167 previously published samples. We combined this with samples from 49 stranded individuals and a new circum-equatorial collection of 476 individual DNA profiles (mtDNA, sex, 13 microsatellite loci), primarily from previously un-sampled equatorial waters, including the Indian Ocean. This combined dataset provided a more comprehensive estimate of differentiation among ocean basins and a finer-scale characterization of differentiation among subtropical/tropical regions within oceans. Using the combined dataset (n = 1,654), significant mtDNA differentiation was found between oceans (F_{ST} 0.119, $p < 0.001$) with pronounced differentiation among equatorial regions of the Indian Ocean (F_{ST} 0.314, $p < 0.001$) and marginal seas of the Atlantic (F_{ST} 0.469, $p < 0.001$). Regional differences were less marked in the Pacific (F_{ST} 0.061, $p < 0.001$), consistent with expansion of formerly allopatric acoustic clans following disruption of populations by whaling. Using the circum-equatorial samples, we also found weak but significant nuclear differentiation between oceans (G”$_{ST}$ 0.016, $p <$

0.001), and among regions within oceans (e.g. Indian Ocean G''_{ST} 0.018, $p < 0.05$; Atlantic 0.052, $p < 0.05$; Pacific 0.012, $p < 0.05$), consistent with male-biased dispersal and gene flow. There was no evidence of phylogeographic structure for the mtDNA at oceanic or regional scales. The differentiation in haplotype frequencies without phylogeographic structure, and low nucleotide diversity of the mtDNA control region (global $\pi = 0.442\%$), suggests a recent global expansion of a single maternal lineage followed by strong local drift due to maternal social organization (a 'maternally-mediated big bang').

2.2. Introduction

Although much effort has been made to ascertain population structure in the globally-distributed and abundant (>360,000 individuals, Whitehead, 2002) sperm whale (*Physeter macrocephalus*), efforts have been hampered by various sampling and biological issues (Dufault et al. 1999). These include the presence of long-range movements, sociality, culture, low mitochondrial DNA (mtDNA) diversity, as well as limited sample sizes and a lack of fine-scale geographic collection of samples. Based on catch records from open-boat ('Yankee') whaling in the 18th-19th century and modern whaling in the 20th century, sperm whales show a strong latitudinal pattern in the distribution of sexes. Females are found in low latitude, tropical and warm temperate waters, and maturing males migrate to high latitude, cold temperate and polar waters (Best 1979, Dufault et al. 1999). Female social groups (also called 'mixed-sex' groups) are thought to consist of long-term (on the order of decades) social units, consisting of females and their juvenile offspring of both sexes (Christal et al. 1998, Coakes and

Whitehead 2004, Dufault and Whitehead 1998, Gero et al. 2013b, Richard et al. 1996a, Whitehead et al. 2012). In contrast, males generally disperse from their natal social units at an age of 3-15 years, returning to warmer waters to breed at around 25 years old (Best 1979, Richard et al. 1996a, Whitehead 2003). Female home ranges (based on behavioral observations) are relatively limited in comparison with those of males and also relative to the ocean basins within which they reside (Dufault et al. 1999, Jaquet et al. 2003, Whitehead 2001).

Attempts to determine stock structure in the sperm whale were particularly contentious during efforts by the International Whaling Commission (IWC) to manage commercial whaling. This was due in part to evidence of long-distance movements, based largely on '*Discovery* marking'. These movements, predominantly of male sperm whales, included individuals moving between the Azores and Spain/Iceland; Japan and the Gulf of Alaska; Mexico to 1,300 km offshore from the Washington/Oregon border; and other large scale (3,200+ km) north/south movements in the Atlantic and Pacific (Aguilar 1985, Brown 1981, Donovan 1991, Dufault et al. 1999, Ivashin 1981, Ivashin and Rovnin 1967, Martin 1980, Mizroch and Rice 2013). These movements contrasted with evidence of restricted dispersal of individuals, also based on whaling data, including differences by area in the size distribution of whales, differences in prey species present in stomachs, as well as the recovery of '*Discovery* marks' in areas close to where individuals were originally marked (Allen 1980, Best 1979, Brown 1967, Clarke 1972, International Whaling Commission 1980, Kasuya and Miyashita 1988). Areas of local depletion, apparently due to whaling, also suggest restricted patterns of dispersal in this species (Whitehead et al. 1997).

With the cessation of commercial whaling, studies of population structure shifted to non-lethal methods including photo-identification, analysis of acoustic codas and genetics. Results generally agreed with whaling data that only limited interchange of individuals occurred between ocean basins, although there are rare instances of direct interchange (summarized in Dufault et al. 1999). Within oceans, there is an indication of site fidelity of marked animals based on photo-identification (Childerhouse et al. 1995, Gero et al. 2013b, Jaquet et al. 2003) and therefore the potential for differentiation between regions to occur within oceans. Genetic data indicates similar patterns at the oceanic level, with analyses of the maternally-inherited mitochondrial DNA control region (mtDNA CR) suggesting a restriction in female gene flow between ocean basins (Lyrholm and Gyllensten 1998). This differentiation was detected despite strikingly low levels of mtDNA CR nucleotide diversity in comparison with other cetaceans (Alexander et al. 2013, Lyrholm et al. 1996, Whitehead 1998). In contrast, no significant nuclear differentiation (based on microsatellite genotypes) has been detected between ocean basins. This has been interpreted as supporting the hypothesis of male dispersal and relative female philopatry (Lyrholm et al. 1999).

Given the limited home range of females based on photo-identification (circa 1,000 km, Whitehead 2001) and previous observations of site fidelity based on photo-identification, mtDNA differentiation should also be expected *within* ocean basins. However, mitochondrial differentiation among regional populations within the Atlantic and Pacific Oceans is weak, except for marginal seas such as the Gulf of Mexico and the Mediterranean within the Atlantic (Drouot et al. 2004a, Engelhaupt et al. 2009, Lyrholm

and Gyllensten 1998, Lyrholm et al. 1999, Mesnick et al. 2011, Rendell et al. 2012, Whitehead et al. 1998). A potential issue is that previous investigations into within-ocean regional structure have been forced to focus on single oceans or to aggregate samples over relatively large scales (thousands of kilometers), due to the logistical difficulty of collecting biopsy samples in deep offshore waters. The aggregation of samples over large areas could potentially mask finer-scale regional patterns.

Paralleling the differences in geographically-based genetic structure within the Pacific and Atlantic, differences in 'acoustic clan' structure have also been identified (Whitehead et al. 2012). An acoustic clan is comprised of different female social groups that show similar repertoires of stereotypical series of clicks referred to as 'codas' (Rendell et al. 2012, Watkins and Schevill 1977, Whitehead et al. 1998). In the Pacific, these acoustic clans appear to be sympatrically distributed, in contrast with the Atlantic where they show an allopatric distribution (Whitehead et al. 2012). It has been suggested that females in the Pacific could show 'social philopatry' to an acoustic clan, rather than geographic philopatry (Rendell et al. 2012). Supporting this, the extent of mtDNA differentiation explained by geographic scale in the Pacific seems to be less than that explained by aggregating samples based on acoustic clan, although neither factor explained a significant proportion of variation (Rendell et al. 2012). To date, regional population structure has not been investigated in the Indian Ocean.

As well as differences in acoustic clan structure, characteristics of female social groups also differ by ocean, including group size and proportion of calves (Whitehead et al.

2012). It is assumed, however, that a consistent pattern in female social groups is the presence of close kin. Kinship within social groups could result in 'non-random sampling' for the purposes of describing geographic structure (Dufault et al. 1999). To account for this potential bias, researchers have used a number of approaches e.g. limiting the sample to only one individual from each group (Lyrholm and Gyllensten 1998), removing individuals that appear to be closely related based on DNA genotypes (Engelhaupt et al. 2009, Mesnick et al. 2011) and accounting for multiple hierarchical levels when partitioning genetic diversity using a nested AMOVA (Rendell et al. 2012).

Here we assemble the largest sperm whale genetic dataset to date, by using both published and previously unpublished data to expand on the number of samples and geographic scope that have limited previous studies. Previously unpublished data included stranding samples, and samples collected by the 'Voyage of the *Odyssey'*: a five year expedition that collected biopsy samples (skin and blubber) from previously under-sampled equatorial regions, including the Indian Ocean (Godard et al. 2003). The *Odyssey* sampling was carried out at relatively discrete (~500 km) intervals within each ocean (Figure 2.1) during circumnavigation of historical 'sperm whaling grounds' in subtropical/tropical waters. Using mtDNA and nuclear microsatellite loci, this dataset provided fine-scale analyses of female and male-mediated gene flow. Previously published data included mtDNA CR haplotypes from 1,167 samples (Engelhaupt et al. 2009, Lyrholm and Gyllensten 1998, Mesnick et al. 2011, Ortega-Ortiz et al. 2012, Rendell et al. 2012, Richard et al. 1996a, Whitehead et al. 1998). Together, the published and previously unpublished data covered most of the known range of males and females.

Aggregation of these datasets was possible because of the concerted efforts of the Cachalote Consortium (Mesnick et al. 2005) to standardize nomenclature for mtDNA CR haplotypes over previous population genetic studies. We used this extensive collection of samples to conduct a global analysis of mtDNA differentiation and diversity.

2.3. Materials and Methods

2.3.1. Assembly of published mtDNA dataset with samples from strandings and the Voyage of the Odyssey

Using the sperm whale mtDNA CR (394 bp) haplotype definitions developed by Mesnick et al. (2005, also see Appendix I, Table I.1), we summarized haplotype information from previous publications by ocean and within-ocean region (Appendix I, Table I.2). In contrast with the *Odyssey* dataset, the lack of standardized nuclear markers did not allow for identification or removal of potential between-study replicates, or the identification of reuse of subsets of samples from other previous studies. However, within-study replicates were removed where identified.

'Voyage of the *Odyssey*' samples were collected from 1999-2005 in circum-equatorial regions (Figure 2.1) using a biopsy dart. Total genomic DNA was extracted as detailed by Godard et al. (2003). DNA aliquots of 895 samples were provided courtesy of Ocean Alliance, sponsor of the 'Voyage of the *Odyssey*'. In addition to the *Odyssey* samples, DNA was also extracted from samples of stranded animals. New Zealand sperm whale skin and tissue samples (n = 89) were collected from strandings by New Zealand Department of Conservation staff from 1994 to 2008 and archived in the New Zealand

Cetacean Tissue Archive (CeTA) at the University of Auckland. One sample originating in Samoa, archived in CeTA, was also included. Oregon sperm whale skin and tissue samples (n = 3) were provided by the Oregon Marine Mammal Stranding Network. DNA was extracted for these stranding samples following a standard phenol/chloroform technique (Sambrook et al. 1989) as modified by Baker et al. (1994).

Amplification of the mtDNA CR using the primers M13dlp1.5 and tphe, and sequencing of a 619 bp consensus length of this fragment on an ABI3730xl DNA analyzer, was carried out as described in Alexander et al. (2013). Sequences were trimmed using PHRED scores and by eye in *Sequencher v. 4.6* (Gene Codes). After trimming, sequences with more than 10% of bases showing a PHRED score of <20 were re-sequenced or removed from the dataset (Morin et al. 2010b). Variable sites between haplotypes were visually confirmed in each sequence using *Sequencher*. After removal of replicates, *Odyssey* and stranding samples were trimmed to the shorter consensus length of 394 bp and combined with the previously published mtDNA data.

2.3.2. **Odyssey** *and stranding sample sex identification*

Samples were sexed using a multiplexed PCR amplifying 152 bp of the *SRY* on the Y chromosome of males (Richard et al. 1994, primers: sperm-whale specific *SRY* primers), and a 442-445 bp fragment of the *ZFX/ZFY* fragment present in both males and females (Aasen and Medrano 1990, primers: P1-5EZ and P2-3EZ). Each reaction consisted of 1 μL of sample DNA, and a final concentration of 0.9× Platinum *Taq* buffer (Invitrogen), 0.36 μM of each of the four primers, 2.27 mM $MgCl_2$, 0.18 mM dNTP and 0.25U of

Platinum *Taq* polymerase (Invitrogen), with ddH_2O to 11 μL total volume. The temperature profile consisted of an initial denaturing step of 3 min at 94°C, followed by 35 cycles of 94°C for 45 s, 58°C for 45 s and 72°C for 60 s, followed by a final extension step of 72°C for 10 min. The PCR products were run on a 1.6% agarose gel (buffer: TBE), stained with ethidium bromide, and visualized under UV light. The presence of two bands indicated a male sample, one band a female sample, and no bands indicated a sex PCR failure.

2.3.3. Microsatellite genotyping, identification of replicates and kin

Thirteen microsatellite loci were selected based on previous genotyping in the sperm whale (Engelhaupt et al. 2009), and in other cetacean species. For the *Odyssey* and stranding samples genotyped in this study, each locus was amplified in an individual reaction, with 1 μL of the sample DNA, a final concentration of 0.9× Platinum *Taq* buffer (Invitrogen), 0.36 μM of each primer and 0.18 mM dNTP. $MgCl_2$ and Platinum *Taq* polymerase (Invitrogen) concentrations varied by locus as detailed in Appendix I, Table I.3, and ddH_2O was added to 11 μL total volume. Temperature profiles consisted of an initial denaturing step of 3-5 minutes at 94-95°C, followed by 35-40 cycles of 94-95°C for 30-40 s, the locus-specific annealing temperature (as detailed in Appendix I, Table I.3) for 30-60 s, and 72°C for 30-60 s, followed by a final extension step of 72°C for 8-30 min. Multiple microsatellite loci were combined based on differing size range and fluorescent label (Appendix I, Table I.3) and co-loaded on an ABI3730xl DNA Analyzer with GS500 LIZ ladder. Output was processed using *GeneMapper v. 3.7* (Applied Biosystems), with a minimum signal strength detection threshold of 50 units. All

automated calls were checked by eye, with a subset of samples cross-checked by a second researcher (D. Steel) to ensure consistency in allele calling.

Other quality control (QC) measures were then carried out following the recommendations of Morin et al. (2010b). Samples were only included if they were genotyped for at least eight microsatellite loci, and successfully sexed or sequenced for mtDNA CR. This QC was to limit the inclusion of samples likely affected by low DNA quantity/quality. Identification of replicate samples was carried out using *Cervus v. 3.0* (Kalinowski et al. 2007). To compensate for genotyping error, we used relaxed matching allowing for mismatches at up to four microsatellite loci, with mismatching loci corrected or repeated. If remaining mismatches were consistent with allelic dropout, samples were considered replicates if they matched at sex and mtDNA CR. Probability of identity ($p_{(ID)}$) was calculated using *Cervus* for pairs showing exact matches, and *Genalex v. 6.501* (Peakall and Smouse 2006, Peakall and Smouse 2012) for those with mismatches (average $p_{(ID)}$ for the combination of exactly-matching markers)**.** The per-allele microsatellite error rate (Pompanon et al. 2005) was estimated using replicates for 110 samples, selected randomly with respect to DNA quality and quantity. Tests for deviation from Hardy Weinberg and linkage disequilibrium were conducted using *Genepop v. 4.2* (Raymond and Rousset 1995a), and tests for the presence of large allele dropout and null alleles using *MICRO-CHECKER v. 2.2.3* (van Oosterhout et al. 2004), following the methods of Carroll et al. (2011). Microsatellite loci were excluded from population genetic analyses if they showed departures from Hardy-Weinberg equilibrium.

A ‘kin restricted’ dataset was also created to account for the sampling of first-order relatives within social groups. For this, we followed Mesnick et al.’s (2011) approach of removing one member of every first-order kin pair (defined as samples sharing at least one allele at every microsatellite locus). To identify pairs of first-order kin, we used *SOLOMON v. 1.0-1* (Christie et al. 2013). The sample with the most complete genotype in each pair was retained. Microsatellite analyses were conducted on both the ‘full’ and ‘restricted’ version of this dataset.

2.3.4. mtDNA diversity, differentiation, and phylogeography

Haplotype and nucleotide diversity (using the Tamura and Nei (1993) correction) were calculated using *Arlequin v. 3.5* (Excoffier et al. 2005). Differences in haplotype diversity between oceans/regions were assessed using a custom *R v. 3.0.2* (R Core Team 2013) script to conduct a permutation test with 1,000 replicates (Appendix I, Figure I.1). Differences in nucleotide diversity were tested using a t-test (Nei 1987) with the standard deviation of nucleotide diversity output by *Arlequin* and degrees of freedom equal to the combined sample size between areas minus two (Patenaude et al. 2007).

A parsimony network to infer the number of substitutions between the mtDNA CR haplotypes observed in this study was created using *TCS v. 1.2.1* (Clement et al. 2000). F_{ST} and Φ_{ST} (using the Tamura and Nei (1993) correction), with 10,000 replicates to assess significance, were calculated using *Arlequin* (Excoffier et al. 2005). To evaluate the potential influence of phylogeographic structure (i.e. divergence as well as drift), we tested differences between G_{ST} and N_{ST} (analogs of F_{ST} and Φ_{ST}: Pons and Petit 1996)

using *PERMUT v. 2.0* (Petit 2010). The presence of non-*a priori* mtDNA structure was investigated using *Alleles in Space (AIS) v 1.0* (Miller 2005), restricted to the *Odyssey* samples with specific spatial coordinates. The following analyses were used: the allelic aggregation index, interpolation of the genetic landscape shape, and Monmonier's (1973) maximum difference algorithm.

2.3.5. Microsatellite diversity and differentiation

For the *Odyssey* microsatellite dataset, observed and expected heterozygosity were calculated using *Cervus*, and allelic richness using *FSTAT v. 2.9.3* (Goudet 2001). To test for significant differences in observed heterozygosity and allelic richness between geographic areas, a custom *R* script (Hamner 2014) was used to implement t-tests or Wilcoxon signed rank tests depending on equality of variances and normality of differences between areas (Appendix I, Figure I.2). F_{IS} by region and ocean (using oceanic and global microsatellite allele frequencies, respectively) was calculated using *FSTAT*, using 10,000 replicates to assess significance. F_{ST} was calculated using *Genepop* (Raymond and Rousset 1995b, Rousset 2008) with the exact test (Raymond and Rousset 1995a) used to assess significance. G''_{ST}, an index that compensates for the diversity of microsatellites (Meirmans and Hedrick 2011), was calculated with *Genodive v. 2.ob25* (Meirmans and van Tienderen 2004), using 10,000 permutations to assess significance. The presence of non-*a priori* population structure was assessed for the microsatellite dataset using *structure v 2.3.4* (Falush et al. 2003, Pritchard et al. 2000). Following Engelhaupt et al. (2009), admixture and correlated allele frequencies were assumed with 500,000 burn-in steps, followed by 1,000,000 steps. Twenty replicates (following the

recommendations of Gilbert et al. 2012) were carried out for $K = 1$ to $K = 13$. The Evanno et al. (2005) method was used to assess the best fitting K through *STRUCTURE HARVESTER* (Earl and vonHoldt 2012). Using *STRUCTURE HARVESTER* output, *CLUMPP* (Jakobssen and Rosenberg 2007) was used to align cluster assignment across replicates.

2.3.6. Isolation by distance

Isolation by distance analyses were carried out for microsatellites and mtDNA CR using *Genalex*, partitioning by within-ocean regions (both at the worldwide scale and within oceans). Geographic distance between regions was calculated as the closest distance between samples from different regions, accurate to the nearest 100 km. This was measured using a custom script (Appendix I, Figure I.3) in *R*, with validity of paths (i.e. not crossing continental landmasses) assessed using *Google Fusion Tables***.** Following Rousset (1997), isolation by distance analyses were conducted using the logarithm of distance against various measures of genetic distance: $F_{ST}/(1 - F_{ST})$ and $\Phi_{ST}/(1 - \Phi_{ST})$ for the mtDNA CR datasets, and $F_{ST}/(1 - F_{ST})$ and $G''_{ST}/(1 - G''_{ST})$ for the microsatellite dataset.

2.3.7. Sex-biased gene flow and dispersal

Analyses of sex-biased gene flow and dispersal were restricted to the samples genotyped in this study, where sex information was available. Explicit tests for sex-biased dispersal were carried out at the oceanic and within-ocean regional level (including all regions with at least 2 individuals of each sex), following the methods of Oremus et al. (2007). To

investigate sex-biased dispersal for mtDNA CR (coding the mtDNA CR as a homozygote locus), and microsatellites, two methods were used in *FSTAT*: (1) a comparison of sex-specific F_{ST} values for both mtDNA and microsatellites and (2) calculation of the sex-specific variance of assignment index (vAIc) based on microsatellites (Goudet et al. 2002). The difference between sex-specific values was tested using 10,000 permutations. The more dispersive sex is expected to have a lower F_{ST} value (method 1), but higher variance (method 2), than the more philopatric sex (Oremus et al. 2007). It should be noted that males in this dataset included immature males that had not dispersed from their natal social group. This is likely to conservatively bias the tests against finding male-biased dispersal. As well as sex-biased dispersal, we calculated sex-specific gene flow using the formulas presented in Hedrick et al. (2013). The input values used were the microsatellite/mtDNA CR F_{ST} values based on the worldwide dataset partitioned by oceans and within-ocean regions, and the individual oceans partitioned by region.

2.3.8. Data archiving statement

For each sample genotyped in the '*Odyssey*' dataset, general location (as well as latitude/longitude where available), individual ID code, microsatellite genotype, and a letter code denoting the mtDNA CR haplotype are archived as detailed in Appendix I. Sequences of each defined haplotype are also archived as detailed in Appendix I.

2.4. Results

2.4.1. Assembly of global mtDNA dataset and definition of spatial scales

Sequence information was summarized from 1,167 samples from previous studies (Engelhaupt et al. 2009, Lyrholm and Gyllensten 1998, Mesnick et al. 2011, Ortega-Ortiz et al. 2012, Rendell et al. 2012, Richard et al. 1996a, Whitehead et al. 1998 as detailed in Appendix I, Table I.2). After removal of replicates, the mtDNA CR of 487 *Odyssey* and stranding individuals were trimmed to a consensus sequence length (394 bp) and combined with this previously published information. This resulted in 1,654 sequences included in analyses of mtDNA differentiation and diversity at the global and oceanic (the Pacific, Indian and Atlantic Oceans) level (Table 2.1).

In addition to analyses at the global and oceanic level with these samples, we also carried out within-ocean regional analyses of mtDNA diversity and differentiation. We restricted these regional analyses to tropical/subtropical areas (38°S to 38°N), as these were the latitudes primarily sampled by the *Odyssey.* The Mediterranean, although higher-latitude, was also included due to its IUCN endangered status (Notarbartolo di Sciara et al. 2012). Regions were defined by aggregating samples that were obtained within ~500 km of each other, with the exception of the Mediterranean that was pooled over the entire sea for consistency with previous publications (Engelhaupt et al. 2009). Regions were also required to have five or more sampled individuals to limit the effect of low sample sizes. This stratification resulted in 1,074 samples divided into 16 regions, including the Mediterranean (Table 2.1). Regional analyses were conducted within each ocean basin and at a worldwide level.

2.4.2. mtDNA diversity, differentiation, and phylogeography

A total of 39 mtDNA CR haplotypes were resolved in the global dataset, including twelve previously unreported haplotypes (Figure 2.2). Except for KK, all of these new haplotypes were rare ($n < 5$) and only found in one region (Figure 2.3). The maximum distance between any two haplotypes was two substitutions and this only occurred twice on the haplotype network (Figure 2.2). Of the 31 variable sites found over the 394 bp mtDNA CR, all were transitions (Appendix I, Table I.1). To investigate the potential for resolving further mtDNA diversity, 400 samples were sequenced for 619 bp of the mtDNA CR. However, a comparison of the two consensus lengths indicated 394 bp captured the majority of variation (Appendix I, Figure I.4). Therefore, even with the addition of Indian Ocean samples, mtDNA CR diversity in the sperm whale remains low compared to other cetaceans (Table 2.1 vs. cetacean mtDNA diversity estimates in Alexander et al. 2013).

Using the haplotype diversity permutation test (Appendix I, Figure I.1), at the oceanic level the Atlantic Ocean had significantly lower haplotype diversity ($h = 0.746$) than the other two oceans (Pacific $h = 0.779$; Indian $h = 0.788$, $p < 0.037$). Within the Atlantic, the Mediterranean ($h = 0.00$, $\pi = 0.00\%$) had significantly lower haplotype diversity than all other regions, unsurprisingly as the Mediterranean showed only a single mtDNA haplotype (Table 2.1). In addition, the Gulf of Mexico ($h = 0.506$, $\pi = 0.216\%$) had lower haplotype diversity ($p < 0.046$) than the Canary Islands ($h = 0.648$, $\pi = 0.333\%$) and western North Atlantic ($h = 0.616$, $\pi = 0.273\%$; Table 2.1). Overall, the Canary Islands and western North Atlantic had higher levels of genetic diversity than the Mediterranean

and Gulf of Mexico (Table 2.1). In the Indian Ocean, Sri Lanka (h = 0.382, π = 0.132%) had significantly lower haplotype diversity than all other regions ($p < 0.003$), except for the Cocos Island (h = 0.451, π = 0.230%). Cocos Island had significantly lower haplotype diversity ($p < 0.0212$) than southwestern Australia (h = 0.791, π = 0.307%), the Seychelles (h = 0.716, π = 0.413%) and Aldabras (h = 0.712, π = 0.364%). Finally, the Maldives/Chagos Archipelago (h = 0.570, π = 0.303) had significantly lower haplotype diversity than southwestern Australia and the Seychelles (p = 0.004). Overall, southwestern Australia, the Seychelles, and Aldabras had the greatest levels of genetic diversity, with Sri Lanka and Cocos Island the least (Table 2.1). Within the Pacific, the Gulf of California (h = 0.788, π = 0.368%) had significantly higher haplotype diversity than Hawai'i (h = 0.643, π = 0.197%, $p < 0.0132$) and the Pacific Crossing (h = 0.679, π = 0.301%, p = 0.045*).* Overall, the Gulf of California and Galapagos had the greatest levels of genetic diversity, and the Pacific Crossing and Hawai'i the lowest (Table 2.1, significance of all regional comparisons given in Appendix I, Figure I.5). No significant differences in nucleotide diversity were detected at the oceanic level or among regions within any ocean basin.

Some sharing of mtDNA haplotypes occurred across all three ocean basins, particularly of A, B, and C: the three most globally common haplotypes (Figure 2.2, Figure 2.3). Despite this, geographic structure was evident at the oceanic scale, with four private haplotypes found at reasonably high frequencies (in >20 individuals) yet restricted to a single ocean basin (haplotype X in the Atlantic, KK in the Indian Ocean, E and D in the Pacific Ocean: Figure 2.2, Figure 2.3). These results were reflected in the significant

inter-ocean mtDNA differentiation observed (F_{ST} 0.119, Φ_{ST} 0.241, $p < 0.001$; Figure 2.3). However, larger mtDNA differentiation values were obtained for within-ocean regional comparisons than at the oceanic level (Figure 2.3). Overall, the Atlantic Ocean showed the highest levels of intra-ocean mtDNA differentiation (F_{ST} 0.469, Φ_{ST} 0.468, $p < 0.001$; Figure 2.3), largely driven by the isolated marginal seas of the Mediterranean and the Gulf of Mexico (Table 2.2). However, significant mtDNA differentiation was also detected within the main Atlantic Ocean (western North Atlantic to the Canary Islands F_{ST} 0.109, Φ_{ST} 0.123, $p < 0.05$; Appendix I, Table I.4). A surprising finding was the high level of mtDNA differentiation within the Indian Ocean (F_{ST} 0.314, Φ_{ST} 0.412, $p < 0.001$; Figure 2.3), similar in magnitude to the levels seen between the marginal seas of the Atlantic. This appears to be driven in part by Sri Lanka (Table 2.2), which has a far higher frequency of the KK haplotype than surrounding regions (Figure 2.3). In contrast, far lower levels of mtDNA differentiation were found in the Pacific Ocean (F_{ST} 0.061, Φ_{ST} 0.067, $p < 0.001$; Figure 2.3). Although Kiribati and Hawai'i showed large numbers of significant pairwise comparisons to other regions, the values of these were lower in magnitude on average than the comparisons involving the Gulf of Mexico, Mediterranean, or Sri Lanka (Table 2.2). Other than private oceanic haplotypes, there was no evidence of a strong pattern of phylogeography in mtDNA. Values for F_{ST} and Φ_{ST} showed similar patterns (Figure 2.3, Table 2.2), and a test of these two indices (using N_{ST} as an analog of Φ_{ST}: Pons and Petit, 1996), showed no significant differences at any hierarchical level.

During examination of non-*a priori* patterns of mtDNA differentiation, the allelic aggregation index analysis performed using *AIS* indicated that mtDNA haplotypes were significantly more spatially clumped than expected under a random distribution of haplotypes ($p < 0.00001$). Genetic landscape features were consistent with grouping samples into the *a priori* within-ocean regions defined using geographic proximity (Appendix I, Figure I.6). In addition, no evidence of barriers to gene flow contradicting *a priori* grouping of regions by oceans was detected (Appendix I, Figure I.6). Using the Monmonier maximum difference algorithm, the only genetic break detected was within the Gulf of California (between individuals in close proximity, with different mtDNA haplotypes).

2.4.3. Assembly of microsatellite dataset and definition of spatial scales

Of the 988 total samples genotyped in this study, 670 passed quality control, with a minimum of 8 microsatellite loci each. Replicates were identified using between 6 and 13 overlapping loci with $p_{(ID)}$s between 3.39E-21 and 1.76E-06, and $p_{(ID\text{-}sibs)}$ between 2.16E-06 and 4.50E-03. The per-allele microsatellite error rate of 1.27% (based on 74 replicate pairs that passed QC) was largely due to allelic drop out (>95%), and was similar to previous studies on sperm whales and other cetaceans (Baker et al. 2013, Carroll et al. 2011, Mesnick et al. 2011). There was no detectable error in designation of mtDNA haplotypes (i.e. an error rate of < 0.7%), and one male/female discrepancy between a replicate pair (e.g. an error rate of 1.69%).

After removal of replicates, the 603 *Odyssey* genotypes that passed QC represented 476 individuals. Using *SOLOMON* to identify pairs of individuals sharing an allele at every microsatellite locus, we found 12 likely first-order kin relationships. One pair from the Chagos Archipelago consisted of two males. Of the remaining relationships, 11 were consistent with a mother-offspring relationship based on sex and mtDNA haplotype. These relationships included three individuals from the Gulf of California that were consistent with a grandmother-mother-offspring. Given the small number of identified potential-first order kin, results for the 'full' and 'restricted' datasets were very similar for all analyses. Consequently, results of the 'restricted' dataset are provided only as Appendix I, Table I.5.

Using the same criteria as for the mtDNA dataset (samples originating from the tropics/subtropics, collected over ~500 km spatial scales); we defined 13 within-ocean regions, including the Mediterranean, in the microsatellite dataset (Table 2.1). In contrast to the mtDNA dataset, three regions were not represented in the microsatellite dataset: Hawai'i, western North Atlantic, and the Gulf of Mexico. No significant deviations from Hardy-Weinberg equilibrium or linkage disequilibrium were detected consistently across the within-ocean regions. In addition, no evidence of scoring error, large allele dropout or null alleles were found for any microsatellite loci consistently across the within-ocean regions. Therefore, all loci were retained for analyses of microsatellite diversity and differentiation (Table 2.3).

2.4.4. Microsatellite diversity and differentiation

In contrast to the mtDNA diversity results, no significant differences were detected between oceans, or regions within oceans, for observed microsatellite heterozygosity or allelic richness using Wilcoxon signed rank and t-tests (depending on equality of variances and normality of differences between pairs) (Table 2.1; locus by locus results Appendix I, Table I.5). Significant differentiation was detected among oceans (F_{ST} 0.003, G''_{ST} 0.016, $p < 0.05$), but this was far lower in magnitude than that seen for mtDNA (Figure 2.3). As well as being the most differentiated at the oceanic level, the Atlantic had the highest within-ocean level of nuclear structure (F_{ST} 0.016, G''_{ST} 0.052, $p < 0.05$), mostly driven by the large number of significant pairwise comparisons involving the Mediterranean (Table 2.2, Appendix I, Table I.4). The Indian Ocean showed lower levels of nuclear structure (F_{ST} 0.005, G''_{ST} 0.018, $p < 0.05$) than the Atlantic Ocean, but higher levels of structure than the Pacific, consistent with results based on mtDNA differentiation (Figure 2.3). A large number of significant pairwise comparisons within this ocean basin involved Sri Lanka (Table 2.2, Appendix I, Table I.4). Finally, the Pacific Ocean showed the lowest overall levels of population structure based on microsatellites (F_{ST} 0.001, $p < 0.05$; G''_{ST} 0.012, N.S., Figure 2.3). Despite this, and surprisingly given the lower level of mtDNA differentiation seen for this area, Papua New Guinea had a large number of significant comparisons to other regions (Table 2.2). Overall, regions with the highest average fixation values were the Mediterranean, southwest Australia, Aldabras, and Kiribati (Table 2.2).

During examination of non-*a priori* patterns of microsatellite differentiation using *structure*, $K = 2$ was identified as the most likely number of clusters using the Evanno et al. (2005) method ($\Delta K = 9.05$ *cf.* $\Delta K = 7.69$ for $K = 3$). However, $K = 1$ had the largest likelihood (mean Ln Pr(X|K) = -18,855.3 *cf.* -18908.9 for $K = 2$). Visual inspection of the structure results for $K = 2$ (Appendix I, Figure I.7) showed no obvious population structure, offering further support for $K = 1$. The implication is that the weak structure revealed by our *a priori* partitions is neither supported nor contradicted by these *structure* results. However, it is not surprising that *structure* failed to detect significant clustering as it is known to perform relatively poorly when levels of genetic differentiation are low (Waples and Gaggiotti 2006).

2.4.5. Isolation by distance

An analysis of isolation by distance was conducted on pairwise comparisons of regions (over a worldwide scale, ignoring ocean) defined for the mtDNA and microsatellite datasets. For the mtDNA CR dataset (using the regions/sample sizes defined in Table 2.1), isolation by distance was weak but significant using Φ_{ST} ($R^2 = 0.028$, $p = 0.01$), and marginal for F_{ST} ($R^2 = 0.017$, $p = 0.06$). For the microsatellite dataset (using the regions/sample sizes defined in Table 2.1), isolation by distance over a worldwide scale was weak but significant for G''_{ST} ($R^2 = 0.039$, $p = 0.04$), and marginal for F_{ST} ($R^2 = 0.030$, $p = 0.06$). An important caveat is that significant isolation by distance was dependent on the inclusion of the Gulf of Mexico and the Mediterranean: regions with high levels of differentiation (Table 2.2) that are also, on average, geographically distant from all other regions (the 1st and 3rd most geographically-distant populations on

average). No significant isolation by distance was detected at the regional level within any of the ocean basins.

2.4.6. Sex-biased gene flow and dispersal

Most regions showed a significant skew towards females, consistent with the assumption that the *Odyssey* largely targeted equatorial social groups dominated by females (Table 2.1). Areas with an excess of sampled males included the Mediterranean, Maldives/Chagos Archipelago, and the Galapagos. Sex-specific estimates of gene flow, calculated from microsatellite and mtDNA F_{ST} values (Hedrick et al. 2013), were low for females and high for males (Table 2.4). The ratio of male N_em to female N_em was between 19:1 (partitioning the worldwide dataset by ocean) and 44:1 (partitioning the Indian Ocean by region, Table 2.4). Given the evidence for sex-biased gene flow, it is perhaps not surprising that tests for sex-biased dispersal indicated males are the more dispersive sex in sperm whales (Table 2.4). For mtDNA, both female-specific and male-specific tests of differentiation were significant at all hierarchical levels (Table 2.4, due to limited sample sizes, a within-ocean regional analysis was not conducted for the Atlantic). However, the magnitude of female-specific F_{ST} exceeded that of males at all levels, and was significantly greater when partitioning by regions over the worldwide dataset (Table 2.4). For the microsatellites, the female-specific estimates of differentiation were weak but significant at all levels. Male-specific differentiation was not significant at any level, perhaps in part due to the lower sample sizes (Table 2.4). The magnitude of female-specific F_{ST} was greater than male-specific differentiation except among regions within the Pacific (however, the female-specific F_{ST} was significantly

different from zero and the male-specific F_{ST} was not). Female-specific F_{ST} significantly exceeded that of males at the oceanic scale (Table 2.4). Surprisingly, the sex-specific variance of assignment tests were not significant.

2.5. Discussion

Even though the mtDNA dataset assembled in this study is one of the largest reported for any cetacean species (n = 1,654), and included samples from under-sampled regions, sperm whale mtDNA diversity remained strikingly low in comparison with many other cetacean species (global $\pi = 0.442\%$, *cf.* Table 1.1). Despite the low levels of mtDNA diversity, in the first regional population genetics study within the Indian Ocean, we found marked patterns of maternal structure. Levels of differentiation were similar to that seen in the Atlantic Ocean (this study; Engelhaupt et al. 2009), but in the absence of obvious geographic boundaries. In contrast, the Pacific Ocean showed far lower levels of regional mtDNA differentiation. This is consistent with previous studies that found no geographically-based mtDNA structure in the Pacific (Lyrholm and Gyllensten 1998, Lyrholm et al. 1999, Rendell et al. 2012, Whitehead et al. 1998), or significant, but low levels of differentiation (Mesnick et al. 2011). The circum-equatorial dataset was more limited in sample size and geographic scope (as it excluded temperate regions) than the combined mtDNA dataset, but the presence of sex and microsatellite genotypes allowed us to address questions of nuclear differentiation and sex-biased dispersal and gene flow. Despite the smaller sample size in relation to the mtDNA dataset, we discovered significant between-ocean nuclear differentiation, indicating that there is restriction in bi-parental gene flow at the oceanic level. This is in contrast to previous research with more

limited sample sizes (Lyrholm et al. 1999), which suggests that larger sample sizes are required to uncover oceanic restrictions in nuclear gene flow in this species.

2.5.1. Why does structure vary by ocean? Geographic vs social structure as a factor in partitioning genetic diversity

It has been previously hypothesized that acoustic clans, which have similar repertoires of vocal codas, shape genetic differentiation (Rendell et al. 2012, Whitehead et al. 1998). In the Atlantic, coda patterns vary based on geography and acoustic clans are allopatric (Whitehead et al. 2012). This correlates with the heightened patterns of geographically-based mtDNA differentiation seen in this ocean. In the Pacific, acoustic clans are distributed sympatrically across broad geographic ranges (Rendell et al. 2012, Whitehead et al. 1998). It has been previously proposed that the lack of geographically-based mtDNA differentiation in the Pacific is because maternal dispersal and gene flow occurs within acoustic clans, but across broad geographic scales i.e. females are socially philopatric rather than geographically philopatric (Rendell et al. 2012, Whitehead et al. 1998). An alternative explanation is that acoustic clans in the Pacific might not have always been sympatric. The Pacific was subjected to a large degree of both legal and illegal whaling (Ivashchenko et al. 2013, Whitehead et al. 2012). Expansions following whaling could be expected to bring formerly allopatric acoustic clans into sympatry (Whitehead et al. 2012), and such an expansion would also be expected to tightly correlate both the maternally-inherited mtDNA and presumably maternally-inherited coda type. This hypothesis could be further tested using linked acoustic and genetic sampling in the Indian Ocean, as previously conducted in the Pacific (Rendell et al.

2012). We would expect that in the less intensively whaled Indian Ocean (due to the formation of the Indian Ocean sanctuary in 1979) that mtDNA genetic variation would be at equilibrium with both acoustic and geographic structure as it is in the Atlantic. Additionally, maternally isolated populations such as Sri Lanka might be expected to have their own coda repertoires, with these repertoires potentially lacking in diversity, similar to the situation seen in the isolated Mediterranean Sea (Pavan et al. 2000, Whitehead et al. 2012).

2.5.2. Female philopatry and male-biased dispersal

The lower levels of both mtDNA and microsatellite differentiation in males compared with females, supports the previous finding of male-biased dispersal in the sperm whale, both in the Atlantic (Engelhaupt et al. 2009) and globally (Lyrholm et al. 1999). It is likely that our estimates of male-biased dispersal are actually an underestimate, due to the inclusion of juvenile males in the dataset that are yet to disperse from their natal unit as well as small overall male sample sizes. The presence of male-biased dispersal helps explain why, in this study, we found that all Atlantic regions were significantly differentiated from each other based on mtDNA, when previous research found significant differentiation only in comparisons with the isolated marginal seas of the Mediterranean and Gulf of Mexico (Engelhaupt et al. 2009). In our study, the 'main basin' of the Atlantic was represented by two subtropical regions: the Canary Islands and western North Atlantic. In Engelhaupt et al. (2009), the 'main basin' was represented by the western North Atlantic and the cold temperate North Sea. The North Sea was represented by mature and maturing large males that likely originated from a mix of

lower latitude regions. This could reduce the likelihood of detecting significant patterns of differentiation between this population and others.

Given that dispersal of individuals does not necessarily equate to gene flow (Palsbøll et al. 2007, Whitlock and McCauley 1999), we also explicitly examined sex-biased gene flow. We demonstrate that the sperm whale shows male-biased gene flow (using the methods of Hedrick et al. 2013). Male-biased gene flow is also supported by between-region geographic comparisons of genetic diversity. While regions within the Atlantic, Indian and Pacific Oceans showed significant differences in mtDNA diversity (Appendix I, Figure I.5), no significant differences in microsatellite diversity were detected. Ongoing male-biased gene flow would homogenize microsatellite diversity between regions, while at the same time female philopatry would maintain differences in mtDNA diversity (Lyrholm et al. 1999). Despite the overall patterns of male-biased dispersal and gene flow, the finding of significant microsatellite differentiation between oceans indicates some restriction in oceanic dispersal and gene flow, even of males.

In addition to sex-biased gene flow and dispersal, we also found regions that appeared to be of differing importance to the sexes, particularly the male-dominated northern Mediterranean, Chagos Archipelago, and the Galapagos. The northern Mediterranean is known as an area that is male-dominated (Drouot et al. 2004b), and male aggregations have been noted to occur at the Galapagos (Christal and Whitehead 1997). Although mixed-sex groups were found around the Maldives, the samples from groups around the Chagos Archipelago were male. Given the occurrence of male aggregations of sperm

whales around the Galapagos, despite its equatorial status, it is not entirely unexpected that some regions in the Indian Ocean would show the same pattern.

2.5.3. Forces influencing diversity and differentiation within oceans

Based on the sex-biased tests of dispersal and differentiation, as well as the strong maternal structure observed at oceanic and regional levels, maternal fidelity appears to be the strongest force influencing global genetic differentiation in the sperm whale. These patterns were particularly strong in comparisons involving regions that could have enhanced philopatry due to particular geographic or bathymetric features, including the Mediterranean, Gulf of Mexico, and Sri Lanka. The mtDNA differentiation of the Gulf of Mexico and Mediterranean in comparison to other regions has previously been attributed to geographic isolation of these areas (Engelhaupt et al. 2009). Geographic isolation also appears to be associated with lower mtDNA diversity, potentially due to founder effects. In the extreme case of the Mediterranean, only one mtDNA CR haplotype was present. In contrast, geographic isolation cannot explain the large degree of mtDNA differentiation seen between Sri Lanka and other regions. Instead a potential explanation lies in Sri Lanka's large number of submarine canyons that lead to enhanced productivity of this region (de Vos et al. 2012). Female sperm whales utilizing the Sri Lankan canyons might not need to range as widely to satisfy nutritional requirements (Gordon 1987, Moors-Murphy 2014). This 'enhanced philopatry' could then lead to the striking mtDNA differentiation observed.

Beyond these regions, a clear pattern within oceans appears to be of female-philopatry and male-biased dispersal, which leads to microsatellite differentiation being less pronounced than mtDNA differentiation. However, two lines of evidence suggest the potential for breeding fidelity of males to some regions. First, significantly elevated levels of homozygosity (F_{IS}) were detected at the oceanic level, and for two within-ocean regions: Sri Lanka and Kiribati (Table 2.1). These elevated F_{IS} patterns indicate non-random mating, potentially due to male fidelity to specific regions. Secondly, 'islands' of nuclear differentiation identified in this study (the Mediterranean, Sri Lanka, Papua New Guinea, and Kiribati) indicate a restriction in bi-parental gene flow, also potentially due to breeding fidelity. Evidence for male fidelity has been found in the Californian Current by Mesnick et al. (2011) using genetic assignment, and in the Caribbean by Gero et al. (2013b) using photo-identification. Some natal return is also suggested by our finding of a possible first order kin relationship between two males in the Chagos Archipelago. Together, these findings could indicate that males pursue different strategies in regards to breeding, with some maintaining return migrations to specific areas while others disperse.

2.5.4. Differentiation, but not divergence: recent global origin of mtDNA diversity?

Despite common shared mtDNA haplotypes across oceans, significant regional and oceanic mtDNA structure was detected through fixation indices, and all oceans showed private mtDNA haplotypes (Figure 2.2). This indicates maternal isolation at the oceanic level. Given the lack of isolation by distance in the dataset, this maternal isolation seems to be driven by local drift. Although Φ_{ST} measures were generally larger than F_{ST} measures for all geographic comparisons, these differences were not significant. This

lack of mtDNA phylogeographic structure (i.e. divergence) between oceans, given the large levels of differentiation in haplotype frequencies within and between oceans, is surprising. One possible explanation would be a recent global origin of sperm whale mtDNA diversity followed by a rapid expansion and assortment of mtDNA haplotypes by female social group fidelity to specific regions. This 'maternally-mediated big bang' would also be consistent with the low overall mtDNA CR diversity found ($\pi = 0.442\%$).

A recent global origin of mtDNA diversity of 103,000 years has been previously suggested by phylogenetic analyses of mtDNA CR and mitogenomes, with the cause attributed to a population bottleneck, expansion or cultural selective sweep (Alexander et al. 2013, Lyrholm et al. 1996, Whitehead 1998). In a remarkable parallel between the maternal lineages of predator and (one of the sperm whale's) prey, the giant squid (*Architeuthis* spp.) also shows extremely low mitogenomic diversity with an origin of global mtDNA diversity of roughly 32,000-730,000 years ago. This raises the possibility that a global expansion of sperm whales could have been predicated on a recent expansion of their prey. However, giant squid are only one of many cephalopod and fish species preyed on by the sperm whale (Whitehead 2003), so this interesting propinquity remains in need of further investigation. In any event, since the expansion or event that reduced the diversity, insufficient time has elapsed for unique ocean-specific or region-specific haplotype lineages to be established through mutation and lineage sorting. However, the marked female philopatry present in the sperm whale has worked on post-expansion mtDNA diversity to establish strong patterns of mtDNA differentiation within oceans.

2.5.5. Management implications

The evidence for sex-biased dispersal and strong maternal population structure in the sperm whale argues for management units based on the more philopatric females, rather than the wider ranging males. Given this, the size of a local population should be based on the effective number of females, requiring female-specific estimates similar to male-specific estimates of humpbacks (*Megaptera novaeangliae)* (Constantine et al. 2012). It is also very important to aggregate samples at the appropriate spatial scale when attempting to detect geographically-based population structure in this species (Donovan 1991, Dufault et al. 1999). Given the clustered sample collection of the *Odyssey*, we arbitrarily chose to group samples that had occurred within 500 km of another sample. This could have inadvertently either split regions that were truly one population, or alternately 'lumped' areas with more than one distinct population. Both of these alternatives present problems. 'Splitting' regional populations could mean that the strong differences between social groups detected by previous studies (Lyrholm and Gyllensten 1998, Rendell et al. 2012) are conflated with regional differentiation. The alternative of 'over-grouping' can also be problematic. An example of this is the Maldives/Chagos Archipelago region, where only males were sampled around the Chagos Archipelago, but both sexes around the Maldives. The two areas were combined for the regional analysis given their spatial proximity, but could represent areas of different importance for males and females. To compensate against errors in geographically partitioning our dataset *a priori*, we also examined the presence of non-*a priori*-based population structure. Although mtDNA results were consistent with our *a priori* classifications, no clustering

of populations based on microsatellites was detected, likely due to the weak nuclear structure observed in this study (Waples and Gaggiotti 2006).

Another important management concern is the identification of isolated populations, as they are more at risk of local anthropogenic threats causing declines. The isolated Atlantic regions of the Gulf of Mexico and the Mediterranean are subject to multiple anthropogenic threats. The Mediterranean is currently listed as an endangered subpopulation by the IUCN in part due to risks including entanglement in marine debris and ship strike (Notarbartolo di Sciara et al. 2012). In addition to these risks, *Odyssey* toxicology results have indicated high levels of both mercury and titanium in the Mediterranean population (Savery et al. 2013a, Wise Sr. et al. 2011). In the Gulf of Mexico, sperm whales have been subjected to pollution associated with oil extraction (e.g. 2010 *Deepwater Horizon* oil spill), as well as potential negative interactions with anthropogenic sound (Mate et al. 1994). We also note here high levels of chromium and silver in the isolated regions of Kiribati and Sri Lanka, respectively (Savery et al. 2013b, Wise Sr et al. 2009).

2.5.6. Conclusion

In this study, we addressed issues that could have previously hindered the detection of global genetic structure in the sperm whale such as limited sample numbers and geographic scale. By doing so, we find for the first time, significant (albeit low) nuclear differentiation between ocean basins, as well as confirming significant mtDNA differentiation between oceans and low overall mtDNA diversity in the sperm whale. In

spite of the low mtDNA diversity, we also demonstrate a high level of within-ocean maternal structure, particularly involving the regions of Sri Lanka, Kiribati, the Gulf of Mexico and the Mediterranean. Identifying such patterns of genetic differentiation and diversity is important as a restriction in movement between local populations could indicate that there is a real risk of long-term declines in response to current anthropogenic threats such as ship strike, marine debris and pollution - despite the sperm whale's large global population. The specific mechanism(s) driving the differing levels of structure by ocean and region require further study: particularly how acoustic codas are structured in the Indian Ocean given the large level of maternal structure based on mtDNA in this region. Overall, the high levels of maternal structure observed in the sperm whale appear to be driven by geography and bathymetry coupled with female philopatry, given the presence of male-biased dispersal and gene flow. Interestingly, despite the apparent limited movement of females, phylogeographic structure at the oceanic or regional scales was not evident. This differentiation, without divergence, suggests a recent 'maternally-mediated big bang' origin of mtDNA diversity in the sperm whale, further supported by the low overall mtDNA diversity found.

2.6. Acknowledgements

Genetic samples from the 'Voyage of the *Odyssey*' were collected under permit #0751-1614 from the US National Marine Fisheries Service. Principle Investigator: Iain Kerr. The authors thank all who served as staff and crew during the Voyage of the *Odyssey*, as well as J. Wise Sr. and C. LaCerte for curation of DNA from samples collected by the *Odyssey*. We thank R. Constantine and K. Thompson for curation of the CeTA database,

and New Zealand Department of Conservation staff for collecting NZ samples used in this study; J. Rice for curation of the OMMSN database that supplied the Oregon samples; Oregon State University Cetacean Conservation and Genomics Laboratory for additional lab support; C. Sislak for assistance with DNA extraction, S. Pierszalowski and R. Hamner for analysis recommendations and assistance, and M. Smith for assistance summarizing *Odyssey* toxicology results. A. Liston and K. O'Malley provided valuable comments on this manuscript. This work was supported by a Mamie Markham Award and a Lylian Brucefield Reynolds Award from the Hatfield Marine Science Center; and a 2008–2011 International Fulbright Science & Technology award to A.A.; and co-funded by the ASSURE program of the Department of Defense in partnership with the National Science Foundation REU Site program to K.H. and C.S.B. [grant number NSF OCE-1004947].

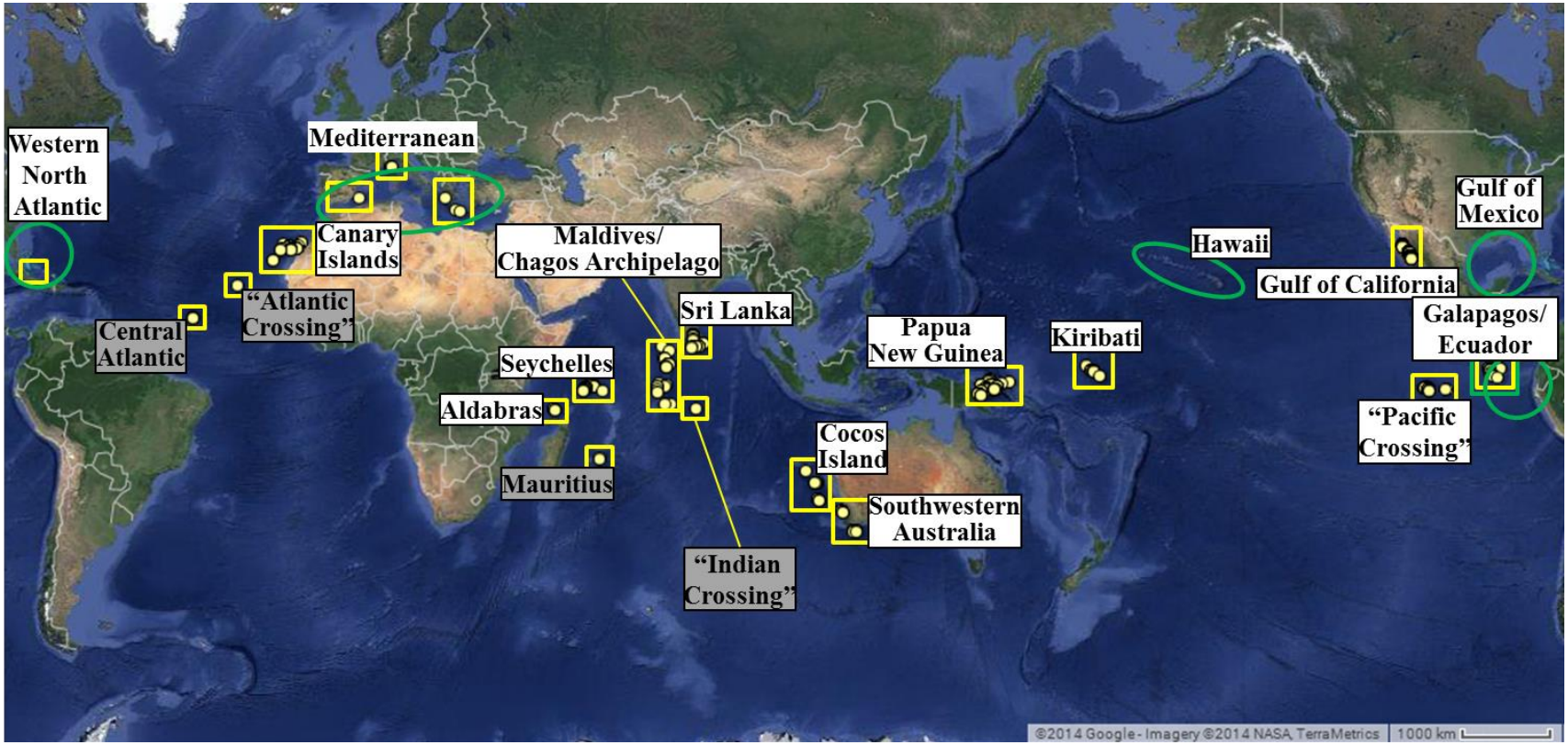


Figure 2.1: White labels indicate areas included in within-ocean mtDNA and microsatellite tropical/subtropical regional analyses. *Odyssey* samples were aggregated together if they occurred within 500 km of another sample. This created the localized regional areas shown in the yellow rectangles. Additional mtDNA samples/regions included in analyses originating from previous studies collected over similar spatial scales are circled by green (references in Table 2.1). *Odyssey* regions not included in regional analyses due to small sample sizes shown with gray labels.

Figure 2.2: Maximum parsimony network based on 394 bp of the mtDNA CR (haplotype definitions in Appendix I, Table I.1). Haplotypes are colored by ocean, with the exception of haplotypes V and W that were not localized to a specific ocean in previous studies. Size of haplotype pie is proportional on a log scale to the total number of samples with the haplotype. Lines represent substitutions (one or two between haplotypes as defined by the key). New haplotypes characterized in this study are outlined in red.

Figure 2.2 (Continued)

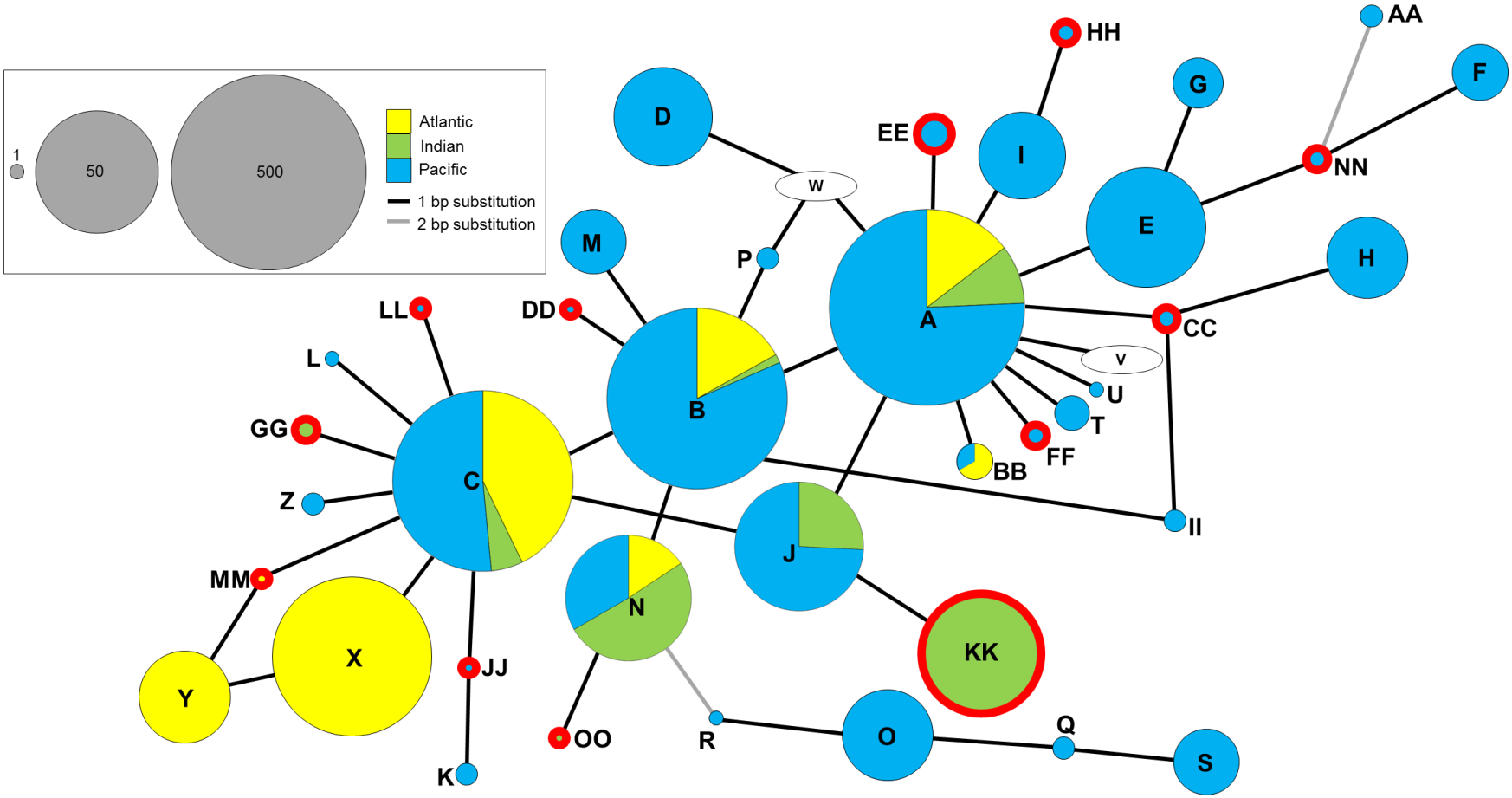


Figure 2.3: mtDNA and microsatellite differentiation values at oceanic and regional levels, including pairwise comparisons between oceans indicated by the arrows. See Table 2.1 for sample sizes used in regional mtDNA and microsatellite analyses. Two analyses were conducted for each set of markers at the worldwide scale: dividing by ocean, and dividing by regions (ignoring ocean as a factor). mtDNA haplotype frequencies by region are given within map, and for oceans to the right of the map. The key to haplotypes is given below the figure (with haplotypes ordered by global frequency). * significant at $p < 0.05$; ** significant at $p < 0.001$.

Figure 2.3 (Continued)

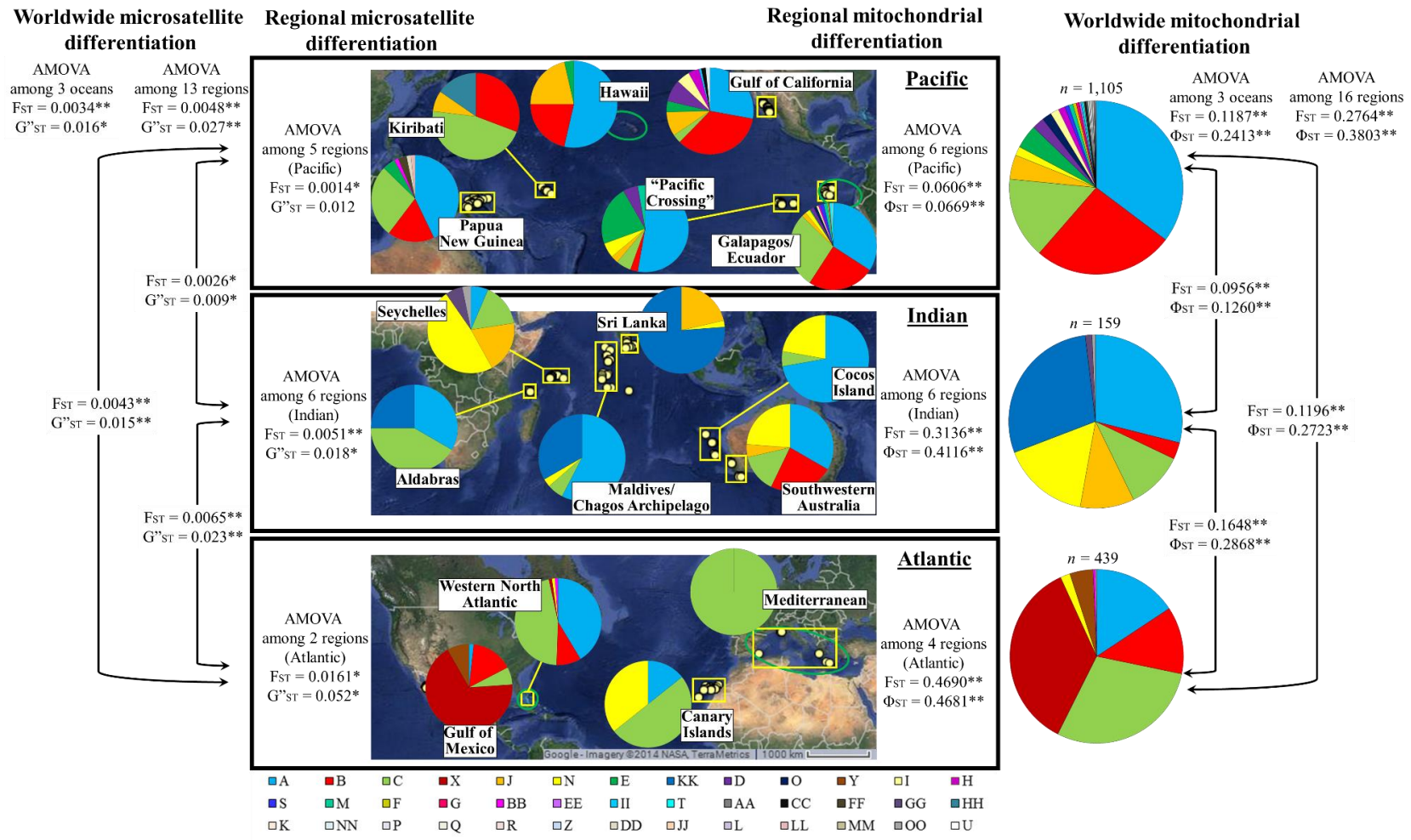


Table 2.1: Regional, oceanic and total global sample sizes (*n)* and diversity metrics for mtDNA and microsatellites. For mtDNA, number of haplotypes (*k*), haplotype diversity (*h*) and nucleotide diversity (in %, π) is presented, with standard deviations calculated in *Arlequin*. For microsatellites, numbers of individuals are given by sex (*F, M*) and total sample size (*n).* A binomial exact test was used to identify areas with a significant bias of females (asterisk after female sample size) or males (asterisk after male sample size), where * significant at $p < 0.05$; ** significant at $p < 0.001$. F_{IS} values are indicated as significant where * significant at $p < 0.05$. Regional allelic richness levels are adjusted by minimum regional sample size, with oceanic allelic richness levels adjusted by minimum ocean sample size. Regions ordered from east to west. 'Unassigned' includes samples not originating from tropical/subtropical regions, from areas with samples sizes too small to include in regional analyses, or those samples without a specific ~500 km regional location. References for mtDNA: [1] This study; [2] Lyrholm and Gyllensten, (1998); [3] Rendell et al. (2012); [4] Richard et al. (1996a); [5] Whitehead et al. (1998); [6] Mesnick et al. (2011); [7] Engelhaupt et al. (2009); [8] Ortega-Ortiz et al. (2012)

Table 2.1 (Continued)

		mtDNA CR					***Microsatellites***					
	Geographic area	*n*	*k*	*h (s.d.)*	π (*s.d.*)	*References*	*F*	*M*	*n*	*Allelic richness*	H_O	F_{IS}
Pacific	*Gulf of California*	122	11	0.788 (*0.024*)	0.368 (*0.250*)	1	96**	20	122	5.2	0.702	0.016
	Galapagos/Ecuador	284	16	0.743 (*0.012*)	0.355 (*0.242*)	1, 2, 3, 4, 5	0	23**	23	5.1	0.677	0.031
	Pacific Crossing	36	8	0.679 (*0.071*)	0.301 (*0.220*)	1	20	14	37	5.1	0.704	0.013
	Hawai'i	28	4	0.643 (*0.068*)	0.197 (*0.165*)	6	--	--	--	--	--	--
	Kiribati	13	4	0.718 (*0.089*)	0.381 (*0.276*)	1	10*	2	13	5.3	0.684	0.092*
	Papua New Guinea	63	8	0.720 (*0.036*)	0.299 (*0.216*)	1	54**	8	65	5.1	0.687	0.031
	Unassigned Pacific	469	--	--	--	1, 2, 3, 5, 6	--	--	--	--	--	--
	Total	***1015***	***33***	***0.779 (0.008)***	***0.388 (0.258)***	--	***180****	***67***	***260***	***8.9***	***0.694***	***0.027****
Indian	*Southwestern Australia*	21	5	0.791 (*0.044*)	0.307 (*0.227*)	1	9	4	21	5.1	0.697	0.007
	Cocos Island	18	3	0.451 (*0.117*)	0.230 (*0.187*)	1	18**	0	18	5.2	0.712	0.001
	Sri Lanka	42	3	0.382 (*0.076*)	0.132 (*0.126*)	1	42**	6	56	5.1	0.671	0.040*
	Maldives/ Chagos Archipelago	33	4	0.570 (*0.061*)	0.303 (*0.222*)	1	9	15	34	5.3	0.700	0.041
	Seychelles	31	6	0.716 (*0.066*)	0.413 (*0.279*)	1	17**	2	31	5.3	0.697	0.020
	Aldabras	12	3	0.712 (*0.069*)	0.364 (*0.268*)	1	6	2	12	5.1	0.677	0.027
	Unassigned Indian	2	--	--	--	1	1	1	3	--	--	
	Total	***159***	***8***	***0.788 (0.015)***	***0.431 (0.280)***	--	***102****	***30***	***175***	***8.8***	***0.686***	***0.035****
Atlantic	*Mediterranean*	40	1	0.000 (*0.000*)	0.000 (*0.000*)	1, 7	1	8*	9	4.6	0.631	0.086
	Canary Islands	14	3	0.648 (*0.081*)	0.333 (*0.247*)	1	15	9	25	5.2	0.690	0.014
	Western North Atlantic	87	6	0.616 (*0.028*)	0.273 (*0.201*)	1, 7	--	--	--	--	--	--
	Gulf of Mexico	230	5	0.506 (*0.035*)	0.216 (*0.170*)	7, 8	--	--	--	--	--	--
	Unassigned Atlantic	68	--	--	--	1, 2	3	0	7	--	--	--
	Total	***439***	***8***	***0.746 (0.010)***	***0.336 (0.232)***	--	***19***	***17***	***41***	***8.8***	***0.669***	***0.051****
	Unassigned Global	41	--	--	--	2	--	--	--	--	--	--
	Global total	**1654**	**39**	**0.826 (0.005)**	**0.442 (0.284)**	--	**301****	**114**	**476**	**13.1**	**0.690**	**0.029**

Table 2.2: Summary of circum-equatorial regional differentiation of mtDNA (F_{ST} and Φ_{ST}) and microsatellites (F_{ST} and G''_{ST}) for sperm whales. For each region, sample size and the number of significant pairwise comparisons to other regions for mtDNA CR (F_{ST} and Φ_{ST}) and microsatellites (exact test) is given. Average fixation index values over all pairwise comparisons are also presented. Regions with the largest number of significant pairwise comparisons/largest average fixation index values are shaded in darkest gray through to white for the regions with fewer numbers of significant comparisons/smaller average fixation index values. Average fixation index values and number of significant pairwise comparisons were calculated from regional pairwise comparisons for mtDNA and microsatellites (listed in Appendix I, Table I.4).

		mtDNA CR					**Microsatellites**			
	Geographic area	*n*	Total no. of sig. pairwise F_{ST} comparisons	Total no. of sig. pairwise Φ_{ST} comparisons	Average pairwise F_{ST}	Average pairwise Φ_{ST}	*n*	Total no. of sig. pairwise comparisons	Average pairwise F_{ST}	Average pairwise G''_{ST}
Pacific	Gulf of California	122	14/15	13/15	0.1811	0.2023	122	5/12	0.0055	0.0186
	Galapagos	284	12/15	12/15	0.1402	0.1598	23	4/12	0.0069	0.0267
	'Pacific Crossing'	36	14/15	13/15	0.2091	0.2899	37	3/12	0.0034	0.0113
	Hawai'i	28	15/15	11/15	0.2225	0.2634	--	--	--	--
	Kiribati	13	15/15	11/15	0.2492	0.2361	13	5/12	0.0088	0.0322
	Papua New Guinea	63	11/15	11/15	0.1515	0.1859	65	12/12	0.0074	0.0251
Indian	Southwestern Australia	21	12/15	10/15	0.1610	0.2167	21	6/12	0.0124	0.0406
	Cocos Island	18	14/15	11/15	0.2780	0.2742	18	4/12	0.0088	0.0280
	Sri Lanka	42	15/15	15/15	0.4774	0.6142	56	10/12	0.0077	0.0260
	Maldives/Chagos Archipelago	33	14/15	14/15	0.2485	0.2917	34	3/12	0.0052	0.0198
	Seychelles	31	14/15	13/15	0.2714	0.2690	31	2/12	0.0029	0.0095
	Aldabras	12	10/15	12/15	0.1867	0.2352	12	2/12	0.0098	0.0346
Atlantic	Mediterranean	40	15/15	15/15	0.5949	0.5946	9	12/12	0.0203	0.0687
	Canary Islands	14	13/15	12/15	0.2413	0.2753	25	4/12	0.0069	0.0230
	Western North Atlantic	87	13/15	12/15	0.1850	0.1936	--	--	--	--
	Gulf of Mexico	230	15/15	15/15	0.4280	0.5661	--	--	--	--

Table 2.3: Summary of locus-specific characteristics of microsatellite genotypes for the 476 *Odyssey* individuals genotyped in this study. *n* gives the number of individuals successfully typed at each locus. *Ho* and *He* (observed and expected heterozygosity, respectively) calculated in *Cervus*. F_{ST} calculated in *Genepop.* Statistically significant F_{ST} values are bolded and italicized, with * significant at $p < 0.05$; ** significant at $p < 0.001$.

Locus	*n*	Reference	Size range (bp)	No of alleles	*Ho*	*He*	Oceanic F_{ST}	Regional F_{ST}
EV1	462	Valsecchi and Amos (1996)	118 – 142	12	0.606	0.648	0.0076	0.0061
EV5	466	Valsecchi and Amos (1996)	148 – 174	11	0.697	0.704	***0.0078****	0.0071
EV14	471	Valsecchi and Amos (1996)	121 - 155	16	0.822	0.800	0.0026	***0.0120***
EV37	460	Valsecchi and Amos (1996)	177 - 250	9	0.485	0.483	***0.0032***	***0.0050****
EV94	423	Valsecchi and Amos (1996)	193 - 225	4	0.598	0.606	0.0025	0.0025
GATA417	397	Palsbøll et al. (1997)	172 - 202	13	0.720	0.763	***0.0129****	0.0019
GT23	340	Bérubé et al. (2000)	75 - 99	3	0.524	0.533	0.0024	0.0000
GT575	377	Bérubé et al. (2000)	131 - 137	5	0.485	0.527	0.0000	0.0104
rw4-10	458	Waldick et al. (1999)	177 - 213	14	0.832	0.834	***0.0018****	0.0037
SW13	446	Richard et al. (1996b)	134 - 176	29	0.852	0.907	0.0000	***0.0092****
464/465	428	Schlötterer et al. (1991)	141 - 145	14	0.685	0.718	0.0053	0.0000
SW19	458	Richard et al. (1996b)	89 - 167	15	0.786	0.835	***0.0028****	***0.0029****
FCB1	450	Buchanan et al. (1996)	107 - 145	31	0.873	0.920	0.0014	***0.0032****
Average	433.5			13.54	0.690	0.714	***0.0034****	***0.0048****

Table 2.4: Results for the sex-specific F_{ST} comparisons by marker and estimates of sex-biased gene flow (*Nm*, m_M/m_F, Hedrick et al. 2013). Regional analyses of sex-specific F_{ST} were limited to areas with more than two identified females and males: the Canary Islands, Aldabras, southwestern Australia, Maldives/Chagos Archipelago, Seychelles, Sri Lanka, Gulf of California, Kiribati, Papua New Guinea, and the Pacific Crossing. Note, that within the Pacific although male-specific microsatellite F_{ST} appears to exceed that of females, the male-specific estimate is not significantly different from zero. Due to limited sample sizes, a within-ocean regional F_{ST} analysis was not conducted for the Atlantic. Hedrick et al.'s (2013) estimates of sex-specific gene flow are based on the fixation indices presented in Figure 2.3. As all variance in assignment tests (vAIc) were not significant, results of these tests are not displayed. Statistically significant values (for sex specific F_{ST} values and for the p-values for the difference in F_{ST}) are bolded and italicized, with * significant at $p < 0.05$; ** significant at $p < 0.001$.

Area		Sex	mtDNA CR			Microsatellites			Gene flow	
			n	F_{ST}	*p* -value	n	F_{ST}	*p* -value	*Nm*	m_M/m_F
Pacific	By region	F	178	***0.1174*****	0.0596	180	***0.0003****	0.8257	7.75	22.01
		M	43	***0.0346*****		44	0.0091		170.57	
Indian	By region	F	70	***0.4892*****	0.1628	83	***0.0061****	0.4434	1.09	43.56
		M	27	***0.2878****		29	0.0050		47.68	
Atlantic	By region	F	--	--	--	--	--	--	0.57	25.99
		M							14.71	
Worldwide	By region	F	257	***0.2757*****	***0.0103****	278	***0.0066*****	0.1916	1.31	38.59
		M	74	***0.1255*****		82	0.0019		50.52	
	By ocean	F	277	***0.1280*****	0.0792	301	***0.0071*****	***0.0211****	3.71	18.74
		M	100	***0.0692*****		114	0.0004		69.57	

3. Lenient matrilineality in groups of sperm whales from equatorial waters of the Pacific, Indian, and Atlantic Oceans

3.1. Abstract

The social structure of the sperm whale has been a focus of scientific interest since the era of "Yankee" whaling. Current understanding of sperm whale social structure suggests the presence of mixed-sex groups ('female social groups', consisting of females and their juvenile offspring of both sexes), bachelor groups (composed of adolescent males that have dispersed from their mixed-sex natal units), and solitary adult bull males. Whaling and behavioral data have provided important insights into the relative size of these groups, and the duration of long-term social bonds within mixed-sex groups. However, several aspects of the molecular ecology of social groups remain undescribed, particularly in the Indian Ocean where sperm whale social groups have not been characterized. In this study, we use microsatellites and mitochondrial DNA (mtDNA) from 106 sperm whale social groups (n = 678 individuals) representing 81 mixed-sex groups, 22 bachelor groups, and 3 groups of unknown composition, in order to investigate differences in how social groups partition genetic diversity by ocean, the mechanisms by which they partition genetic diversity, and differences between bachelor and mixed-sex groups. We find over a worldwide scale, social groups partition more mtDNA (F_{ST} 0.434) and microsatellite (F_{ST} 0.015) diversity than geographically-based partitioning schemes (Chapter 2). Although this affirms their importance in partitioning genetic diversity, the percentage of mtDNA variance explained by social group varied by ocean in hierarchical analyses nesting social group within geographic regions. Strong

effects of social groups were seen in the Pacific, and strong effects of geographic regions in the Indian Ocean. By developing a new index for assessing the matrilineality of social groups within geographic regions (the 'standardized matrilineality index'), we show that the mechanism for the partitioning of mtDNA diversity by social groups appears to be 'lenient matrilineality'. That is, although the sperm whale does not show strict matrilineality, most groups include fewer matrilines than expected by chance given regional mtDNA haplotype frequencies. Genetic and ecological differences also occur by group type, with bachelor groups significantly smaller and containing more mtDNA haplotypes per group size than mixed-sex groups. Finally, although evidence of lenient matrilineality of social groups was detected, social groups appear to contain too many matrilines to be the unit through which global sperm whale mtDNA diversity is reduced by the previously proposed cultural hitchhiking and environmental heterogeneity lineage extinction hypotheses.

3.2. Introduction

The social structure of sperm whales has been the subject of interest since the days of 18-19th century "Yankee" whaling (Beale 1839). Whalers and earlier scientists assumed that sperm whale social structure was characterized by warmer-water 'harems' consisting of females and a bull sperm whale 'master', and cooler-water 'bachelor' groups of adolescent males not yet mature enough to obtain a harem (Beale 1839, Best 1979). Later, it was recognized that the male bull sperm whales are only ephemerally associated with female-dominated social groups, spending the remainder of their time in polar waters. Social bonds between females appear much more important for maintaining the

cohesiveness of these ‘mixed-sex’ groups (Christal et al. 1998, Coakes and Whitehead 2004, Dufault and Whitehead 1998, Gero et al. 2013b, Richard et al. 1996a, Whitehead et al. 2012, Whitehead and Arnbom 1987). In fact, the presence of stable mixed-sex groups and ‘roving males’ led some researchers to speculate that the sperm whale and elephant showed convergent social evolution with dispersal of males and strictly matrilineal mixed-sex groups (Best 1979, Letteval et al. 2002, Weilgart et al. 1996, Whitehead 1998).

Matrilineal social structure refers to patterns of kinship structured along maternal lines. While assorting into social groups could offer protection and defense against predators, assorting specifically with female kin is also thought to offer benefits including increased reproductive performance and reduced competition with conspecifics (Silk 2007). In African savannah elephants (*Loxodonta africana*), matrilineality is seen within long-term, stable, female social groups (Archie et al. 2006). Females remain with their natal group when mature, leading to stable core groups of 1-20 adult females related through a matriarch (Archie et al. 2006). In contrast, males disperse, as seen in sperm whales. These patterns of matrilineality are associated with significant genetic differentiation between groups (Archie et al. 2008). The African forest elephant (*L. cyclotis*) and Asian elephant (*Elephas maximus*) also show matrilineal group structure (Fernando and Lande 2000, Schuttler et al. 2014). The duration of group bonds is less well studied in the forest elephant than the savannah elephant; but the Asian elephant appears to show less cohesive social structure than the savannah elephant (de Silva and Wittemyer 2012). However, one consistent finding across the geographic range of both elephant genera is

that deep phylogeographic divergences are evident based on mtDNA (Eggert et al. 2002, Vidya et al. 2009).

In addition to the obvious parallels with elephants, other mammalian species with strictly matrilineal social groups include the wild boar (*Sus scrofa*), primates such as the savannah baboon (*Papio cynocephalus*), gelada (*Theropithecus gelada*), howler monkeys (*Alouatta* spp.), macaques (*Macaca* spp.), grey mouse lemur (*Microcebus murinus*), and human (*Homo sapiens*); bats such as the northern myotis (*Myotis septentrionalis*) and Bechstein bat (*M. bechsteinii*); carnivores such as the spotted hyena (*Crocuta crocuta*); and cetaceans such as the long-finned pilot whale (*Globicephala melas*) and killer whale (*Orcinus orca*) (Amos et al. 1993, de Ruiter and Geffen 1998, Hage and Marck 2003, Hoelzel et al. 2007, Holekamp et al. 2012, Johnson et al. 2014, Kaminski et al. 2005, Kerth et al. 2000, Möller 2014, Patriquin et al. 2013, Silk 2007, 2009, Van Horn et al. 2004, 2007). Although the killer whale and long-finned pilot whale have a closer evolutionary relationship with the sperm whale than the other species listed here, both males and females appear to show fidelity to their natal group in these species (Amos et al. 1993, Hoelzel et al. 2007). This contrasts with the male dispersiveness observed in the sperm whale.

In the sperm whale, assumed matrilineality of mixed-sex groups led to hypotheses to explain the low mitochondrial DNA diversity (mtDNA) observed in the global sperm whale population (Lyrholm et al. 1996, Whitehead 1998). Whitehead (1998, 2005) focused on the possibility of mtDNA control region (CR) haplotypes “hitchhiking” with

cultural innovations (beneficial behaviors transmitted in matrilineal social groups). Through modeling, he showed that the linkage of cultural innovations within matrilineal groups with the maternal inheritance of mtDNA haplotypes could reduce mtDNA CR diversity through a selective sweep (Whitehead 1998; 2005). An alternative socially-mediated explanation could be that of lineage extinctions. Two hypotheses based on this have been proposed: (1) environmental heterogeneity coupled with matrilineal groups could reduce mtDNA CR diversity through the reduction of fitness of some social groups (Tiedemann and Milinkovitch 1999), and (2) mass strandings of matrilineal mixed-sex groups could cause lowered mtDNA CR diversity through random lineage extinctions (Alexander 2006, Siemann 1994). However, these hypotheses appear contingent on the level of matrilineality shown by social groups.

Direct observation of mixed-sex groups in the Galapagos revealed that they themselves are made up of temporary associations of smaller groupings termed ‘social units’ (Christal et al. 1998, Whitehead et al. 1991). The observations of long-term social group cohesiveness (Gero et al. 2013b, Gordon 1987, Whitehead et al. 1991) likely refer to the ‘social unit’, as units appear to associate with each other in ‘social groups’ over shorter time periods on the order of days to weeks (Whitehead et al. 1991). Further behavioral observation indicated that dispersal and merging of individuals could occur even at the ‘social unit’ level (Christal et al. 1998). These results - and investigations into the number of matrilines present using the maternally-inherited mtDNA - are inconsistent with either social units or groups as purely matrilineal social structures (Christal et al. 1998, Mesnick et al. 1999, Richard et al. 1996a). However, lack of strict matrilineality has been

suggested to be an artifact of limited geographic sampling or ongoing social disruption resulting from whaling (Whitehead 1999), with studies investigating matrilineality in social groups restricted to the eastern Pacific, Caribbean, and Sargasso Sea (Whitehead et al. 2012).

Despite the lack of strict matrilineality shown by mixed-sex groups, the social group appears to be an important factor influencing the genetic structure of the sperm whale based on analyses of mtDNA and microsatellites (Lyrholm and Gyllensten 1998, Lyrholm et al. 1999, Rendell et al. 2012). In fact, in previous studies, partitioning datasets by social group explained more mtDNA variance than partitioning by geography (Lyrholm and Gyllensten 1998, Rendell et al. 2012). However, to date, the investigation of genetic partitioning by social group has been restricted to the Pacific and Atlantic, with no studies on differentiation by social group conducted in the Indian Ocean. An additional unanswered question is *how* social groups partition genetic diversity. The presence of large significant mtDNA and microsatellite differentiation between groups suggests a degree of maternal isolation, as well as potentially elevated levels of relatedness, driving the significant microsatellite differences even if social groups do not appear to be strictly matrilineal (Mesnick 2001, Richard et al. 1996a).

In the above description of social groups, we have focused largely on the mixed-sex group. However, despite the fact that female sperm whales appear to be more social and gregarious than adult male sperm whales, adolescent males do form groups based on size, with larger individuals becoming more solitary until they join the ranks of the solitary

bulls (Best 1979). These bachelor groups are characterized from behavioral and whaling data by a lack of preferred companions or social organization, and a lack of individuals occurring in close proximity, in contrast to characteristics of mixed-sex groups (Letteval et al. 2002, Whitehead et al. 1992). Even though research has indicated that some males show fidelity to specific feeding areas over a period of years (Childerhouse et al. 1995, Whitehead et al. 1992), multiple males potentially cluster together only over a period of days (Letteval et al. 2002). However, previous research has focused on larger males distributed more coastally, rather than smaller males found in bachelor groups containing larger numbers of individuals further offshore (Letteval et al. 2002).

In this study, we investigate the partitioning of genetic diversity by social groups in different ocean basins, for the first time including the Indian Ocean. We also make opportunistic observations on distance and duration of social group bonds. To investigate the mechanism by which social groups partition genetic diversity, we examine the existence of first-order relatives and relatedness within social groups, and develop the 'standardized matrilineality index' (based on the proportion of dyads within social groups with the same haplotype, using the maternally-inherited mtDNA CR as a proxy for matrilines). We contrast these results between mixed-sex groups and bachelor groups. To achieve these objectives, we use social groups sampled during the 'Voyage of the *Odyssey*', a five-year expedition sponsored by the Ocean Alliance that sampled sperm whales in circum-equatorial regions (Godard et al. 2003), as well as social groups characterized in the literature.

3.3. Materials and Methods

3.3.1. Sample collection and genotyping

Skin biopsy samples were collected from sperm whales located in tropical/subtropical regions of the Pacific, Indian, and Atlantic Oceans, by the 'Voyage of the *Odyssey*' (Figure 3.1). DNA was extracted from these samples as detailed in Chapter 2. The *Odyssey* samples were then genetically sexed, genotyped at 13 microsatellite loci, and had a 394 bp fragment of their mtDNA CR amplified and sequenced, as detailed in Chapter 2. Published genetic information was also included for 39 social groups located in the Gulf of Mexico, eastern North Pacific and Ecuador (Figure 3.1, Table 3.1: Engelhaupt 2004, Mesnick 2001, Ortega-Ortiz et al. 2012, Richard et al. 1996a).

3.3.2. Definition of social groups

Field data on spatial and temporal proximity of *Odyssey* biopsy samples were used to identify samples collected during a single encounter with a social group, including when attempts were made to follow the same group overnight. In addition, due to the chance that previously sampled groups might also have been unintentionally re-encountered, we combined groups that had genetic replicates between them (replicate identification process detailed in Chapter 2). Groups were included in analyses if they had two or more individuals pass the measures of genetic quality control presented in Chapter 2. Recaptures of multiple individuals from the same social group, based on the genetic replicates identified in Chapter 2, were used to opportunistically look at association time of individuals and distance traveled between recapture events.

The groups identified in the *Odyssey* dataset likely correspond to a mix of 'social groups' and 'social units' as defined in previous publications (Christal et al. 1998, Whitehead 2003). To assess group size, acoustic and visual field estimates were used. However, to classify groups into 'mixed-sex' (female social groups) and 'bachelor' groups we relied on genetic sex identification (as described in Chapter 2). Genetic sex identification was used because although photo-identification is used for mark-recapture abundance estimates in sperm whales, it is difficult to visually assess the sex of sperm whales, especially immature animals. If female samples were present, groups were classified as 'mixed-sex'. If only male samples were characterized, groups were classified as 'bachelor groups'. This classification is likely to contain some ascertainment error, as females could have been present but un-sampled in "bachelor groups". If genetic sex information was not available for a group, it was classified as an 'unknown'. Mixed-sex groups (*Odyssey* and previously published data) were only included in sex-ratio comparisons where more than half of the group had sex information available.

3.3.3. Genetic differentiation among social groups

F_{ST} and Φ_{ST} (using the Tamura and Nei (1993) correction) were calculated for mtDNA using *Arlequin v. 3.5* (Excoffier et al. 2005), with 10,000 replicates to assess significance. F_{ST} was calculated for microsatellites using *Genepop v. 4.2* (Raymond and Rousset 1995a, Rousset 2008), assessing significance of the results with the exact test (Raymond and Rousset 1995a). G''_{ST}, an index which compensates for the diversity of microsatellites (Meirmans and Hedrick 2011), was calculated using *Genodive v. 2.ob25* (Meirmans and van Tienderen 2004), with 10,000 permutations to assess significance. F_{IS}

of social groups by region, based on microsatellites, was calculated using *FSTAT v 2.9.3.2* (Goudet 2001), using 10,000 replicates to assess significance. To conduct hierarchical analyses, social groups were nested within (a) ocean and (b) region. Regional analyses were conducted nesting social group within region at both a worldwide scale, and within each ocean. Hierarchical analyses were conducted using *Arlequin* for mtDNA (for F_{ST} and Φ_{ST}), and *GDA v 1.0* (Lewis and Zaykin 2001) for microsatellites (F_{ST} only). To limit the effect of small sample sizes, analyses of genetic differentiation were restricted to the subset of social groups defined in the section above that had five or more individuals with genetic data available.

3.3.4. Partitioning of microsatellite diversity by social group: identification of kin and average relatedness

Average relatedness (*R;* Queller and Goodnight 1989) between members of social groups where more than five individuals were sampled was calculated using *GROUPRELATE* (Valsecchi et al. 2002). *R* can range between -1.0 (no alleles shared between group members) through 0.0 (individuals are not more related than expected by chance given background allele frequencies) to +1.0 (group is comprised of clones). A group comprised of siblings will (on average) have an *R* value of 0.5 (Queller and Goodnight 1989). To assess whether members of a group were significantly more related than expected by chance, *GROUPRELATE* was also used to randomly permute individuals among groups within each geographic region (geographic region definitions follow Chapter 2, also see Figure 3.1). Potential first-order kin were identified in the *Odyssey*

dataset, as described in Chapter 2, by identifying pairs of individuals that shared at least one allele at every microsatellite locus using *SOLOMON v. 1.0-1* (Christie et al. 2013).

3.3.5. Partitioning of mtDNA diversity by social group: number of matrilines and the standardized matrilineality index

To assess matrilineality of social groups, the number of unique mtDNA CR haplotypes was recorded for each social group. As unrelated females can possess the same mtDNA CR haplotype, the number of unique mtDNA haplotypes in a social group represents an estimate of the minimum number of matrilines present. To control for the effects of uneven sampling among groups we characterized the number of haplotypes in a group divided by social group size (in this case, the total number of individuals within a group with mtDNA CR information). One haplotype/group size indicated every individual in that social group had a different haplotype, and 1/*n* haplotypes/group size indicated that every individual had the same haplotype, where *n* equaled the number of individuals sequenced for mtDNA in that social group. This measure was used instead of haplotype diversity as some previous publications recorded only the number of haplotypes in social groups, and not the frequency of those haplotypes.

To further quantify 'lenient matrilineality' by social group, we developed the 'standardized matrilineality index'. This method is similar to that previously used by Weinrich et al. (2006) for examining the mtDNA CR haplotypes of associated humpback whales (*Megaptera novaeangliae*). However, unlike Weinrich et al. (2006), who developed an index based on the actual identity of the haplotypes in question, our

matrilineality index only considered if haplotypes in a dyad were identical or not. To calculate the standardized matrilineality index, the proportion of within-social group dyads that share the same haplotype is first calculated (Table 3.2). This estimate is referred to as Hi_{obs} (the observed proportion of haplotype identity). Monte Carlo sampling is then used to generate a null distribution of the expected proportion of within-social group dyads that share the same haplotype, given the background regional haplotype frequencies. A custom script (Appendix II, Figure II.1) was used to generate this distribution in *R v. 3.0.2* (R Core Team 2013) with 1,000 replicates. The mean of this distribution is referred to as the expected mean proportion of haplotype identity (Hi_{exp}). The Monte Carlo sampling is repeated for each social group of differing size within a region (Table 3.2).

The null distribution can be used to calculate whether a social group is significantly more matrilineal than expected, given the regional haplotype frequencies, by:

$$\text{Significance} = \frac{\text{no of. permutations} \geq Hi_{obs}}{\text{Total number of permutations}}$$

Finally, in order to calculate the standardized matrilineality index itself, the following simple formula is used:

$$\text{Standardized matrilineality index} = Hi_{obs} - Hi_{exp}$$

The standardized matrilineality index will vary between -1 and 1, with positive values indicating groups are more matrilineal than expected given regional haplotype frequencies. Because the standardized matrilineality index controls for differing haplotype frequencies between regions, similar to the matrilineality index (*q*) proposed by Whitehead et al. (2012), it allows for comparisons across regions and oceans of the level of matrilineality of social groups. However, in the case of a region with one haplotype, the standardized matrilineality index cannot be calculated, as deviation from Hi_{exp} is not possible (i.e. there is no power to detect matrilineality). Likewise, background haplotype frequencies for a region need to be based on more than the social group being tested, as there will also be no power to detect a deviation from Hi_{exp}. The standardized matrilineality calculations can also be carried out at a regional level, by summing across the total observed number of within-social group dyads, and the number of same-haplotype within-social group dyads (Table 3.2). An example of the calculation of the social group and regional standardized matrilineality index is available in Table 3.2. In the current study, indices were conducted on both the full dataset, and restricting the sample to females (due to the possibility males could have originated from other social groups).

3.4. Results

3.4.1. Definition of social groups and geographic regions

Among the *Odyssey* dataset, 67 social groups (n = 420 individuals) had at least two individuals pass genetic quality control measures (see Chapter 2 for QC measures). These groups originated from the Gulf of California, Galapagos, ‘Pacific Crossing’, Kiribati,

Papua New Guinea, southwestern Australia, Cocos Island, Sri Lanka, Maldives/Chagos Archipelago, Aldabras, Seychelles, Mediterranean, the Canary Islands and the central Atlantic (Figure 3.1). After inclusion of published social group information, this led to 106 social groups (n = 678 individuals), representing 20 regions and all 3 oceans, included in downstream analyses of genetic differentiation, relatedness and matrilineality (Figure 3.1; Table 3.1). Of the 106 groups, 27 had 2 individuals sampled, 31 had 3 - 4 individuals sampled, and 48 groups ≥ 5 individuals sampled (Table 3.1). Using genetic sex information, 22 of the groups were characterized as bachelor groups, 81 as mixed-sex groups, and 3 as unknown (Table 3.1).

Replicates between *Odyssey* groups allowed for the identification of 'tandem movements' where multiple individuals of a social group were initially sampled at the same time, and were then later sampled together anywhere from 1 to 51 days later, and 5 to 508 km from the initial sampling event (Table 3.3). Of the 8 tandem movements, the majority (75%) of inferred movements occurred within 14 days and 200 km of where the social group was initially sampled (Table 3.3). This reflects the sampling strategy of the *Odyssey*, as it generally spent limited time in specific areas. Exceptions included a Sri Lankan social group (SRI002) recaptured over a longer period (38 days), but within a reasonably short (< 25 km) distance (Table 3.3: "h"). The other outlier in Table 3.3 was a Papua New Guinea social group (PNG010) that moved a considerable distance between resampling events (~507 km) but over an extended period of time (51 days). The longer period between recapture events reflects the additional time the *Odyssey* spent sampling in these locations. Group code, latitude, and longitude for the social groups genetically

characterized in this study are archived as described in Appendix I (along with DNA profiles for individuals within these groups).

3.4.2. Comparisons of group size and sex ratio

Based on field data, mixed-sex groups had a significantly larger average group size (16.9 individuals) in comparison with bachelor groups (7.0 individuals, t-test $p < 0.001$, Figure 3.2). There were no significant differences in average bachelor group size among the Pacific, Indian, and Atlantic Oceans. However, significant differences occurred between all oceans for mixed-sex groups, with the Atlantic showing a significantly smaller average group size (11.4 individuals) than the Indian Ocean (15.6) which in turn had a significantly smaller average group size than the Pacific Ocean (24.1, see Figure 3.2 for *p*-values and number of groups involved in comparisons). Bachelor groups were defined as groups with no females sampled (100% male). We therefore only examined the percentage of females for the 21 mixed-sex groups where more than half of the group had sex information available. Seven 100% female 'mixed-sex' groups were observed: two each in the Atlantic and Pacific, and three in the Indian Ocean (Table 3.1). The remaining 14 groups included males, although all groups were at least 50% female. Our results were consistent with sex ratio information for 'mixed-sex' groups described from whaling data by Best (1979). Sample sizes were too limited to carry out comparisons by ocean.

3.4.3. Genetic differentiation among social groups

As detailed in Table 3.1, 23 groups had ≥5 individuals with both microsatellite and mtDNA information. An additional 3 groups had only microsatellite information for ≥5

individuals, while 11 groups had only mtDNA information (groups from the literature as detailed in Table 3.1). The analyses of genetic differentiation were somewhat limited in inclusion of bachelor groups, as 95% of the bachelor groups had fewer than five individuals genotyped for microsatellites or mtDNA. Of the 26 groups with microsatellite information where F_{IS} was calculated (using background regional allele frequencies), only one group showed a significant excess of homozygotes (SRI002), and one group showed a significant excess of heterozygotes (CNI009) (Table 3.1). Both SRI002 and CNI009 were defined as mixed-sex groups.

Differentiation among social groups within regions was extremely high (Figure 3.3), with significant differentiation values as high as $F_{ST} = 0.703$ for mtDNA (Cocos Island, $n = 2$ groups) and $F_{ST} = 0.033$ for microsatellites (Maldives/Chagos Archipelago, $n = 3$ groups). Despite these high levels of differentiation, some regions showed no significant differentiation between social groups for mtDNA (Seychelles) or microsatellites (Gulf of California, Cocos Island). In non-nested analyses of social group differentiation at an oceanic and at a worldwide scale, social group consistently explained greater levels of variation than partitioning by geographic regions and oceans (Figure 3.3). However, given the more fine-scale partitioning of social groups compared to higher-level geographic scales, it is expected that social groups would explain more genetic variance. To account for this, we used hierarchical AMOVAs with two different partitioning schemes (nesting social group within regions and within oceans). No matter what hierarchical partitioning scheme was used, social group was the only level (compared with regions and oceans) that explained any significant variance in the microsatellite

dataset (Figure 3.3). For nested social group analyses of mtDNA at a worldwide scale, region and ocean explained a significant amount of variance, but less than that explained by social group (Figure 3.3). When restricting analyses to social groups nested within geographic regions, analyzing the Pacific Ocean and Indian Ocean separately, an interesting pattern was found. In the Pacific, only social group (compared with region) explained any significant amount of mtDNA variance (Figure 3.3). In contrast, within the Indian Ocean region was more important in explaining mtDNA variance than social group (Figure 3.3).

3.4.4. Partitioning of microsatellite diversity by social group: identification of kin and average relatedness

All 67 *Odyssey* groups were investigated for the presence of first-order kin, using microsatellite data. Average $p(_{ID})$ and $p(_{ID\text{-}sibs})$ for the 13 microsatellite markers were 3.10E-14 and 1.18E-05 respectively (as described in Chapter 2). Potential first-order kin were identified using *SOLOMON* to find pairs of individuals that shared at least one allele per microsatellite locus (as described in Chapter 2). For the 25 published Gulf of Mexico groups, information on kin relationships was taken from Engelhaupt (2004) and Ortega-Ortiz et al. (2012). Of the total 92 groups assessed, 14 groups had one or more pairs of potential first-order kin (Table 3.1). All but one of these occurred within mixed-sex groups, involving at least one female (one male-male pair was inferred within the Chagos Archipelago). Of the 13 mixed-sex groups with potential first-order kin, 10 groups had a single pair. One Gulf of Mexico group (G11) had two pairs of potential first-order kin. Two groups, a Gulf of Mexico group (G3), and a Gulf of California group

(GCA003) had potential first-order kin consistent with a grandmother-mother-daughter relationship. GCA003 also had an additional separate pair of potential first-order kin present (Table 3.1). With 17 pairs of potential first-order kin detected in mixed-sex groups, and just one in bachelor groups, mixed-sex groups appear significantly more likely to contain kin than bachelor groups (chi-squared test, $p = 0.0364$). In addition to the kin relationships identified above, three pairs of first-order kin were found across different mixed-sex groups: two within the Gulf of California (GCA002 dispersal to/from GCA024, GCA026 dispersal to/from GCA029) and one within Sri Lanka (SRI004 dispersal to/from SRI005).

In comparison with the groups investigated for kin, one social group from the *Odyssey* dataset (C_ATL006) was excluded from relatedness analyses, as it was the only sample from the central Atlantic (and therefore background microsatellite allele frequencies could not be calculated). This left 91 social groups (66 *Odyssey* groups, 25 Gulf of Mexico groups). Of these, only two groups were found to be more related than expected by chance given regional microsatellite allele frequencies: one located in Sri Lanka ($R = 0.456$, $p = 0.004$) and one located in the Gulf of California ($R = 0.226$, $p = 0.005$). Both of these groups were classified as mixed-sex. Over all 91 groups, average relatedness was low and not significantly different from zero (average $R = 0.017$, t-test $p = 0.918$). When stratified by group type, average relatedness did not differ significantly from zero for mixed-sex groups ($R = 0.026$, t-test $p = 0.9902$) or bachelor groups ($R = 0.002$, $p = 0.5335$, Figure 3.4). No significant difference in average social group relatedness was detected between group types (t-test $p = 0.5228$, Figure 3.4). No significant differences in

average social group relatedness were detected between oceans for the combined bachelor/mixed-sex group dataset (t-test, $p > 0.1115$) or by group type (bachelor-by-ocean $p > 0.2211$, female-social-by-ocean $p > 0.3161$).

3.4.5. Partitioning of mtDNA diversity by social group: number of matrilines and the matrilineality index

Of the 99 groups with more than one individual with mtDNA information, there were 38 groups (29 mixed-sex groups, 7 bachelor, 2 unknown) with only a single haplotype (38.4%), including 11 groups with >5 sampled individuals (all mixed-sex groups). The remainder showed up to eight mtDNA CR haplotypes (Table 3.1). Bachelor groups had a significantly higher number of haplotypes/group size ($\overline{x} = 0.690$) than mixed-sex groups ($\overline{x} = 0.430$, t-test $p < 0.0004$, Figure 3.5). For both bachelor and mixed-sex groups there were no significant differences in the number of haplotypes/group size among groups located in the Pacific, Indian, and Atlantic Oceans.

To account for differences in mtDNA diversity between regions, we calculated the standardized matrilineality index for social groups within regions, and then by region across all social groups. Due to the concern that mature males might have been associated with the mixed-sex groups (which could reduce apparent matrilineality), we conducted the matrilineality indices on the total dataset, and on a dataset restricted to only females found within mixed-sex groups. Given the limited overall number of males in mixed-sex groups, there was only a small difference between the total dataset and the dataset restricted to females (correlation coefficient = 0.820, $p < 0.001$, $n = 38$ groups).

Therefore, the results presented here are based on the total dataset, as this also allows an assessment of the matrilineality of bachelor groups (which are defined as not containing females).

For all social groups that had mtDNA haplotype frequencies based on more than one individual (n = 87), 69.0% had standardized matrilineality indices greater than zero i.e. were more matrilineal than expected by chance (Table 3.1). Of the 60 groups more matrilineal than expected by chance, 38 were the strictly matrilineal groups characterized above. Even though the remainder of these groups were not necessarily strictly matrilineal, their standardized matrilineality index values indicated they were more matrilineal than expected given the background haplotype frequencies of the regions they were located in. Based on the permutation test, 29.9% of groups (23 mixed-sex groups, 1 bachelor, 2 unknown) had indices that were significantly greater than expected by chance (Table 3.1). This included 20 of the groups characterized as strictly matrilineal. Based on the permutation test, 19.5% of groups (10 mixed-sex groups, 7 bachelor groups) were identified as having standard matrilineality indices that were significantly smaller than expected by chance i.e. they were less matrilineal than expected based on haplotype frequencies of regions (Table 3.1).

A comparison of mixed-sex groups and bachelor groups revealed that mixed-sex groups were qualitatively more matrilineal (Figure 3.6). However, this difference was not significant (t-test, p = 0.1631). A comparison of the proportion of mixed-sex groups and bachelor groups that were more matrilineal than expected by chance (standardized

matrilineal index values greater than zero) was also non-significant, although mixed-sex groups showed a greater proportion of groups more matrilineal than expected (47/65 = 72.3% of mixed-sex groups, 10/19 = 52.6% of male groups). Given the apparent difference in matrilineality by group type, comparisons among oceans were restricted to mixed-sex groups due to the greater sample sizes across oceans compared with bachelor groups (mixed-sex Pacific n = 28, Indian 14, Atlantic 23). However, no significant differences were seen in the average standardized matrilineality index for mixed-sex groups by ocean (Pacific 0.223, Indian 0.214, Atlantic 0.276, t-test $p > 0.5790$). The standardized matrilineality index was also constructed summing across the within-social group dyads for each region. Of the 13 regions, 10 were positively matrilineal, with 8 significantly so (Table 3.1). Only the Pacific Crossing, Seychelles and Sri Lanka showed standardized matrilineality indices less than one (Table 3.1), indicating that the 'matrilineality' of groups in these regions was similar to or less than that expected given the background haplotype frequencies. The Pacific Crossing was significantly less matrilineal than expected by chance ($p = 0.011$), driven by the large mixed-sex group PX003 that was also significantly less matrilineal than expected by chance (Table 3.1). No significant differences were observed between oceans based on the regional standardized matrilineality index, but this was limited by the number of regions included (n = 13, Table 3.1).

3.5. Discussion

This study represents the first comparative examination of the molecular ecology of sperm whale social groups from all three major ocean basins. We demonstrated

differences in the molecular ecology between bachelor groups and mixed-sex groups. In addition, we showed that the amount of geographic differentiation, after accounting for social group differentiation, varies by ocean. We also showed that the mechanism for the partitioning of mtDNA diversity by social groups appears to be lenient matrilineality. Here, we define 'lenient matrilineality' as groups containing more maternal kin than expected by chance, using the maternally-inherited mtDNA as a proxy for matrilineal relatedness. However, partitioning of microsatellite diversity does not appear to be driven by increased relatedness among individuals within social groups. We detected differences in social group size by ocean, with the Pacific having the largest social groups and the Atlantic having the smallest.

3.5.1. Bachelor groups versus mixed-sex groups

Consistent with previous research suggesting that bachelor groups are smaller in size than mixed-sex groups (Best 1979), we found that bachelor groups in our dataset had significantly fewer individuals than mixed-sex groups. Bachelor groups had significantly more mtDNA haplotypes per group size than mixed-sex groups, and decreased matrilineality based on the standardized matrilineality index (although this difference was not significant). These results are consistent with bachelor groups representing a 'more random' assortment of individuals than the mixed-sex group. First-order kin appeared more likely to occur within mixed-sex groups than in bachelor groups, and the two groups that were found to be significantly more related than expected by chance were both mixed-sex groups. However, no significant difference in relatedness values between bachelor and mixed-sex groups was detected. This might be driven by male groups being

more related than previous research has suggested. This could be driven by male fidelity to areas that might increase the likelihood of related males assorting into bachelor groups. In support of this, a pair of male first-order kin was found within a group (CHG004) located at the Chagos Archipelago in the Indian Ocean.

3.5.2. How do social groups partition genetic diversity?

Over a worldwide scale, partitioning the *Odyssey* dataset by social group explained more genetic variance than previous attempts to partition the dataset by geography (Chapter 2). This confirms previous research indicating the importance of the social group in partitioning genetic diversity (Lyrholm and Gyllensten 1998, Rendell et al. 2012). What is less clear is the mechanism by which social groups partition genetic diversity. Male dispersal and close-knit sociality of female units in the sperm whale have led past researchers to draw parallels between this species and elephants (Family Elephantidae) (Best 1979, Whitehead 2003). In elephants, mixed-sex units are matrilineal, consisting of a matriarch and her offspring (Vidya and Sukumar 2005). However, previous molecular ecology studies have shown a mixture of matrilines present in sperm whale social groups (Mesnick et al. 1999, Richard et al. 1996a), as well as in individual social units (Christal et al. 1998). These results are inconsistent with either social units or groups as strictly matrilineal social structures (Christal et al. 1998, Mesnick et al. 1999, Richard et al. 1996a), and are reflected by our own results showing the presence of multiple mtDNA haplotypes within social groups. Limited female dispersal is presumably the mechanism by which different matrilines combine, as identified through previous behavioral research (Christal et al. 1998). This is supported by three cases of apparent dispersal of female

relatives from their natal mixed-sex group in the *Odyssey* dataset: one in Sri Lanka, and two in the Gulf of California.

Despite the presence of multiple matrilines in our social groups, the large values of mtDNA differentiation between social groups suggests that dispersal of females between groups occurs only on a limited basis. This is also supported by the majority of potential first-order kin detected in this study being found within the same mixed-sex group, as well as finding relationships consistent with a grandmother-mother-offspring in multiple groups, indicative of a female remaining with her mother for at least part of her own reproductive lifetime. Limited dispersal of females is also supported by 38.4% of groups containing a single mtDNA haplotype. To assess the 'lenient matrilineality' of groups comprised of multiple matrilines, we presented a new measure of the matrilineality of groups: the standardized matrilineality index. The standardized matrilineality index is more general than Weinrich et al.'s (2006) test in that it only considers whether the within-social group dyads have the same haplotype, and not what the specific identity of the haplotype is. Another attempt to quantify matrilineality in cetaceans was made by Whitehead et al. (2012), using a method similar to Jost's (2008) *D*. The formulation of Whitehead et al.'s (2012) index provides a single matrilineality index (q) based on all social groups present in an area. We believe our matrilineality index (based on the proportion of same-haplotype dyads in a social group) offers an advantage as it can provide an estimate of matrilineality on a per-social group basis, rather than just at a regional level.

Using the standardized matrilineality index, we found 69.0% of the sperm whale social groups with mtDNA haplotype frequency information (n = 87) were more matrilineal than expected by chance given regional mtDNA haplotype frequencies. This included 72.3% of mixed-sex groups and 52.6% of male groups. Except for one bachelor group, all of the 26 groups significantly more matrilineal than expected by chance were mixed-sex groups. The matrilineality index results indicate that the sperm whale can be characterized as 'leniently matrilineal', and it is likely that lenient matrilineality is the mechanism by which mtDNA differentiation between groups is enhanced.

Social groups characterized in the sperm whale ranged from strictly matrilineal, through more matrilineal than expected by chance, to not matrilineal. This variation in levels of matrilineality - in that some social groups appear to be strictly matrilineal while others are not - appears to be a feature of other 'matrilineal' species such as the wild boar, human, spotted hyena, howler monkeys, macaques, long-finned pilot whale, and the Bechstein bat (de Ruiter and Geffen 1998, Holekamp et al. 2012, Kaminski et al. 2005, Kerth et al. 2000, Oremus et al. 2013, Silk 2007, 2009, Van Horn et al. 2004, 2007). The development of the standard matrilineality index provides a method to relate within-species variation in matrilineality between social groups to reproductive fitness consequences (Silk 2007). However, the use of the standardized matrilineality index is restricted to species where the mtDNA frequencies of social groups and regions are known. Given the utility of mtDNA in tracing patterns of maternal kinship, it is surprising that some of the studies listed above have not utilized this marker.

The development of a metric to quantify matrilineality also allows comparisons within and across different species to characterize the ecological and phylogenetic factors associated with matrilineality. The standard matrilineality index would also allow for species with social groups that show only lenient matrilineality, but not strict matrilineality, to be included in such comparisons. The standardized matrilineality index could also be implemented at the higher hierarchical levels of social organization shown by species such as the elephants (de Silva and Wittemyer 2012, Wittemyer et al. 2009) and the sperm whale ('acoustic clan': Rendell and Whitehead 2003). Finally, in some species such as the African elephant, a breakdown of matrilineality has been observed in areas subjected to high levels of poaching (Archie and Chiyo 2012, Gobush and Wasser 2009). In this situation, the standardized matrilineality index could be used to demonstrate the breakdown of matrilineality due to anthropogenic pressures.

Despite the overall findings of lenient matrilineality in the sperm whale, 17 groups (10 mixed-sex, 7 bachelor) were found to be significantly less matrilineal than expected by chance. Of the 10 mixed-sex groups that were significantly less matrilineal than expected by chance (Table 3.1), 2 were above the 95^{th} percentile in terms of the number of individuals sequenced for mtDNA per group. These larger groups could have represented multiple groups that were mistakenly aggregated based on field or genetic data, reducing apparent matrilineality. An additional 5 groups were at or below the 5^{th} percentile of number of individuals sequenced for mtDNA per group (n = 2 individuals), although total estimated group size was often much larger (Table 3.1). The sampling/sequencing of additional individuals within these groups could have increased the standardized

matrilineality index if the number of same-haplotype within social-group dyads increased.

While we detected significant nuclear differentiation between sperm whale social groups, we found few instances of relatedness within groups exceeding that expected by chance given regional allele frequencies. This finding is similar to previous studies in the Gulf of Mexico (Ortega-Ortiz et al. 2012) and the Azores (Pinela et al. 2009), where the majority of sampled social groups did not show high levels of within group relatedness. This lack of relatedness was also found in the groups characterized as 'leniently matrilineal', potentially due to the presence of multiple unrelated matrilines. The lack of significant relatedness, despite high levels of microsatellite differentiation between social groups, could be driven by the different objectives of tests for genetic differentiation versus tests for relatedness. In tests of relatedness, the relatedness of group members to each other is assessed in light of background regional microsatellite allele frequencies. However, these tests do not consider whether social groups are more or less related to *other* such groups than expected by chance. In contrast, tests of genetic differentiation explicitly partition the dataset, to examine the overall degree of differentiation between groups.

3.5.3. Oceanic differences

Differences in group size and regional differentiation - after accounting for social group - were observed by ocean, with levels of social group genetic differentiation also varying by region within oceans. The Atlantic had significantly smaller mixed-sex group sizes than the Indian and Pacific Oceans, with the Pacific Ocean showing the largest average

mixed-sex group size. This is consistent with previous research by Whitehead et al. (2012) who attributed this to a tendency for multiple social units to cluster together in the Pacific, whereas units rarely form groups in the Atlantic. That the Indian Ocean also has significantly smaller group sizes than the Pacific is interesting in light of the relative differences in geographically-based genetic structure among the three oceans. Both the Atlantic and Indian Ocean show higher degrees of maternal structure (Chapter 2), and both have relatively smaller group sizes, suggesting a correlation between social group size and overall geographic structure.

Increased geographically-based genetic structure in the Indian Ocean in comparison with the Pacific was also supported by hierarchical AMOVA analyses nesting social group within region. Only the Indian Ocean showed significant geographic variation after accounting for social group differentiation. One mechanism that would explain the greater levels of geographic variation in the Indian Ocean would be a restriction in the geographic range of social groups compared with the Pacific Ocean. This is consistent with previous behavioral research showing limited dispersal of individuals around Sri Lanka (Gordon 1987), the most studied area in terms of sperm whale biology in the Indian Ocean. This is also consistent with our anecdotal findings of a social group re-sampled in Papua New Guinea (Pacific Ocean) moving a far greater distance than the social group re-sampled around Sri Lanka (Indian Ocean). This suggests that differences in social group movement and organization could drive differences in geographic structure in the sperm whale. Previously, differences in geographically-based structure and social group size between the Atlantic and Pacific Oceans have been attributed to

oceanography, predation, whaling or culture (Whitehead et al. 2012). Our data suggests that a consideration of the factors driving differences in geographic structure and social group size should also be extended to the Indian Ocean.

However, the differences between oceans do not appear to extend to overall differences in matrilineality of mixed-sex groups, suggesting that groups within different oceans represent similar overall entities in terms of the number of matrilines they contain. Long-term behavioral studies on social units in the Indian Ocean would help determine whether, like the Atlantic, units do not often combine to form larger groups in contrast with the Pacific. The differences in social group characteristics listed above seem to be largely limited to mixed-sex groups, with differences in group size by ocean not detected for bachelor groups. Microsatellite differentiation by social group and the sex ratio of mixed-sex groups was consistent across different oceans. This could suggest a similar age of dispersal for males, and similar levels of male-mediated dispersal and gene flow across social groups when comparing different oceans. Although this is suggestive that bachelor groups and males show similar ecology across oceans, even if mixed-sex groups differ, further study is warranted.

3.5.4. Could social structure drive a reduction in mtDNA diversity?

Social group structure could be influential in reducing mtDNA diversity through a sweep involving the linkage of cultural innovations within matrilineal groups with the maternal inheritance of mtDNA haplotypes (Whitehead 1998; 2005), or through lineage extinctions based on environmental heterogeneity (Tiedemann and Milinkovitch 1999) or

mass strandings of matrilineal mixed-sex groups (Alexander 2006, Siemann 1994). Previous challenges to these hypotheses have included the lack of matrilineality shown by the sperm whale (Mesnick 2001) and low overall rates of mass stranding in the sperm whale (Whitehead 1999).

In this study, we find evidence that 69.0% of sperm whale social groups contain fewer mtDNA haplotypes than expected by chance i.e. lenient matrilineality. Even though this indicates the importance of maternal social structure, the percentage falls well short of the 95.5% *strict* matrilineality required by the older model of cultural hitchhiking (Whitehead 1998). When this challenge to the hypothesis was previously raised (Mesnick 2001), it was suggested that the lack of matrilineality observed was an artifact of limited geographic sampling or ongoing social disruption resulting from whaling in the Pacific (Whitehead 1999). However, here we assessed the number of haplotypes/group size in social groups from all three oceans, and found no ocean where the matrilineality threshold required for the cultural hitchhiking model was reached. The lineage extinction hypothesis of Tiedemann and Milinkovitch (1999) appears to require strict matrilineality, and therefore also seems inconsistent with the lower levels of lenient matrilineality found in the sperm whale. The apparent level of dispersal between the social groups/units characterized in this study is also inconsistent with social groups being the assemblages on which the newer model of cultural hitchhiking occurs (Whitehead 2005). However, the data from this current study cannot speak to whether ‘higher order’ social groupings, such as ‘acoustic clan’ (clans contain social groups that have similar vocalizations), could be the assemblages on which cultural hitchhiking acts (Whitehead 2005).

Although the mass stranding hypothesis for reduced mtDNA diversity does not require strict matrilineality, it does require annual rates of mortality due to mass strandings to be around 1% (Amos 1999, Siemann 1994, Whitehead 1998; 1999). Whitehead (1998) previously noted that this rate was unlikely in the sperm whale. We reviewed documented mass stranding events (Appendix III) and agree such a high mortality rate is unlikely. The estimate of the number of mass stranding events in Appendix III is likely to understate the true number of mass strandings, as not all mass strandings are likely to have been accurately recorded. However, given the current sperm whale population size of 360,000 (Whitehead 2002), our estimate of the number of strandings since 1405 (77 strandings) would have to be an approximate 25,000× underestimate for the mass stranding hypothesis to hold. We therefore agree with Whitehead (1998): mass stranding rates do not appear to be frequent enough in the sperm whale to cause a reduction in mtDNA diversity.

3.5.5. Conclusion

Extending previous research to also include groups from the Indian Ocean, we found significant variation among the genetic and ecological characteristics of social groups by ocean and group type. Differences occurred by group type, with bachelor groups significantly smaller, containing more mtDNA haplotypes and fewer kin than mixed-sex groups. However, even within mixed-sex groups, variation in group size and genetic differentiation occurred by ocean, including the previously uncharacterized Indian Ocean. This suggests that research into the causes of differences in social group characteristics and geographic structure by ocean (e.g. Whitehead et al. 2012) should also be extended to

the Indian Ocean. Despite oceanic differences, we confirmed the importance of the social group in partitioning genetic diversity within the sperm whale (Lyrholm and Gyllensten 1998, Rendell et al. 2012), and find indications of 'lenient matrilineality' in all three oceans. Variation in relatedness, matrilineality and inclusion of kin also occurred within oceans, with some social groups more strongly kin-structured than others (as also found by Mesnick 2001), and some regions showing larger levels of genetic differentiation by social group. Taken in aggregate, our results strongly support a "kith and kin" model of mixed-sex groups consisting of both related and unrelated females as indicated by previous researchers (Gero et al. 2008, Mesnick 2001, Ortega-Ortiz et al. 2012, Richard et al. 1996a).

Even though evidence of lenient matrilineality of social groups was detected, social groups appear to contain too many matrilines to be the unit on which cultural hitchhiking or the environmental heterogeneity lineage extinction hypotheses could act to reduce mtDNA diversity. Our summary of stranding literature (Appendix III) also suggested that mass stranding events in the sperm whale are too rare for the mass stranding hypothesis to reduce mtDNA diversity. However, even given the absence of 'strict matrilineality', group size and nested AMOVA results are strongly suggestive that underlying differences in social group size and movement might drive differences in geographically-based genetic structure between oceans. Sperm whales appear to be a fascinating example of what happens when social organization drives worldwide geographic structure in a species with few boundaries to dispersal.

3.6. Acknowledgments

Genetic samples from the 'Voyage of the *Odyssey*' were collected under permit #0751-1614 from the US National Marine Fisheries Service. Principle Investigator: Iain Kerr. The authors thank all who served as staff and crew during the Voyage of the *Odyssey*, as well as J. Wise Sr. and C. LaCerte for curation of DNA from samples collected by the *Odyssey*. We thank the Oregon State University Cetacean Conservation and Genomics Laboratory for additional lab support and S. Pierszalowski for analysis recommendations. A. Liston and K. O'Malley provided valuable comments on this manuscript. This work was supported by a Mamie Markham Award and a Lylian Brucefield Reynolds Award from the Hatfield Marine Science Center; and a 2008–2011 International Fulbright Science & Technology award to A.A.

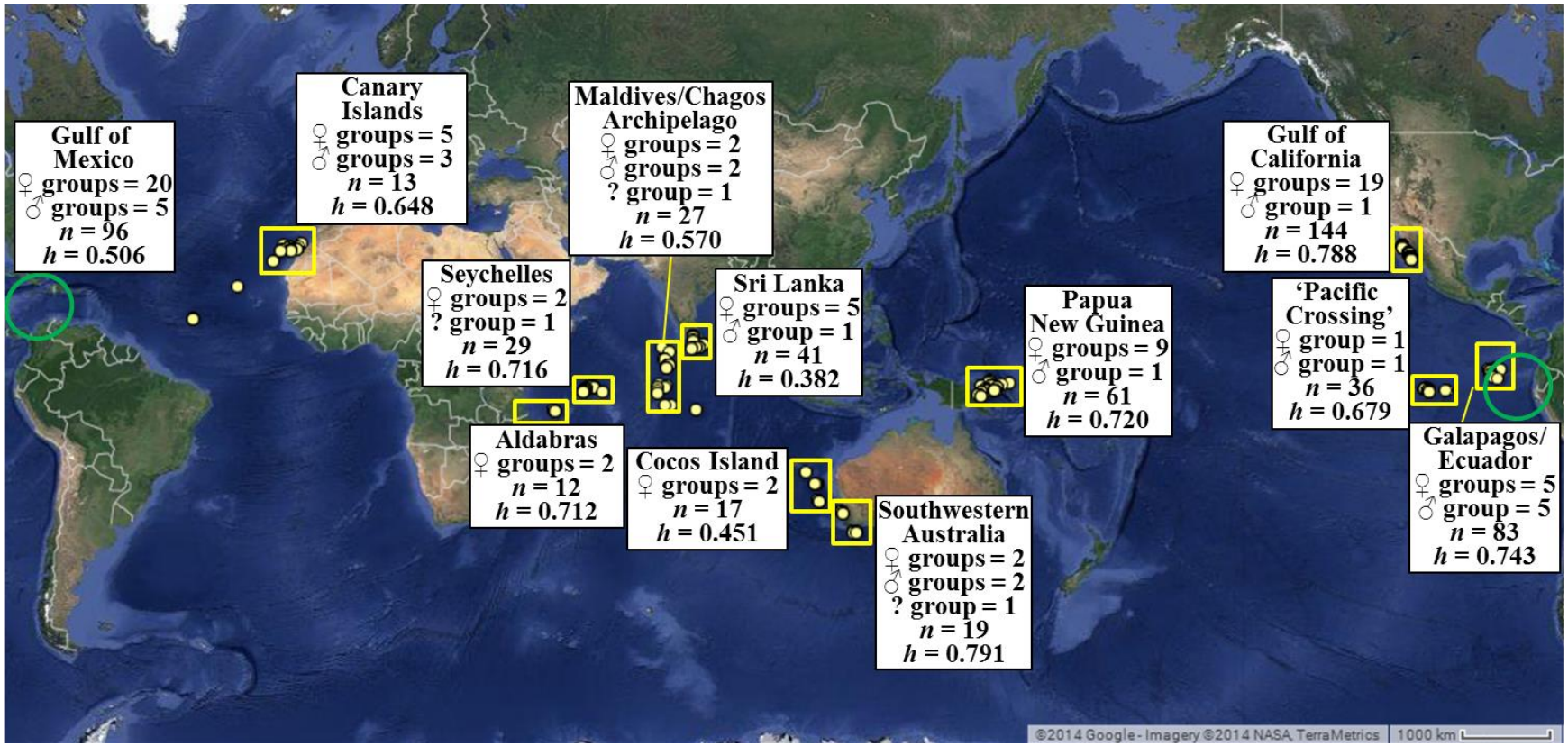


Figure 3.1: Regions with more than one sampled social group. Group type (♀ mixed-sex group; ♂ bachelor group; ? unknown). The total number of individuals (*n*) over all social groups is given, and the haplotype diversity (*h*) of each region is presented from Table 2.1. *Odyssey* based samples are depicted in yellow, other publications in green. In addition to the regions displayed on the map, the following regions were represented by just one social group each: central Atlantic (♀), central Eastern North Pacific (♀), Kiribati (♀), Mediterranean (♂), west of Colombia (♀), west of Gulf of California (♀), west of Mexico (♀), west of Peru (♀).

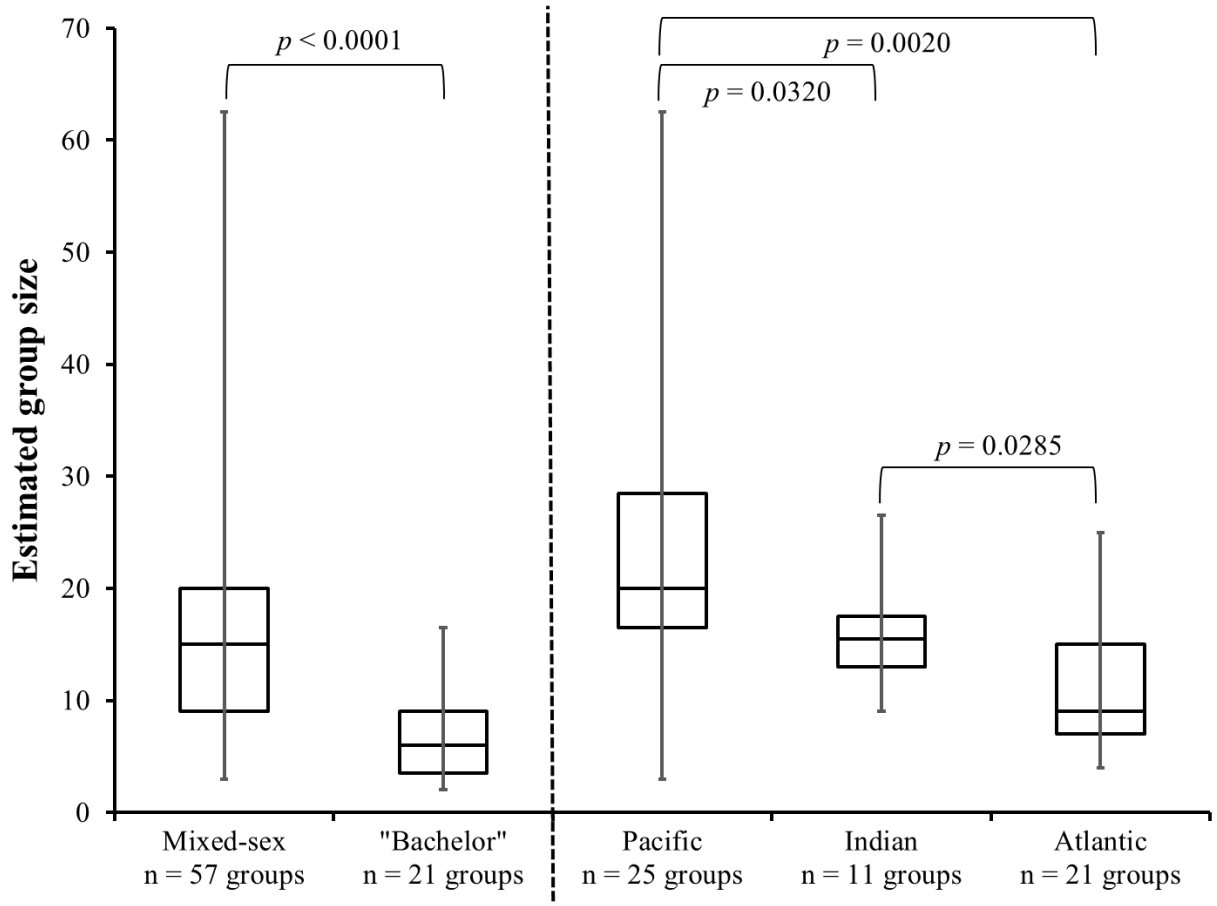
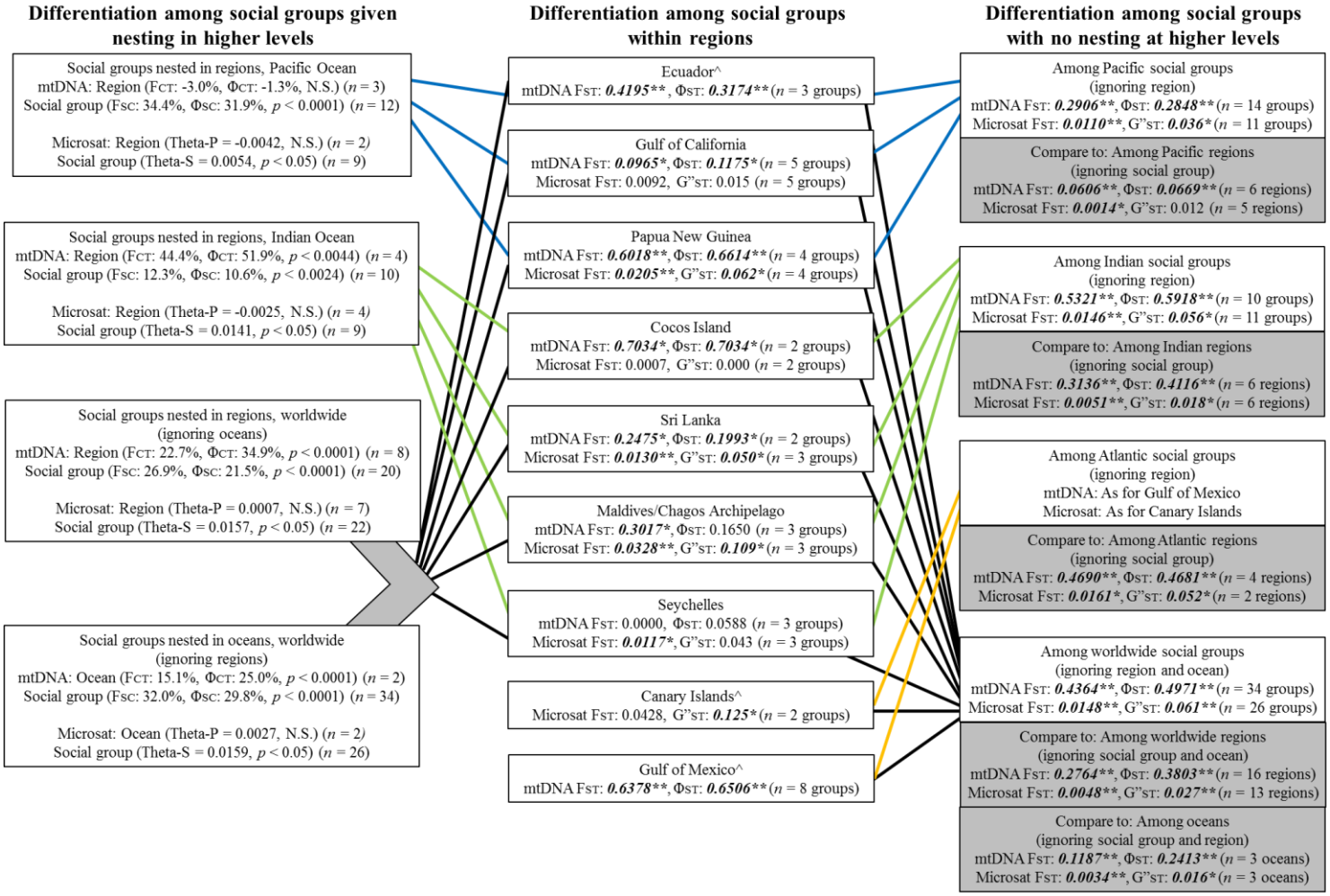


Figure 3.2: An assessment of differences in observed sperm whale social group size between mixed-sex groups and bachelor groups, and between the Pacific, Indian, and Atlantic Oceans restricted to mixed-sex groups. Minimum, lower quartile, median, upper quartile and maximum estimated group size given for each group category. *'n'* denotes number of groups in each category.

Figure 3.3: Genetic differentiation by social group within regions (center column), hierarchical analyses nesting social group within region and within ocean (left column), and non-hierarchical analyses of social group differentiation at oceanic and worldwide levels, ignoring region (right column). All groups with ≥5 individuals were included (mixed-sex, bachelor groups, and unknown), but the majority of groups (32/34 for mtDNA and 24/26 for microsatellites) were mixed-sex. Hierarchical analyses nesting social group within region were conducted both at the worldwide scale (ignoring ocean), and within the Pacific and Indian Oceans (color coding connections by oceans for within-ocean analyses). The Atlantic was excluded from nested hierarchical analyses due to a lack of sampled regions. Aldabras, southwest Australia, Kiribati, and the Pacific Crossing were all represented by only one social group so were not analyzed by region, but were included in oceanic nested analyses. For microsatellites the Theta-S (social group) and Theta-P (region or ocean) components of variance and their significance are presented. For mtDNA, the percentage of variation explained by each hierarchical level and their significance are presented. The right column gives non-nested analyses of social group at oceanic and worldwide levels, ignoring region as a factor (color coding connections by oceans for within-ocean analyses). For the center and right columns, significant AMOVA values are bolded and italicized with * significant at $p < 0.05$, ** significant at $p < 0.001$. ^ Regions included in either mtDNA of microsatellite analyses, but not both. For comparative purposes, genetic differentiation values for geographic partitioning of the dataset are also provided from Chapter 2, shown in the grey boxes.

Figure 3.3 (Continued)



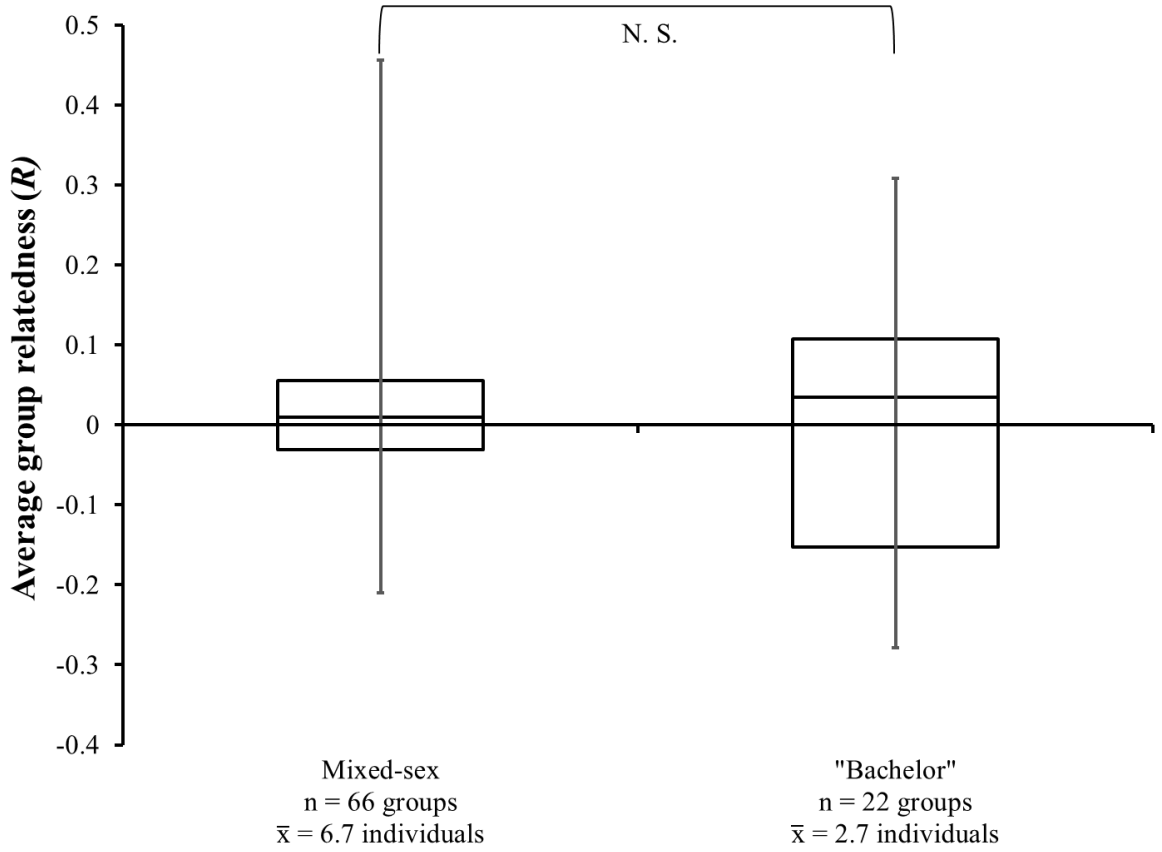


Figure 3.4: An assessment of differences in average relatedness of sperm whale mixed-sex groups versus bachelor groups, based on Queller and Goodnight's (1989) relatedness index constructed with regional microsatellite baselines in *GROUPRELATE*. Minimum, lower quartile, median, upper quartile and maximum average social group relatedness for each group type presented. Number of groups (*n*) and mean number of individuals per group included in relatedness analyses ($\bar{x}$) given.

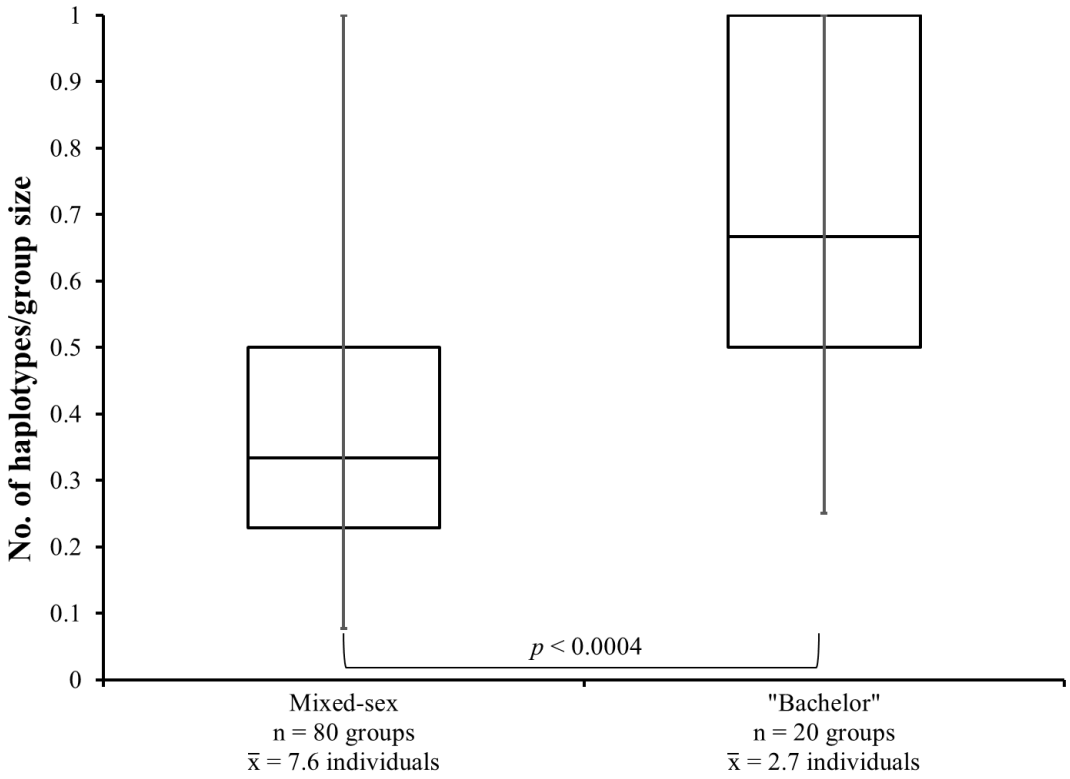


Figure 3.5: An assessment of differences in the number of mtDNA CR haplotypes divided by group size for sperm whale mixed-sex groups versus bachelor groups. Minimum, lower quartile, median, upper quartile and maximum number of mtDNA haplotypes divided by group size is given for each group type. Number of groups (*n*) and mean number of individuals sequenced for mtDNA per group ($\overline{x}$) given.

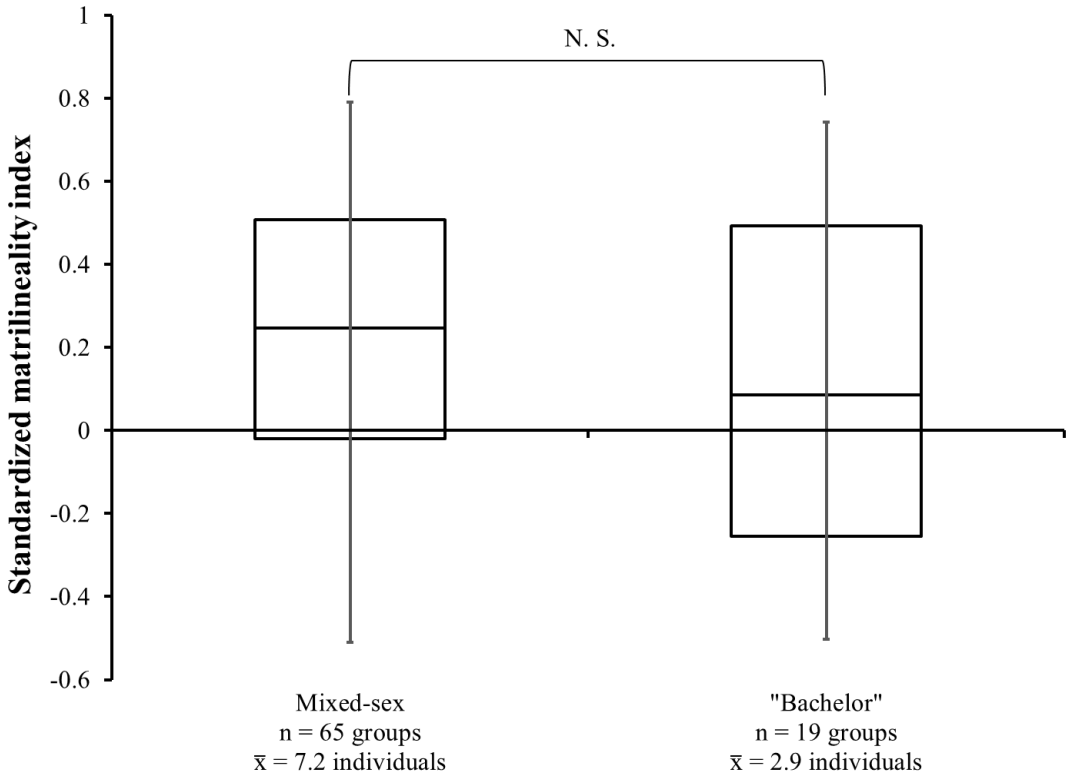


Figure 3.6: An assessment of differences in average standardized matrilineality indices for sperm whale mixed-sex groups versus bachelor groups. Minimum, lower quartile, median, upper quartile and maximum standardized matrilineality indices are given for each group type. Number of groups (*n*) and mean number of individuals sequenced for mtDNA within groups ($\overline{x}$) given. A standardized matrilineality index of zero indicates that the social group has as many same-mtDNA CR haplotype dyads as would be expected by chance given regional haplotype frequencies. Standardized matrilineality index values over zero indicate that the group is more matrilineal than expected by chance.

Table 3.1: Social groups used in genetic analyses, separated by headers indicating region, ocean, and year sampled (for some previously published groups, date unknown). For oceans, the average standardized matrilineality index over all groups is given. For region, the regional standardized matrilineality index has been calculated, with n(mtDNA) and *k* taken from Chapter 2 (not restricted to data based on social groups). Within region, groups ordered by group type (♀ female social group; ♂ bachelor group; ? unknown). Number of individuals that were included in nuclear analyses (where applicable) indicated by *n* (QC). Number of females/males in each group was determined using genetic data. Relatedness (*R)* and significance of relatedness calculated using *GROUPRELATE*. A 'Y' in the Kin column indicates first-order relationships were detected in that group, a 'D' that a first-order relationship was detected across groups (all kin relationships are detailed in the text). Number of individuals sequenced for mtDNA shown in *n*(mtDNA) column and number of mtDNA CR haplotypes per group/region in *k* column. Standardized matrilineality index (SMI) is presented by group and region. F(mtDNA), F *k*, and F standardized matrilineality index are the same analyses restricted to just confirmed female samples. *R*-values, F_{IS} and standardized matrilineality index values bolded, italicized and marked with asterisk if significant at * p <0.05; ** $p < 0.001$). '--' indicates missing data/social group not included in analysis. The AMOVA column indicates which AMOVA analyses social groups were included in (mtDNA, microsatellite: usat or both). References: [1] This study; [2] (Engelhaupt 2004); [3] (Ortega-Ortiz et al. 2012); [4] (Mesnick 2001); [5] (Richard et al. 1996a).

Group type	Group name	Estimated size	*n* (QC)	*n* (F)	*n* (M)	*R*	F_{IS}	Kin	*n* (mtDNA)	*k*	SMI (Hi_{obs} - Hi_{exp})	F (mtDNA)	F *k*	F SMI (Hi_{obs} - Hi_{exp})	AMOVA (mtDNA, usat, both)	Ref
PACIFIC											0.223			0.275		
Gulf of California, Pacific Ocean (1999, Unknown)									122	11	***0.058****	96	10	***0.078****		
♀	MES_GCA_2	--	8	--	--	--	--	--	8	4	--	--	--	--		[4]
♀	MES_GCA_3	--	10	--	--	--	--	--	10	2	--	--	--	--		[4]
♀	MES_GCA_4	--	11	--	--	--	--	--	11	2	--	--	--	--		[4]
♀	MES_GCA_5	--	6	--	--	--	--	--	6	2	--	--	--	--		[4]
♀	GCA002	27	7	7	0	0.051	0.006	D	7	3	0.111	7	3	0.084	Both	[1]
♀	GCA003	8-11	45	38	5	-0.003	-0.053	Y	45	6	***0.043****	38	5	***0.070****	Both	[1]
♀	GCA005	4	8	6	2	-0.015	0.035		8	4	-0.004	6	4	-0.046	Both	[1]
♀	GCA007	3	3	3	0	0.111	--		3	2	0.113	3	2	0.090		[1]
♀	GCA010	--	4	3	1	0.035	--		4	3	-0.043	3	2	0.090		[1]
♀	GCA011	--	4	3	1	0.023	--		4	2	0.290	3	1	***0.757****		[1]
♀	GCA012	--	2	2	0	-0.014	--		2	1	0.791	2	1	0.765		[1]
♀	GCA017	--	2	1	1	-0.142	--		2	2	***-0.209****	1	1	--		[1]
♀	GCA019	--	13	11	1	-0.018	-0.012	Y	13	3	***0.193****	11	3	0.100	Both	[1]
♀	GCA020	20+	2	2	0	0.112	--		2	1	0.791	2	1	0.765		[1]

Table 3.1 (Continued)

Group type	Group name	Estimated size	*n* (QC)	*n* (F)	*n* (M)	*R*	F_{IS}	Kin	*n* (mtDNA)	*k*	SMI (Hi_{obs} - Hi_{exp})	F (mtDNA)	F *k*	F SMI (Hi_{obs} - Hi_{exp})	AMOVA (mtDNA, usat, both)	Ref
Gulf of California, Pacific Ocean (1999, Unknown) (Cont.)									122	11	***0.058****	96	10	***0.078****		
♀	GCA023	18	2	1	1	0.229	--		2	2	***-0.209****	1	1	--		[1]
♀	GCA024	60	5	4	1	-0.109	-0.004	D	5	3	0.080	4	2	0.252	Both	[1]
♀	GCA026	30	3	2	0	-0.026	--	D	3	2	0.113	2	2	***-0.235****		[1]
♀	GCA029	--	4	3	1	***0.226****	--	D	4	1	***0.790****	3	1	***0.757****		[1]
♀	GCA030	--	3	2	1	0.063	--		3	3	***-0.222****	2	2	***-0.235****		[1]
♂	GCA009	--	2	0	1	0.162	--		2	2	***-0.209****	--	--	--		[1]
Galapagos/Ecuador, Pacific Ocean (1991: Ecu, 2000: GPG)									284	16	***0.242****	43	6	***0.260****		
♀	RIC_Ecu_A	28.5	18	12	6	--	--	--	17	4	***0.247****	11	3	***0.390****	mtDNA	[5]
♀	RIC_Ecu_B	24.3	20	16	4	--	--	--	20	3	***0.461****	16	3	***0.493****	mtDNA	[5]
♀	RIC_Ecu_C	--	18	16	2	--	--	--	18	5	-0.02	16	5	-0.032	mtDNA	[5]
♀	MES_GPG_1	--	10	--	--	--	--	--	10	2	--	--	--	--		[4]
♀	MES_GPG_2	--	5	--	--	--	--	--	5	2	--	--	--	--		[4]
♂	GPG001	9	3	0	3	-0.190	--		3	2	0.085	--	--	--		[1]
♂	GPG012	9-14	2	0	2	0.090	--		2	2	***-0.286****	--	--	--		[1]
♂	GPG015	12	3	0	3	-0.210	--		3	2	0.085	--	--	--		[1]
♂	GPG016	5	2	0	2	-0.078	--		2	2	***-0.286****	--	--	--		[1]
♂	GPG022	6	3	0	3	0.059	--		3	2	0.085	--	--	--		[1]
'Pacific Crossing', Pacific Ocean (2000)									36	8	***-0.042****	--	--	--		
♀	PX003	50-75	35	20	12	-0.031	--		34	8	***-0.042****	--	--	--	Both	[1]
♂	PX001	2-3	2	0	2	-0.279	--		2	1	0.661	--	--	--		[1]
Central eastern North Pacific, Pacific Ocean (Date unknown)									21	3	--	--	--	--		
♀	MES_CENP_1	--	21	--	--	--	--	--	21	3	--	--	--	--		[4]
West of Baja California, Pacific Ocean (Date unknown)																
♀	MES_WGC_1	--	6	--	--	--	--		6	3	--	--	--	--		[4]
West of Colombia, Pacific Ocean (Date unknown)									9	3	--	--	--	--		
♀	Mesnic_WCol_1	--	9	--	--	--	--		9	3	--	--	--	--		[4]
West of Mexico, Pacific Ocean (Date unknown)									5	2	--	--	--	--		
♀	MES_WMX_1	--	5	--	--	--	--		5	2	--	--	--	--		[4]
West of Peru, Pacific Ocean (Date unknown)									15	4	--	--	--	--		
♀	MES_WPer_1	--	15	--	--	--	--		15	4	--	--	--	--		[4]
Kiribati, Pacific Ocean (2000)									13	4	0.005	10	3	0.003		
♀	KR02001	18-23	11	9	1	-0.061	--		11	3	0.005	9	3	0.003	Both	[1]

Table 3.1 (Continued)

Group type	Group name	Estimated size	*n* (QC)	*n* (F)	*n* (M)	*R*	F_{IS}	Kin	*n* (mtDNA)	*k*	SMI (Hi_{obs} - Hi_{exp})	F (mtDNA)	F *k*	F SMI (Hi_{obs} - Hi_{exp})	AMOVA (mtDNA, usat, both)	Ref
Papua New Guinea, Pacific Ocean (2001)									63	8	***0.373****	52	6	***0.345****		
♀	PNG002	18	2	2	0	-0.210	--		1	1	--	1	1	--		[1]
♀	PNG006	15-18	10	10	0	0.039	0.008		10	3	0.025	10	3	-0.014	Both	[1]
♀	PNG007	50	8	8	0	-0.037	0.022		8	1	***0.713****	8	1	***0.673****	Both	[1]
♀	PNG010	22-23	16	12	1	0.012	0.032		16	3	***0.468****	12	2	***0.511****	Both	[1]
♀	PNG011	5-25	4	3	1	0.076	--	Y	4	1	***0.713****	3	1	***0.703****		[1]
♀	PNG012	18	4	4	0	0.054	--		3	2	0.042	3	2	0.036		[1]
♀	PNG013	10	4	3	1	0.046	--		4	1	***0.713****	3	1	***0.703****		[1]
♀	PNG014	18-20	7	7	0	0.015	-0.022		7	2	***0.233****	7	2	***0.199****	Both	[1]
♀	PNG018	30	4	4	0	0.019	--		4	2	0.046	4	2	0.004		[1]
♂	PNG008	4-5	4	0	4	0.054	--		4	3	-0.121	--	--	--		[1]
INDIAN											0.214			0.191		
Southwestern Australia, Indian Ocean (2001-2002)									21	5	***0.159****	9	3	-0.107		
?	AUS001	11-13	3	0	0	-0.097	--		3	1	***0.759****	--	--	--		[1]
♀	AUS005	17-20	2	1	0	0.043	--		2	2	***-0.225****	--	--	--		[1]
♀	S_AUS008	18	8	8	0	-0.059	--		8	3	0.001	8	3	-0.107	Both	[1]
♂	S_AUS006	9	2	0	1	-0.213	--		2	2	***-0.225****	--	--	--		[1]
♂	S_AUS007	2-3	4	0	3	0.033	--		4	1	***0.743****	--	--	--		[1]
Cocos Island, Indian Ocean (2002)									18	3	***0.314****	18	3	***0.314****		
♀	COC001	11	6	6	0	-0.048	-0.010		6	2	-0.104	6	2	-0.104	Both	[1]
♀	COC002	11-20	11	11	0	-0.064	0.032		11	1	***0.427****	11	1	***0.427****	Both	[1]
Sri Lanka, Indian Ocean (2003)									42	3	-0.014	30	2	-0.053		
♀	SRI002	--	25	20	1	-0.013	***0.075****	Y	21	3	***-0.155****	16	2	***-0.181****	Both	[1]
♀	SRI004	--	2	2	0	-0.077	--	D	2	1	0.371	2	1	0.314		[1]
♀	SRI005	--	13	10	0	-0.039	0.005	D	13	1	***0.370****	10	1	***0.278****	Both	[1]
♀	SRI006	--	10	9	1	0.007	-0.023		2	1	0.371	2	1	0.314	usat	[1]
♀	SRI009	13	2	1	1	***0.456****	--		--	--	--	--	--	--		[1]
♂	SRI003	8	3	0	3	0.113	--		3	2	***-0.308****	--	--	--		[1]
Maldives/Chagos Archipelago, Indian Ocean (2002)									33	4	***0.399****	8	1	--		
?	MAL002	5	4	0	0	-0.056	--		4	1	***0.557****	--	--	--		[1]
♀	MAL008	16	9	5	2	-0.031	-0.010		9	1	***0.551****	5	1	--	Both	[1]
♀	MAL012	14	6	4	0	0.097	-0.003		5	1	***0.547****	3	1	--	Both	[1]
♂	MAL013	7	3	0	2	-0.065	--		3	1	0.562	--	--	--		[1]
♂	CHG004	1-13	6	0	6	0.022	0.029	Y	6	3	-0.186	--	--	--	Both	[1]

Table 3.1 (Continued)

Group type	Group name	Estimated size	*n* (QC)	*n* (F)	*n* (M)	*R*	F_{IS}	Kin	*n* (mtDNA)	*k*	SMI (Hi_{obs} - Hi_{exp})	F (mtDNA)	F *k*	F SMI (Hi_{obs} - Hi_{exp})	AMOVA (mtDNA, usat, both)	Ref
Aldabras, Indian Ocean (2002)									12	3	***0.138****	6	3	0.280		
♀	ALD001	9	3	3	0	-0.002	--		3	1	0.647	3	1	0.615		[1]
♀	ALD003	13	9	3	2	-0.032	--		9	2	0.099	3	2	-0.051	Both	[1]
Seychelles, Indian Ocean (2002)									31	6	-0.021	17	4	-0.077		
?	SEY003	3-23	7	0	0	-0.001	-0.023		7	3	0.027	--	--	--	Both	[1]
♀	SEY004	23-30	7	2	0	-0.034	0.057		7	3	0.170	2	1	0.674	Both	[1]
♀	SEY002	14-20	15	15	0	-0.017	0.022		15	4	***-0.070****	15	4	***-0.082****	Both	[1]
ATLANTIC											0.276			0.437		
Mediterranean, Atlantic Ocean (2004, 2009)									2	1	--	--	--	--		
♂	N_MED015	2	2	0	2	-0.188	--		--	--	--	--	--	--		[1]
Canary Islands, Atlantic Ocean (2005)									14	3	***0.498****	9	2	0.443		
♀	CNI005	12-15	5	4	1	0.060	-0.009	Y	4	1	***0.607****	3	1	0.433	usat	[1]
♀	CNI007	4-5	2	1	1	0.110	--		1	1	--	1	1	--		[1]
♀	CNI009	12-15	5	5	0	0.043	***-0.123****		2	1	0.618	2	1	0.440	usat	[1]
♀	CNI012	6-11	2	1	1	-0.012	--		2	2	***-0.382****	1	1	--		[1]
♀	CNI014	7-10	2	1	1	-0.124	--		1	1	--	1	1	--		[1]
♂	CNI002	8-13	2	0	2	0.308	--		1	1	--	--	--	--		[1]
♂	CNI010	15-18	2	0	2	0.175	--		--	--	--	--	--	--		[1]
♂	CNI013	3	2	0	1	-0.145	--		2	1	0.618	--	--	--		[1]
Central Atlantic, Atlantic Ocean (2005)									2	2	--	--	--	--		
♀	C_ATL006	10-20	3	2	0	--	--	Y	2	2	--	--	--	--		[1]
Gulf of Mexico, Atlantic Ocean (2000-01)									230	5	***0.270****	--	--	--		
♀	G0	6	3	2	1	-0.009	--		3	1	0.497	--	--	--		[2, 3]
♀	G1	6	5	4	1	0.07	--	Y	5	1	***0.507****	--	--	--	mtDNA	[2, 3]
♀	G2	--	3	3	0	0.27	--	Y	3	1	0.497	--	--	--		[2, 3]
♀	G3	7	7	7	0	0.095	--	Y	7	1	***0.509****	--	--	--	mtDNA	[2, 3]
♀	G4	4	4	3	1	-0.004	--		4	1	***0.505****	--	--	--		[2, 3]
♀	G5	7	5	4	1	0.036	--	Y	5	1	***0.507****	--	--	--	mtDNA	[2, 3]
♀	G10	11	5	4	1	0.056	--		5	1	***0.507****	--	--	--	mtDNA	[2, 3]
♀	G11	9	4	3	1	0.242	--	Y	4	2	-0.161	--	--	--		[2, 3]
♀	G13	12	3	2	1	0.201	--	Y	3	2	-0.170	--	--	--		[2, 3]
♀	G14	15	6	5	1	0.007	--		6	1	***0.509****	--	--	--	mtDNA	[2, 3]
♀	G15	25	4	4	0	0.021	--		4	2	-0.162	--	--	--		[2, 3]
♀	G16	22	4	4	0	0.005	--		4	1	***0.505****	--	--	--		[2, 3]
♀	G17	22	3	3	0	0.061	--		3	1	0.497	--	--	--		[2, 3]

Table 3.1 (Continued)

Group type	Group name	Estimated size	*n* (QC)	*n* (F)	*n* (M)	*R*	F_{IS}	Kin	*n* (mtDNA)	*k*	SMI (Hi_{obs} - Hi_{exp})	F (mtDNA)	F *k*	F SMI (Hi_{obs} - Hi_{exp})	AMOVA (mtDNA, usat, both)	Ref
Gulf of Mexico, Atlantic Ocean (2000-01) (Cont.)									230	5	***0.270****	--	--	--		
♀	G18	10	8	7	1	0.024	--	Y	8	3	-0.102	--	--	--	mtDNA	[2, 3]
♀	G19	8	2	2	0	-0.032	--		2	1	0.488	--	--	--		[2]
♀	G20	18	5	1	4	0.025	--		5	4	***-0.393****	--	--	--	mtDNA	[2]
♀	G21	8	2	2	0	-0.004	--		2	1	0.488	--	--	--		[2]
♀	G22	9	7	7	0	-0.04	--		7	1	***0.509****	--	--	--	mtDNA	[2]
♀	G23	7	2	1	1	0.019	--		2	2	***-0.510****	--	--	--		[2]
♀	G24	15	2	2	0	-0.042	--		2	1	0.488	--	--	--		[2]
♂	G6	2	2	0	2	0.261	--		2	1	0.488	--	--	--		[2, 3]
♂	G7	3	3	0	3	0.037	--		3	3	***-0.503****	--	--	--		[2, 3]
♂	G8	6	3	0	3	0.055	--		3	1	0.497	--	--	--		[2, 3]
♂	G9	6	2	0	2	-0.155	--		2	2	***-0.500****	--	--	--		[2, 3]
♂	G12	12	2	0	2	0.188	--		2	1	0.488	--	--	--		[2, 3]

Table 3.2: A worked example of the calculation of the standardized matrilineality index. This example consists of four social groups, and four haplotypes present in a region (A, B, C, D). The number of same-haplotype dyads within a social group are calculated by haplotype, and then summed to give the total number of same-haplotype dyads found in a social group. The proportion relative to the total number of within group dyads is then calculated (Hi_{obs}). Using the regional haplotype frequencies (given in the 'regional totals' column), Monte Carlo sampling of the expected proportion of same-haplotype within-social group dyads is carried out (code given in Appendix II, Figure II.1). The mean of this distribution (Hi_{exp}) can be used to calculate the standardized matrilineality index (Hi_{obs} - Hi_{exp}). Significance can be assessed as the number of permutations that exceed Hi_{obs} divided by the total number of permutations. As two different sizes of social group (n = 4, 6 dyads and n = 8, 28 dyads) were present in this example, two separate Monte Carlo simulations were carried out. The same calculations can also be carried out at the regional level by summing over the same-haplotype within-social group dyads and total within social group dyads calculated on a per-group basis (shown in the right column of the table).

Table 3.2 (Continued)

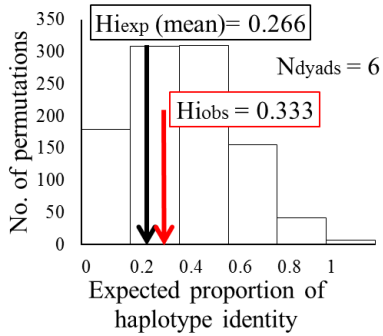
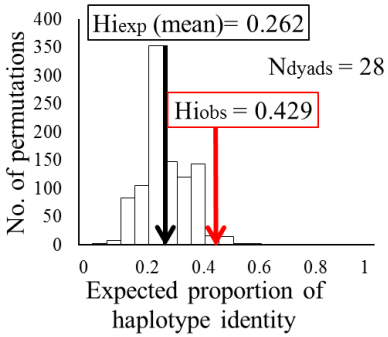
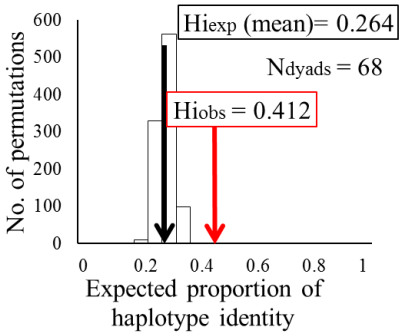
Haplotype	Group 1	Group 2	Group 3	Group 4	Regional totals
A	2			4	6
B	2	2			4
C		2	4		6
D			4	4	8
Total group size	4	4	8	8	
Observed haplotype A-A dyads	= (2*(2-1))/2 = 1	0	0	6	= 1 + 0 + 0 + 6 = 7
Observed haplotype B-B dyads	= (2*(2-1))/2 = 1	1	0	0	= 1 + 1 + 0 + 0 = 2
Observed haplotype C-C dyads	= (0*(0-1))/2 = 0	1	6	0	= 0 + 1 + 6 + 0 = 7
Observed haplotype D-D dyads	= (0*(0-1))/2 = 0	0	6	6	= 0 + 0 + 6 + 6 = 12
Total observed same-haplotype dyads	= 1 + 1 + 0 + 0 = 2	2	12	12	= 2 + 2 + 12 + 12 = 28
Total within group dyads	= (4*(4-1))/2 = 6	6	28	28	= 6 + 6 + 28 + 28 = 68
Observed proportion of haplotype identity (Hi_{obs})	= 2/6 = 0.333	0.333	0.429	0.429	= 28/68 = 0.412
					
Expected mean prop. of haplotype identity (Hi_{exp})	0.266 (from graph)	0.266	0.262	0.262	0.264
Standardized matrilineality index = Hi_{obs} - Hi_{exp}	= (0.333 - 0.266) = 0.067	0.067	0.167	0.167	0.148
Significance = $\frac{\text{no. of permutations} \geq Hi_{obs}}{\text{total no. of permutations}}$	= 309/1000 = 0.309	0.309	0.035	0.035	< 0.001

Table 3.3: Inferred "tandem movements" of multiple individuals from a single social group. In some instances, a range in time and distance is given for each movement, as the multiple individuals present in a movement event were not always simultaneously sampled.

Tandem movement	Number of individuals and sample codes	Region and social group	Total time (days) between first and last sample	Total distance (km) between first and last sample
a	2 (GCA_IND_004, GCA_IND_005)	Gulf of California (GCA003)	10.1 - 11.2	4.9 - 11.5
b	2 (GCA_IND_010, GCA_IND_012)	Gulf of California (GCA003)	2.9 - 3.0	37.8 - 40.3
c	2 (PX_IND_001, PX_IND_002)	'Pacific Crossing' (PX003)	6.8 - 7.1	61.8 - 83.3
d	2 (PX_IND_001, PX_IND_003)	'Pacific Crossing' (PX003)	3.9	142.8 - 151.8
e	2 (PX_IND_003, PX_IND_005)	'Pacific Crossing' (PX003)	2.8	95.2 - 97.9
f	2 (PNG_IND_003, PNG_IND_004)	Papua New Guinea (PNG008)	1.3	101.0 - 102.0
g	2 (PNG_IND_014, PNG_IND_017)	Papua New Guinea (PNG010)	50.7 - 50.9	506.8 - 508.2
h	4 (SRI_IND_002, SRI_IND_003, SRI_IND_005, SRI_IND_011)	Sri Lanka (SRI002)	38.0 - 38.2	8.9 - 22.2

4. Low diversity in the mitogenome of sperm whales revealed by next-generation sequencing

Alana Alexander, Debbie Steel, Beth Slikas, Kendra Hoekzema, Colm Carraher, Matthew Parks, Richard Cronn, and C. Scott Baker

Genome Biology and Evolution, 2013

Online only (Oxford University Press)

5:113-129

4. Low diversity in the mitogenome of sperm whales revealed by next-generation sequencing

4.1. Abstract

Large population sizes and global distributions generally associate with high mitochondrial DNA control region (CR) diversity. The sperm whale (*Physeter macrocephalus*) is an exception, showing low CR diversity relative to other cetaceans, however diversity levels throughout the remainder of the sperm whale mitogenome are unknown. We sequenced 20 mitogenomes from 17 sperm whales representative of worldwide diversity using Next Generation Sequencing (NGS) technologies (Illumina GAIIx, Roche 454 GS Junior). Resequencing of three individuals with both NGS platforms and partial Sanger sequencing showed low discrepancy rates (454-Illumina: 0.0071%; Sanger-Illumina: 0.0034%; Sanger-454: 0.0023%) confirming suitability of both NGS platforms for investigating low mitogenomic diversity. Using the 17 sperm whale mitogenomes in a phylogenetic reconstruction with 41 other species, including 11 new dolphin mitogenomes, we tested two hypotheses for the low CR diversity. First, the hypothesis that CR-specific constraints have reduced diversity solely in the CR was rejected as diversity was low throughout the mitogenome, not just in the CR (overall diversity π=0.096%; protein-coding 3rd codon=0.22%; CR=0.35%), and CR phylogenetic signal was congruent with protein-coding regions. Second, the hypothesis that slow substitution rates reduced diversity throughout the sperm whale mitogenome was rejected as sperm whales had significantly higher rates of CR evolution and no evidence of slow coding region evolution relative to other cetaceans. The estimated time to most recent

common ancestor for sperm whale mitogenomes was 72,800 to 137,400 years ago (95% highest probability density interval), consistent with previous hypotheses of a bottleneck or selective sweep as likely causes of low mitogenome diversity.

4.2. Introduction

Sperm whales (*Physeter macrocephalus)* are distributed throughout all oceans of the world, from the Arctic to the Southern Ocean (Best 1979, Whitehead 2003) and are abundant, with a world-wide estimate of ~360,000 individuals (Whitehead 2002). Despite being hunted intensively over the last two centuries, sperm whales appear to have suffered only localized declines due to whaling (Best 1979, Whitehead 2002; 2003). Generally large population sizes and widespread distributions are associated with high mitochondrial DNA (mtDNA) diversity (Mulligan et al. 2006). However, despite their global distribution and high abundance, sperm whales have been noted for extremely low mtDNA control region (CR) diversity relative to other cetaceans (Lyrholm and Gyllensten 1998, Whitehead 1998), both in terms of number of haplotypes and nucleotide diversity (Table 4.1). To date, only 28 mtDNA haplotypes have been reported from 1,167 samples of sperm whales, based on 400 bp of the CR (Table 4.2). The widespread geographic sampling of these haplotypes, and the relatively high abundance of sperm whales, argues against explanations of biased geographic sampling or a whaling-mediated bottleneck (Whitehead 1998). The low CR diversity is also reflected in the weak mtDNA CR phylogeographic structure found in the sperm whale (Lyrholm and Gyllensten 1998): the three most common worldwide haplotypes ("A", "B" and "C: see Table 4.2 for haplotype definitions) are shared in whales from the Pacific and Atlantic

Oceans (Engelhaupt et al. 2009, Lyrholm and Gyllensten 1998, Mesnick et al. 2011, Ortega-Ortiz et al. 2012, Rendell et al. 2012, Richard et al. 1996a, Whitehead et al. 1998).

A number of hypotheses have been proposed to explain the low diversity in the mtDNA CR of the sperm whale, including (1) an overall slowing of mtDNA substitution rates (Lyrholm et al. 1996, Whitehead 1998), as found for other large whale species (Jackson et al. 2009); (2) a population bottleneck (Lyrholm and Gyllensten 1998, Lyrholm et al. 1996); (3) a selective sweep of mtDNA by 'hitchhiking' either with maternal cultural innovations in matrilineal social groups (Whitehead 1998; 2005), or variation in ocean productivity affecting the success of different social groups (Tiedemann and Milinkovitch 1999); and (4) lineage extinctions relating to the mass stranding of matrilineal groups (Whitehead 1998). We suggest a fifth hypothesis: that constraints in the CR can restrict sites that accumulate variation, leading to saturation of sites free to vary in the CR relative to mtDNA protein-coding regions. This hypothesis has been supported in other groups such as killer whales (*Orcinus orca)* and fishers (*Martes pennanti)* (Knaus et al. 2011, Morin et al. 2010a), and is also consistent with previous observations of substitutional hot-spots in the sperm whale CR (Lyrholm et al. 1996).

Here we extend estimates of sperm whale mtDNA CR diversity using Next Generation Sequencing (NGS) technologies (Illumina GAIIx and Roche GS Junior 454) to sequence 17 sperm whale mitogenomes chosen to represent the worldwide diversity of CR haplotypes (Table 4.2). Given the expectation of low diversity in the sperm whale

mitogenome, and the potential for NGS error to inflate diversity estimates (Shen et al. 2010), we first investigated discrepancy rates between the two NGS platforms by resequencing three mitogenomes. We also calculated NGS discrepancy rates with targeted comparison to Sanger sequencing for all sperm whale mitogenomes included in this study. Additionally, we developed quality control (QC) criteria to validate variable sites among the mitogenomes, and took a number of precautions against co-amplification of numts. We then aligned these sperm whale mitogenomes to the available NCBI Reference Sequence (RefSeq) cetacean mitogenomes and to the mtDNA protein-coding genes and CR from a further 11 dolphin species/subspecies contributed by this study (including 8 previously unpublished species/subspecies) for a total of 42 cetacean species included in our analyses.

Using this dataset, we investigated two of the hypotheses for low sperm whale mtDNA CR diversity: (1) constraints acting solely on the CR resulting in reduced variation relative to the remainder of the mitogenome, or (2) slow substitution rates resulting in low diversity either in the CR or throughout the entire sperm whale mitogenome. To investigate the constraint hypothesis (hypothesis 1), we quantified intraspecific sperm whale diversity in the mtDNA CR relative to other regions of the mitogenome. We also compared intraspecific phylogenetic signal from the CR to that of the combined protein-coding regions and tested for phylogenetic congruence (Farris et al. 1995).

To test the hypothesis of slow substitution rates in the CR and protein-coding regions of the sperm whale mitogenome (hypothesis 2), we calculated Bayesian estimates of

mitogenomic substitution rates utilizing one of the largest cetacean mitogenome phylogenies constructed to date: 44 mitogenomes representing 42 species (Appendix IV, Table IV.1). The sperm whale CR and protein-coding substitution rates were then contrasted with other cetacean species. Finally, we also investigated the likelihood of population bottlenecks or selective sweeps acting on this species by calculating the time to most recent common ancestor (TMRCA) of sperm whale mitogenomes.

4.3. Material and Methods

4.3.1. Sample collection and PCR amplification of mitogenomes

The 17 sperm whale skin samples were collected from stranded individuals represented in the New Zealand Cetacean Tissue Archive (CeTA) at the University of Auckland (Thompson et al. In Press) and from the Oregon Marine Mammal Stranding Network (OMMSN) Tissue Archive (Appendix IV, Table IV.2). DNA was extracted from these tissue samples using a standard phenol/chloroform method (Sambrook et al. 1989), modified by Baker et al. (1994) for smaller samples. Sequences of the CR and mtDNA protein-coding genes from 11 dolphin species/subspecies described in Carraher (2004), but previously unpublished, were also generated from skin samples curated in CeTA using the same DNA extraction methods.

The mitogenome of each sperm whale was amplified in five overlapping fragments ranging from 3.0 kbp to 4.3 kbp in long-range PCR (LR-PCR) reactions using high fidelity Phusion® Polymerase (New England Biolabs, USA). Thermoprofiles consisted of an initial denaturation step of 98 °C for 30 s followed by 35 cycles of 98 °C for 8 s, the

specific annealing temperature for each fragment for 30 s, and 72 °C for 1 min 15 s; followed by a final extension of 72 °C for 10 min. Reagent concentrations for each reaction were as follows: 1× Phusion® HF Buffer (NEB, USA); 0.5 μM of each primer (IDT); 2% DMSO (NEB, USA); 15-30 ng of template DNA, dNTP (Promega, USA) and Phusion® Polymerase (NEB, USA) concentration varied by fragment as described in Appendix IV, Figure IV.1 (primer sequences and fragment-specific annealing temperatures are also detailed in this Appendix) and ddH_2O to 20 μL.

As described by Carraher (2004), the CR and mtDNA protein-coding genes from 11 dolphin species/subspecies were amplified in 11 shorter (1 to 2.5 kbp), and partially overlapping, fragments (Appendix IV, Table IV.3). Fragments were amplified using the Expand Long Template PCR System (Roche Diagnostics) as per the manufacturer's recommendations. Thermoprofiles consisted of an initial denaturation step of 93 °C for 2 min followed by 10 cycles of 93 °C for 30 s, 60 °C for 30 s, 68 °C for 45 s per kilobase, then 24 cycles of 93 °C for 30 s, 60 °C for 30 s, 68 °C for 45 s per kilobase, increased by 20 s/cycle, with a final extension of 68 °C for 10-15 min.

4.3.2. Illumina sequencing, assembly and quality control (QC)

Thirteen sperm whale mitogenomes (Appendix IV, Table IV.2) were sequenced using a single-end 40 bp run on one lane of an Illumina GAIIx. The input library was prepared using multiplex identifier (MIDs) adapters to identify each sample (Cronn et al. 2008), following standard Illumina library preparation v 1.12, combined in equimolar quantities and run on the Illumina at a combined concentration of 5 pM. Output data was filtered

using Illumina's *GA pipeline v 1.5.0*. Reads were then sorted by MIDs using the program *bcsort* (Jennings et al. 2011) before being converted to the PHRED scoring system using a custom PERL script (Davies 2009). Reads were trimmed from the 3' end, using *BWA v 0.5.7* (Li and Durbin 2009), until a base with a PHRED quality above 20 was reached, to ensure only high-quality fragments of reads were included. Reads were then assembled to the sperm whale mitogenome available from NCBI (NC_002503: (Arnason et al. 2000)) using the program *BWA*. To compensate for the artifact of low coverage at the beginning and end of the arbitrarily linearized alignment (as the mitogenome is circular), a second assembly was conducted with the reference genome re-linearized at 8,213 bp relative to the original reference. *Samtools v 0.1.7* (Li et al. 2009a) was used to generate the consensus sequence (>51%) from the *BWA* assembly.

4.3.3. Roche 454 sequencing, assembly and quality control (QC)

Mitogenomes of seven sperm whales, including three that had been sequenced using the Illumina technology, were sequenced on a Roche GS Junior 454 (Appendix IV, Table IV.2). Sample libraries were prepared using MIDs and the standard Roche GS Junior 454 rapid library preparation, emulsion PCR and sequencing steps (June 2010 version). Reads were separated by their MIDs using the program *sfftools* (Roche, USA). For each sample, reads were then trimmed and assembled against the sperm whale mitogenome (NC_002503) using the default settings of *GS Reference Mapper* (Roche, USA). As for the Illumina sequencing described above, combined assemblies were generated by re-linearization of the reference and the consensus base call at each site was used.

4.3.4. Sanger sequencing of dolphin mitogenomes and sperm whale CRs

As previously described above, the CR and mtDNA protein-coding genes from 11 dolphin species/subspecies were amplified using LR-PCR. Excess primers and nucleotides were removed using a standard SAP dephosphorylation protocol (USB®, USA). Products were then sequenced with BigDye™ dye terminator chemistry (Applied Biosystems, USA) with 43 internal primers detailed in Appendix IV, Table IV.3, followed by dye-terminator removal using Agencourt CleanSEQ beads (Beckman Coulter, USA). Sequences were run on an ABI3730xl and assembled manually with a published dolphin mitogenome sequence, the white-beaked dolphin (*Lagenorhynchus albirostris*, NC_005278: (Arnason et al. 2004)). To evaluate the risk of numts affecting the analysis, sequences were reviewed to ensure overlap of long-range fragments.

Initial CR haplotype identity of sperm whales used in this study was based on PCR amplification and Sanger sequencing using the primers dlp1.5 (Baker et al. 1998) and tphe as described by Carroll *et al.* (2011). Substitutions in the first (5') 400 bp of sequence were used to define the CR haplotypes referred to in previous publications (see Table 4.2 for haplotype definitions and haplotype distributions).

4.3.5. Estimation of NGS discrepancy rates

To estimate NGS discrepancy rates, three sperm whale samples were independently amplified and sequenced on both the Illumina and 454 platforms, and limited Sanger resequencing was carried out using all sperm whales included in this study. Final consensus sequences were aligned using *Sequencher® v 4.6* (Gene Codes Corporation,

USA) and discrepancy rates calculated by comparing all individuals with sequence available from two or more sequencing (NGS or Sanger) platforms.

4.3.6. QC and validation of variable sites

Coverage plots were examined for gaps in coverage indicative of the incorporation of linear numts. Mitogenomes were then aligned using *Sequencher® v 4.6*, and variable sites in the multiple sperm whale alignment were accepted if all mitogenomes met a minimum sequencing depth of 15× (Smith et al. 2008) and the minimum measure of mapping quality determined for each platform, as described below (mapping quality is the probability that reads are correctly placed during the assembly process, as opposed to the quality of individual reads assessed during the assembly process). At variable sites, Illumina-sequenced mitogenomes were required to have a mapping quality of at least 20 at the variable site (determined by *BWA* using a PHRED scale), with ≥ 15 for the 5 bp stretch of non-variable sites adjacent to the variable site (Li et al. 2008, Li et al. 2010). If an Illumina-sequenced mitogenome had a mixed-base signal present at a variable site, identical (duplicate) reads were removed from that assembly in order to minimize the potential impact of PCR over-sampling. In addition, for Illumina-sequenced mitogenomes, the possibility of strand bias was tested using a Fisher's exact test to test for a difference in the distribution of reads for each nucleotide in the forward and reverse directions of variable sites. For 454-sequenced mitogenomes variable sites had to fulfill the *454HCDiffs* (Roche, USA) criteria: the variable site was required to be covered by both forward and reverse reads (which addressed the problem of strand bias), or at least

five reads with quality scores over 20 (using the *GS Reference Mapper* scale), or 30 if the difference between reads involved a homopolymer ≥5 bp in length.

If mapping quality criteria were not met, or NGS coverage was <15× at a variable site for a particular mitogenome, Sanger sequencing was used to confirm the sequence using the primers listed in Appendix IV, Figure IV.1 or Appendix IV, Table IV.4. Due to the known error rate of 454 in homopolymeric regions (Kircher and Kelso 2010), Sanger and Illumina sequencing were used to confirm sequences wherever variable homopolymers ≥5 bp in length were present. All variable sites within primer regions were masked.

4.3.7. Annotation and analysis of intraspecific diversity in the sperm whale

The sperm whale mitogenomes were annotated using the reference sequence sperm whale mitogenome (NC_002503) available from NCBI (Arnason et al. 2000). Nucleotide diversity analyses (including codon and gene-specific analyses) were conducted in *MEGA v5.0.5* (Tamura et al. 2011) on the *Sequencher® v 4.6* alignment of the 17 sperm whale mitogenomes using a standard error computation of 10,000 bootstrap replicates, the maximum composite likelihood substitution model, homogenous rates among lineages and different rates among sites (with a gamma parameter of 1.5, as indicated by Bayesian modeling).The relative levels of diversity in the CR and other mitogenome regions were compared using the standard error computations calculated above to provide 95% confidence intervals. The same parameters were also used in *MEGA v5.0.5* to calculate the maximum pairwise divergence by gene between the sperm whale mitogenomes generated in this study in order to examine numt influence. To test for

congruence in phylogenetic signal of the CR and protein-coding regions we used an incongruence length difference (ILD) test (Farris et al. 1995), implemented with 1,000 replicates and the default search parameters in *TNT v1.1* (Willi Hennig Society version, (Goloboff et al. 2008)) using a customized script (Siddall). Phylogenetic reconstructions were also carried out to directly visualize the comparison of CR and protein-coding gene phylogenetic signal, as described below.

The number of heteroplasmies in the sperm whale mitogenomes was assessed by reviewing all positions where the minority base represented at least 30% of reads at a site. In addition, the Illumina mitogenomes were required to have a minimum coverage of two reads in each direction for each alternate base (Li et al. 2010), and the 454 mitogenomes had to satisfy the *454HCDiffs* criteria (Roche, USA).

4.3.8. Interspecies alignment of cetacean mitogenomes

Reference sequence mitogenomes of the 33 cetacean species available from GenBank were downloaded in order to complement the dolphin species and sperm whales sequenced in this study, leading to a total of 42 cetacean species available for phylogenetic analyses (Appendix IV, Table IV.1). Interspecies alignments of the GenBank sequences, the dolphin sequences generated by Carraher (2004), and the sperm whale sequences generated in this study were constructed using *MUSCLE v3.6* (Edgar 2004) and checked by eye. *ND6* was excluded from phylogenetic analyses for the same reasons given in Ho and Lanfear (2010): distinctive patterns of evolution and location on the light strand opposite all other mtDNA protein-coding genes. Where the start and end

of protein-coding genes overlapped, the overlapping portions were excluded. The *MEGA v5.0.5* data viewer was then used to examine protein-coding regions of mitogenomes generated in this study for premature stop codons indicative of numts.

4.3.9. Bayesian phylogenetic estimation of substitution rates and TMRCA

Species-specific substitution rates were estimated using Bayesian phylogenetic reconstructions conducted in *BEAST v1.6.2* (Drummond and Rambaut 2007) with a GTR + I + Γ model of nucleotide substitution and estimated base frequencies. Two independent runs of 90,000,000 states, sampling every 3,000 states, were completed for each analysis, using an uncorrelated lognormal relaxed clock (Drummond et al. 2006) and a linked Yule tree prior across the CR and protein-coding region partitions. The rate analyses were calibrated using the fossil calibration dates given by Ho and Lanfear (2010). Unlike Ho and Lanfear (2010), a lognormal distribution, with the means represented in real space and a standard deviation of one million years, was used as the prior shape for all calibration points with the exception of the root (note that this contrasts with use of minimum-age priors e.g.(Ho and Lanfear 2010)). The root node was calibrated with a uniform prior of 34 to 46 million years as in Ho and Lanfear (2010). Independent runs were checked for effective sample sizes (ESS) of more than 200 and convergence of posterior values in *Tracer v1.5* (Rambaut and Drummond 2007), and convergence of tree topologies in *FigTree v1.3.1* (Rambaut 2009) before combining.

Phylogenetic relationships for 42 cetacean species (including the sperm whale) were reconstructed and lineage-specific substitution rates were estimated for the two separate

partitions of the CR and codon-partitioned protein-coding genes using *TreeAnnotator v1.6.2* visualized in *FigTree*. The sperm whale substitution rates were then compared to the 41 other cetacean species included in this study. Each species was represented once, with the exception of the bottlenose dolphin (*Tursiops truncatus),* where a previous mitogenome was available from GenBank, and the Hector's dolphin (*Cephalorhynchus hectori*), where the two subspecies *C. h. hectori* and *C. h. maui* were both included. The sperm whale was represented by a randomly selected mitogenome sequenced in this study, PmaNZ005. The median values for point estimates are reported in this paper with associated 95% highest probability density (HPD) intervals. Copies of the *.xml files used in all BEAST analyses, containing the priors defined, are given in Appendix IV, Figure IV.2.

The TMRCA of sperm whale mitogenomes was estimated using sequences of mtDNA protein-coding regions in *BEAST*. The sperm whale-specific substitution rate for the protein-coding regions calculated in the analysis described above was used, along with a skyline tree prior as this model is independent of the demographic history of the population (Drummond et al. 2005). The intraspecific protein-coding tree constructed in this analysis was also used to further investigate CR-specific constraints. The 400 bp mtDNA CR haplotype for each sperm whale mitogenome was mapped on to the intraspecific protein-coding tree to visually assess whether the mtDNA CR was reflective of underlying mitogenomic patterns of evolutionary relatedness.

4.4. Results

4.4.1. Summary of NGS and Sanger sequencing coverage

Over 93 Mbp of NGS sequence data was generated for the 20 sperm whale mitogenomes representing 17 individuals. For the thirteen mitogenomes sequenced with Illumina, this provided an average sequencing depth of 359×. For the seven mitogenomes sequenced with 454, this provided an average sequencing depth of 174×. Average mapping quality exceeded 36 *(BWA*: PHRED quality) for sperm whale mitogenomes sequenced with Illumina, and exceeded 63.5 (*GS Reference Mapper* 454 quality) for samples sequenced with 454. An additional 43 kbp of Sanger sequence was used to validate variable sites in the multiple sperm whale alignment and to estimate sequencing error of the NGS platforms. As described in Carraher (2004), the CR and protein-coding genes of 11 dolphin mitogenomes were successfully amplified and sequenced to an average depth of 2.64× by conventional Sanger methods.

4.4.2. Discrepancy rates in NGS sequencing

Mitogenomes of the three sperm whales sequenced on both the Illumina and 454 platforms showed good agreement. After independent assembly, the alignment of the 3 pairs revealed only 41 inconsistencies in the total comparison of 98,568 bp: this included 34 inconsistencies in homopolymers of ≥5 bp in length, a known source of error in 454 sequencing. Of the remaining seven inconsistencies, three were resolved when duplicate Illumina reads were removed; these presumably reflected PCR re-amplification of a minority nucleotide for a low-frequency heteroplasmy, or PCR-based polymerase error. Three additional inconsistencies occurred between 454 and Illumina in the CR of

PmaNZ034. At all three positions, the 454 sequence showed mixed-bases (between 50 - 59%) in comparison with the Illumina sequence, possibly indicative of a 454 library artifact. The final inconsistency in 454 sequence relative to the Illumina sequence was also in PmaNZ034, located adjacent to a 20 bp homopolymer. After excluding homopolymer inconsistencies, which were verified using Sanger sequencing (see below), the 454-Illumina discrepancy rate was seven in 98,568 bp, or 0.0071%.

Of the 43 kbp of Sanger sequence, 28,620 bp was available for comparison to the thirteen Illumina-sequenced sperm whale mitogenomes, and 21,971 bp was available for the seven 454-sequenced mitogenomes. Comparisons of Illumina and Sanger sequencing revealed two inconsistencies: site 23 in PmaNZ058, and site 12,982 in PmaNZ005 (no longer inconsistent when duplicate Illumina reads were removed), giving a total Illumina-Sanger discrepancy rate of two inconsistencies in 57,240 bp compared or 0.0034%. Comparisons of 454 and Sanger found eight inconsistencies, all of which were associated with homopolymers. After excluding homopolymer inconsistencies that were verified using Sanger sequencing (see below), the 454-Sanger discrepancy rate was <0.0023%, i.e. no errors detected in 43,942 bp compared.

4.4.3. QC and validation of variable sites

After alignment of the 17 mitogenomes, additional QC measures were applied to variable sites, including verification of low coverage/quality sequence with Sanger sequencing, removal of duplicate Illumina reads and testing for strand bias. After application of a Bonferroni correction (by sample), no significant strand bias was detected at variable

sites in Illumina samples. Over 354 instances of high NGS QC for variable sites were covered by Sanger sequencing, showing complete agreement. In contrast, 54 instances of low NGS QC were found over 92 putative variable sites. Sanger sequencing confirmed 44 of these 54 to be identical to the nucleotide called by NGS. Two sites in Illumina sequences (site 23 in PmaNZ058 and site 12,982 in PmaNZ005: as described above) were corrected based on Sanger sequencing/removing Illumina duplicate reads (assuming that Sanger sequencing was correct). An additional eight sites in 454 sequences, all at homopolymer sites (described above), were corrected based on Sanger sequencing. There were also two additional lengths of homopolymer that could not be resolved using Sanger sequencing for the 454-generated mitogenomes included in this study: a C_{12-21} homopolymer at position 1,129 (located in the 16S rRNA) from four mitogenomes (PmaNZ013, PmaNZ076, PmaNZ082, and PmaOR001) and a C_{9-10} homopolymer at position 16,270 (located in the CR) from three mitogenomes (PmaNZ013, PmaNZ076 and PmaNZ082). Due to the uncertainty of the true length of these homopolymer regions, potential differences in homopolymer length were masked at these positions. A summary of sites that differed between sequencing platforms in the error-checking and validation of variable sites sections of this paper are available in Appendix IV, Table IV.5. Average NGS sequencing depth on validated variable sites is given for each sperm whale in Appendix IV, Figure IV.3.

We found no evidence of numts: all sperm whale mitogenomes assembled with no gaps in coverage, as did the protein-coding region fragments for the dolphin mitogenomes sequenced in this study (Carraher 2004). There were no abrupt or discordant regions of

high divergence among the sperm whale mitogenomes which would indicate inadvertent assembly of numt sequences (maximum pairwise divergence observed was 1.65% for tRNA-Lys: Appendix IV, Table IV.6). No premature stop codons in protein-coding regions were observed for any mitogenome (sperm whale or dolphin) generated for this study.

4.4.4. Mitogenome haplotypes

After completion of QC review, the multiple alignment of 17 sperm whale mitogenomes showed only 82 variable sites, 76 of which were transitions, and 6 of which were transversions. No sites showed both transitions and transversions, and no indels were required in the alignment after masking the small number of homopolymers that could not be successfully Sanger sequenced (Table 4.3). Despite the low number of variable sites, all sperm whale mitogenomes in this study were resolved as having unique haplotypes, even those with identical CRs (Figure 4.1). Pairwise differences between mitogenomes ranged from 2 to 33. The largest number of singleton substitutions (sites differing in only one individual) found among the 17 sperm whale mitogenomes sequenced in this study was 11 (Appendix IV, Table IV.7). By comparison, a large number of singleton substitutions (n = 27) were found in the only sperm whale mitogenome available from NCBI, NC_002503 (Arnason et al. 2000). Given the potential for cloning or sequencing errors relative to mitogenomes sequenced in this study, NC_002503 was excluded from further intraspecific diversity analyses in this paper.

4.4.5. Sperm whale mitogenome diversity by gene and codon position

The overall mitogenome diversity among the 17 sperm whales sequenced in this study was low ($\pi = 0.096\%$). In contrast to expectations under the CR-specific constraint hypothesis, the CR (previously noted for its low diversity in comparison with other species: Table 4.1) actually showed the highest diversity of any mitogenomic partition ($\pi = 0.35\%$). CR nucleotide diversity was more than 1.5× greater than that of the combined 3rd codon protein-coding positions ($\pi = 0.22\%$), despite the 3rd codon position and 3rd codon diversity significantly exceeding all other remaining mitogenomic partitions (as determined by non-overlapping 95% confidence intervals, Figure 4.2). As the 95% confidence intervals of the CR and 3rd codon position partitions overlap, we could not conclude that the CR has significantly higher diversity than the 3rd codon position. The percentage of variable sites in the CR (1.57%, or 15 out of 954 sites), exceeded that of the 3rd codon (including *ND6*; 1.18%, or 45 out of 3,803 sites), although again not significantly based on a Fisher's exact test ($p = 0.3315$). In a gene by gene comparison, only 3 out of the 13 protein-coding genes (*COX2, ND3, ND4L*) had 3rd codon diversity that exceeded CR diversity, and these differences were not statistically significant based on overlapping 95% confidence intervals (Appendix IV, Table IV.8). We note that the estimates of nucleotide diversity presented here are likely to be biased upwards because of the non-random selection of samples based on CR haplotypes. Nucleotide diversity for a random sample of sperm whales is likely to be considerably lower as even distantly related individuals could inherit identical mitogenomes.

Based on our criteria of at least 30% of reads representing the secondary base at a site, we detected between 1-6 heteroplasmies in three mitogenomes, located in the 12srRNA, 16srRNA; protein-coding genes, and the CR. Sanger sequencing was available for one of these sites which confirmed the heteroplasmy present (Appendix IV, Table IV.9). Heteroplasmies occurring in the protein-coding regions were either synonymous or required only the substitution of an amino acid with similar properties (Adachi and Hasegawa 1996, Gilis et al. 2001). The other fourteen mitogenomes showed no heteroplasmies that satisfied the detection criteria. Two of these mitogenomes were sequenced with both Illumina and 454 and showed no evidence of heteroplasmies with either NGS platform (Appendix IV, Table IV.9). These overall rates of heteroplasmy are consistent with studies on humans (Sosa et al. 2012) and other cetacean species (Vollmer et al. 2011).

4.4.6. Sperm whale mtDNA protein-coding genes and CR phylogenetic signal

As a further test of the CR-specific constraint hypothesis, we compared the phylogenetic signal of the sperm whale CR and protein-coding regions. From the hypothesis, we expected there would be a lack of congruence between the CR and protein-coding regions due to homoplasy in the CR. In fact, the phylogenetic signal from the CR and protein-coding regions was found to be congruent with a significant correlation between pairwise differences over the 400 bp CR haplotype and pairwise differences over the rest of the mitogenome (Pearson's product-moment correlation = 0.54, $p < 0.001$). CR haplotypes were also generally congruent with phylogenetic reconstructions based on protein-coding regions within the sperm whale, albeit with less resolution in the CR haplotypes (Figure

4.1). When either protein-coding or CR characters were traced on to the phylogenetic tree, only a small number of characters showed evidence of homoplasy, reflected by the high consistency (Kluge and Farris 1969) and retention indices (Farris 1989) when either partition was mapped on to the protein-coding tree (Figure 4.1). Finally, the ILD test detected no phylogenetic incongruence between the CR and protein-coding regions (p = 1.000).

4.4.7. Patterns of purifying selection in the sperm whale mitogenome

The sperm whale mitogenomes showed the expected pattern of purifying selection reported in other mammalian mitogenomes (Stewart et al. 2008). Of the 61 variable sites occurring in protein-coding regions, 46 were synonymous, and 15 were non-synonymous, with twice as many replacement substitutions found in 1st versus 2nd codon positions (10 vs. 5; Table 4.3). Over the combined protein-coding genes, the dN/dS ratio was 0.25; significantly lower than the value expected under neutral evolution (Fisher's exact test, p < 0.001). The transitional bias observed in the coding regions (57 transitions, 4 transversions) was similar to that of both the CR and RNA partitions (Table 4.3). The 15 amino acid substitutions present in this alignment (substitutions shown in Appendix IV, Table IV.7) generally had a low cost of replacement based on several substitution-cost matrices (Adachi and Hasegawa 1996, Gilis et al. 2001).

4.4.8. Cetacean Bayesian phylogenetic reconstructions and substitution rates

The two interspecific *BEAST* runs for estimating substitution rates showed ESS values over 200 for all parameter estimates, and convergence for both parameters and tree

topologies. Given this agreement, runs were combined, which gave ESS values over 400 for all parameter estimates. The phylogenetic tree obtained was well supported, with 38/43 clades supported by posterior probabilities exceeding 95% (Figure 4.3). Estimates of the age of the Mysticeti clade were younger in our analyses (95% HPD: 11.96-17.86 mya) than in other reconstructions (Dornburg et al. 2012, Ho and Lanfear 2010, Jackson et al. 2009), likely due to the differences in prior shape for date calibration at this node. In addition, both our phylogeny and previous studies showed uncertain relationships within Delphinidae, with the main difference being positioning of the white-beaked dolphin (*L. albirostris*) between reconstructions (Dornburg et al. 2012, Ho and Lanfear 2010, Vilstrup et al. 2011). Apart from these differences, our tree showed very similar relationships in comparison with previous publications utilizing mitogenomes (Figure 4.3) and the two samples of the two species with multiple representatives in the phylogeny (Hector's dolphin: *C. hectori*; bottlenose dolphin: *T. truncatus*) grouped together with high confidence. As well as phylogenetic concordance, overall cetacean substitution rates obtained from this analysis were broadly comparable with those obtained from previous studies (Table 4.4). Over all cetaceans, substitution rate estimates were 1.12%/mya or 1.12×10^{-2} substitutions/site/million years (95% HPD: 0.99 to 1.26%/mya) for the CR, and 1.08%/mya (95% HPD: 0.99 to 1.18%/mya) for the coding region. First codon positions evolved at 0.45× this average protein-coding rate (95% HPD: 0.42 to 0.48), 2^{nd} codon positions at 0.12× (95% HPD: 0.11 to 0.14) and 3^{rd} codon positions at 2.4× this rate (95% HPD: 2.39 to 2.46).

To test the hypothesis of a slow substitution rate in the sperm whale, estimates of substitution rates for the sperm whale obtained from age-calibrated Bayesian phylogenetic reconstructions in *BEAST* were compared to estimates from the other 41 cetacean species included in the analysis. Under the hypothesis of reduced substitution rates in the sperm whale, the expectation was slow rates in the sperm whale relative to other cetaceans. All species-specific substitution rates fell within 0.5%/mya of each other in the protein-coding regions and CR, with the exception of the sperm whale (an outlier with a fast CR substitution rate) and the Franciscana (*Pontoporia blainvillei*; an outlier with fast CR and protein-coding region rates) (Figure 4.4). Sperm whales ranked 25th out of 42 in terms of protein-coding region rates (ranking from fast to slow), implying an average (but not slow) substitution rate for this region relative to other cetaceans. Surprisingly, sperm whales had the fastest rate of substitutions in the CR for any of the 42 cetaceans included in the analysis (2.60%/mya; Figure 4.4).

4.4.9. Bayesian estimates of sperm whale mitogenome TMRCA

In addition to investigating the hypotheses of CR-specific constraints and rate variation in the sperm whale mitogenome, the TMRCA of the protein-coding mtDNA genes of the sperm whale was estimated to investigate the potential for a population bottleneck or selective sweep. Assuming clock-like intraspecific substitution rates within the protein-coding regions, and using the sperm-whale protein-coding substitution rate estimated from the interspecies phylogeny, the TMRCA for the combined sperm whale mtDNA protein-coding regions was estimated as 103,000 years ago (95% HPD: 72,800 to 137,400 years ago).

4.5. Discussion

This study represents the first population level survey of mitogenome diversity in the sperm whale. We found that the low diversity previously characterized for the sperm whale CR (Lyrholm and Gyllensten 1998, Whitehead 1998) is a feature of the entire sperm whale mitogenome. Furthermore, intraspecific phylogenetic signals from the protein-coding region and CR were congruent. Overall, these data indicate that evolution of the CR has not been significantly constrained compared with the rest of the sperm whale mitogenome. While substitution rates of sperm whale mtDNA protein-coding regions were similar to rates in other cetacean species, the sperm whale CR substitution rate was the fastest of all 42 cetacean species included in this analysis. This suggests that slow sperm whale-specific substitution rates are not responsible for low CR or overall mitogenomic diversity.

4.5.1. Sources of sequencing error

Studies of population-level diversity can be sensitive to NGS sequencing errors or lack of quality control enforced on variable sites (Shen et al. 2010), and there are few standardly reported quality control measures in NGS sequences (Goto et al. 2011a). Given the previously reported low levels of CR diversity in the sperm whale (Lyrholm and Gyllensten 1998), it was important to ensure that sequencing errors did not impact our mitogenomic diversity estimates. We carried out a three-way comparison of NGS and Sanger sequencing technologies and found relatively low discrepancy rates. The highest discrepancy rate (454 to Illumina: 0.0071%) was an order of magnitude below the overall estimate of diversity obtained for the sperm whale mitogenome ($\pi = 0.096\%$). We

compensated for the small number of discrepancies, including apparent PCR duplicates; polymerase, library and assembly artifacts; and differences in homopolymer lengths (a known artifact of 454 technology (Kircher and Kelso 2010)) by Sanger sequencing variable homopolymers and low QC NGS sites. The complete agreement between high QC NGS and Sanger sequence validates the QC threshold used in this study.

Another source of potential error in mtDNA sequencing is the inadvertent inclusion of numts (Dunshea et al. 2008, Sorenson and Quinn 1998, Zhang and Hewitt 1996). As a precaution against this, we used LR-PCR with fragment sizes that exceeded the typical length of mammalian numts, as well as high annealing temperatures in order to reduce non-specific amplification (Li et al. 2012, Richly and Leister 2004). Assemblies showed no evidence of gaps in coverage that could be indicative of linear numts incorporating into the alignment (Sorenson and Quinn 1998, Thalmann et al. 2004), nor were there anomalously divergent sperm whale mitogenome regions or premature stop codons that would indicate the incorporation of ancient pseudogenes (Nabholz et al. 2010, Sorenson and Quinn 1998). Along with the low discrepancy rate, robust QC at variable sites, and lack of detectable strand bias, the absence of detectable numts provided confidence in the estimates of diversity obtained from our sperm whale mitogenome alignments.

4.5.2. Are CR-specific constraints operating on the sperm whale mitogenome?

Like sperm whales, killer whales and fishers also have very low levels of mitogenomic diversity relative to other mammal species for which estimates are available (Appendix IV, Table IV.10). In killer whales and fishers, the CR presents misleading

phylogeographic interpretations, as deep divergences among non-CR regions of the mitogenome are not accurately represented by the CR (Knaus et al. 2011, Morin et al. 2010a). Instead, the CR seems to have become saturated with substitutions due to constraints, thus making it less accurate at predicting intraspecific relationships than the full mitogenome (Knaus et al. 2011). In sperm whales, this is not the case. Not only does the CR accurately reflect the intraspecific phylogeny shown by the protein-coding genes, but the CR (previously noted for its low diversity compared with other cetacean species (Lyrholm and Gyllensten 1998, Whitehead 1998)) actually showed the greatest nucleotide diversity among the partitions compared. Only a modest increase in diversity was recovered by sequencing the full mitogenome in comparison with the CR (10 haplotypes distinguished from 400 bp of CR; 14 haplotypes distinguished from 954 bp of CR; 17 haplotypes from the full 16,428 bp mitogenome). This further indicates that the CR does not have a large number of sites placed under mutational constraints, leading to saturation, in comparison with the protein-coding regions in the sperm whale.

4.5.3. Are substitution rates slow in the sperm whale mitogenome?

While tests of substitution rates in the sperm whale mitogenome have been conducted previously (Lyrholm et al. 1996, Whitehead 1998), these tests were limited to only a fraction of the mitogenome. They were also conducted assuming Kimura's (1980) nucleotide substitution model and no variation in substitution rates among sites. This might not accurately model evolution in cetacean mitogenomes. The use of the domestic cow (*Bos taurus*), a relatively distant outgroup, could also have limited the power to detect significant rate variation (Bromham et al. 2000). We investigated variation in

substitution rates across the CR and all protein-coding genes (with the exception of *ND6*); and used *BEAST*, as this allowed our analyses to accommodate differences in base composition, substitution biases, and rate heterogeneity (Drummond et al. 2006). We also calibrated our rate analyses with cetacean fossil calibration points (in comparison to the domestic cow outgroup used in previous analyses). We found no evidence of a pervasive pattern of slow substitution rates in the sperm whale mitogenome relative to other cetaceans. This supports previous studies that found no evidence for sperm whale-specific slowing of substitution rates in the limited number of genes examined (Lyrholm et al. 1996, Whitehead 1998). In fact, our study indicated that the sperm whale CR had a significantly elevated rate of evolution, a finding that is also consistent with previous studies utilizing different relative rate tests (Lyrholm et al. 1996). These results suggest that a slow substitution rate is not responsible for the low mitogenome diversity seen in the sperm whale.

Estimates of cetacean substitution rates in this study were also broadly comparable with previous studies of cetaceans (Table 4.4), including supporting previous findings of slow rates in cetaceans compared with other mammals such as primates and rodents (Jackson et al. 2009, Martin and Palumbi 1993, Nabholz et al. 2008). The correlation between CR and coding region rates among cetaceans was weak and not statistically significant (Pearson's product-moment correlation = 0.155, $p = 0.326$) probably reflecting substitutional saturation of the CR between species, or different selective constraints on specific genes, across cetaceans.

4.5.4. Can a genetic bottleneck or selective sweep explain low sperm whale mtDNA diversity?

The Bayesian analyses employed in this study estimated a TMRCA for the sperm whale mitogenome of 103,000 years ago, consistent with previous research based on the CR (Lyrholm et al. 1996). It is likely that the actual age of the TMRCA is younger than this estimate due to the use of external calibration points to estimate the substitution rate within the sperm whale (Ho et al. 2008). This would most likely put the age of the TMRCA in the Pleistocene, a period characterized for its glaciation (Steeman et al. 2009). We consider this TMRCA to be younger than expected considering the age of the sperm whale lineage (divergence between the sperm whale and pygmy sperm whale 95% HPD: 17.6-27.9 mya) and the sperm whale's abundance and worldwide distribution (Best 1979, Whitehead 2002; 2003). Further evidence for a relatively recent TMRCA is the low number of transversions (Ts/Tv ratio = 12.5) and complete lack of Sanger/Illumina-confirmed indels in the sperm whale mitogenomic sequence. The Ts/Tv ratio is far higher than seen in mtDNA (mostly based on CR and *CytB*) from other cetacean species such as the common minke whale (*Balaenoptera acutorostrata;* Ts/Tv = 5; (Pastene et al. 2007)), the killer whale (Ts/Tv = 3.75; (Hoelzel et al. 2002)), and even the Hector's dolphin (*C. hectori*) (Ts/Tv = 3.3; (Pichler et al. 1998)), a species that is known to have undergone a population reduction.

Our estimate of the TMRCA is consistent with a population bottleneck or selective sweep having acted on the sperm whale in the past, but is not consistent with the impact of commercial whaling that exploited sperm whales from the 18th century onwards (Best

1979, Whitehead 2003). A whaling-induced sperm whale population bottleneck has also previously been discounted due to estimates of current sperm whale abundance (Whitehead 1998). However, sperm whales occupy a very specialized niche (Best 1979, Whitehead 2003), and past climatic influences, particularly glaciation during the Pleistocene, may have altered prey distribution (Steeman et al. 2009), or otherwise impacted on the abundance of the sperm whale.

Selective sweeps could also have reduced sperm whale mitogenomic diversity through the previously proposed hypothesis of 'hitchhiking' with maternal cultural innovation (Whitehead 1998). In addition, sperm whales make routine foraging dives in excess of 1,000 m (Watkins et al. 1993). Unique physiological adaptations in the sperm whale appear to have resulted in response to the selective pressure of deep diving including increases in muscle myoglobin levels; greater utilization of blood and muscle versus lung storage of oxygen; and collapsing of lungs at shallow depths (Kooyman and Ponganis 1998). Additional adaptations might include positive selection on mtDNA-encoded proteins due to the mitochondria's role in oxidative phosphorylation (Ballard and Dean 2001). Mutations with novel adaptive properties could lead to a selective sweep of mitogenomes. However, in order for a selective sweep to globally reduce mtDNA diversity, there must be inter-ocean dispersal of female sperm whales. Despite female social units being relatively philopatric, they are known to carry out large-scale dispersal events in apparent response to poor ocean conditions (e.g. Galapagos to Ecuador/Gulf of California) over short time scales of less than a decade (Whitehead et al. 1997). These documented movements and the shared mtDNA haplotypes between ocean basins

confirms at least occasional inter-ocean dispersal in female sperm whales, providing the potential for a selective sweep to reduce mtDNA diversity on a global scale.

Although a population bottleneck or selective sweep are both plausible causes of the low mitogenomic diversity, distinguishing between these hypotheses will require comparative population-level nuclear DNA datasets (which are not yet available for the sperm whale or its sister taxon, the pygmy sperm whale). A population bottleneck would be expected to reduce mtDNA and nuclear diversity in the sperm whale relative to the pygmy sperm whale, while a selective sweep should reduce only mtDNA diversity in the sperm whale (Rokas et al. 2001). Within the sperm whale, a population bottleneck would be expected to reduce the time to coalescence for both mtDNA and nuclear DNA, whereas a selective sweep should only reduce the time to mtDNA coalescence (Charlesworth et al. 2003), although these inferences would depend heavily on an accurate assessment of mutation rates (Karl et al. 2012).

4.5.5. Conclusions

The low diversity of sperm whale mitogenomes found in this study, and consistency of the CR with protein-coding intraspecific reconstructions, refutes CR-specific constraints as an explanation for low CR diversity. Furthermore the pervasive low mitogenome diversity requires a hypothesis that can explain the low mitogenome-wide variation. In this study we eliminated slow substitution rates in the sperm whale mitogenome as an explanation for low diversity. The lack of "rare" substitutions such as indels and transversions, and the recent TMRCA for sperm whale mitogenomes suggests that the

previously proposed hypotheses of selective sweeps or population bottlenecks are the most likely candidates for explaining the low mitogenome diversity seen in the sperm whale.

4.6. Acknowledgements

This work was supported by a Mamie Markham Award and a Lylian Brucefield Reynolds Award from the Hatfield Marine Science Center; and a 2008-2011 International Fulbright Science & Technology award to A.A; and co-funded by the ASSURE program of the Department of Defense in partnership with the National Science Foundation REU Site program to K.H and C.S.B. [grant number NSF OCE-1004947]. Publication of this paper was supported, in part, by the Thomas G. Scott Publication Fund. The authors would like to thank R. Constantine and K. Thompson for curation of the CeTA database, and New Zealand Department of Conservation staff for collecting NZ samples used in this study; J. Rice for curation of the OMMSN database that supplied the Oregon sample; M. Vant for primer development; D. Jacobsen for assistance with 454 prep; K. Jones for assistance with 454 analysis, the National Oceanic and Atmospheric Administration Southwest Fisheries Science Center and A. Liston for consultation on cetacean mitogenomics; M. Dasenko and C. Sullivan (Oregon State University Center for Genomics Research and Biocomputing) for Illumina sequencing and computational support; S. Wu and WLS. Li for consultation on Bayesian analysis, Oregon State University Cetacean Conservation and Genomics Laboratory for additional lab support; and A. Sremba, A. Hickey, E. Carroll, E. Harvey, P. Alexander, K. Makova and two anonymous referees for providing valuable comments on this manuscript.

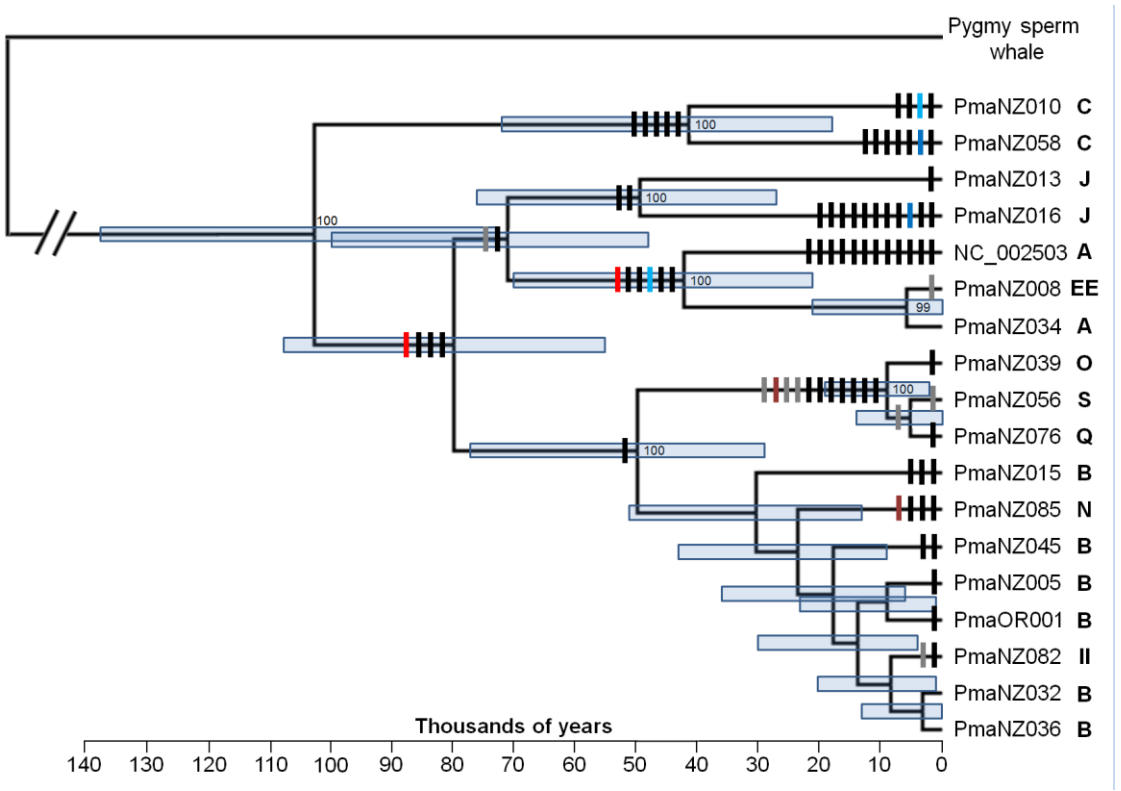


Figure 4.1: Bayesian phylogenetic reconstructions of the relationship between the 17 sperm whale mitogenomes in this study based on protein-coding genes (10,815 bp), including the NCBI GenBank sample, NC_002503 (Arnason et al. 2000). Terminal taxa are labeled by their individual sample codes and CR haplotypes (400 bp) are listed in bold after sample codes. Clades with more than 95% posterior probability are indicated on the tree, and 95% HPD intervals for node ages are shown by the horizontal blue boxes. Parsimony-inferred character changes (using *Mesquite v2.75* (Maddison and Maddison 2011)) in the protein-coding regions are shown by black cross bars. Inferred character changes in the 400 bp region of the CR used to define haplotypes are shown in gray. Homoplastic characters are shown in shades of blue for the protein-coding region (two sites requiring two changes each), and in shades of red for the CR (also two sites requiring two changes each). Consistency Indices: 0.9722 for protein-coding regions and 0.8333 for CR. Retention Indices: 0.9565 for protein-coding regions and 0.8333 for CR.

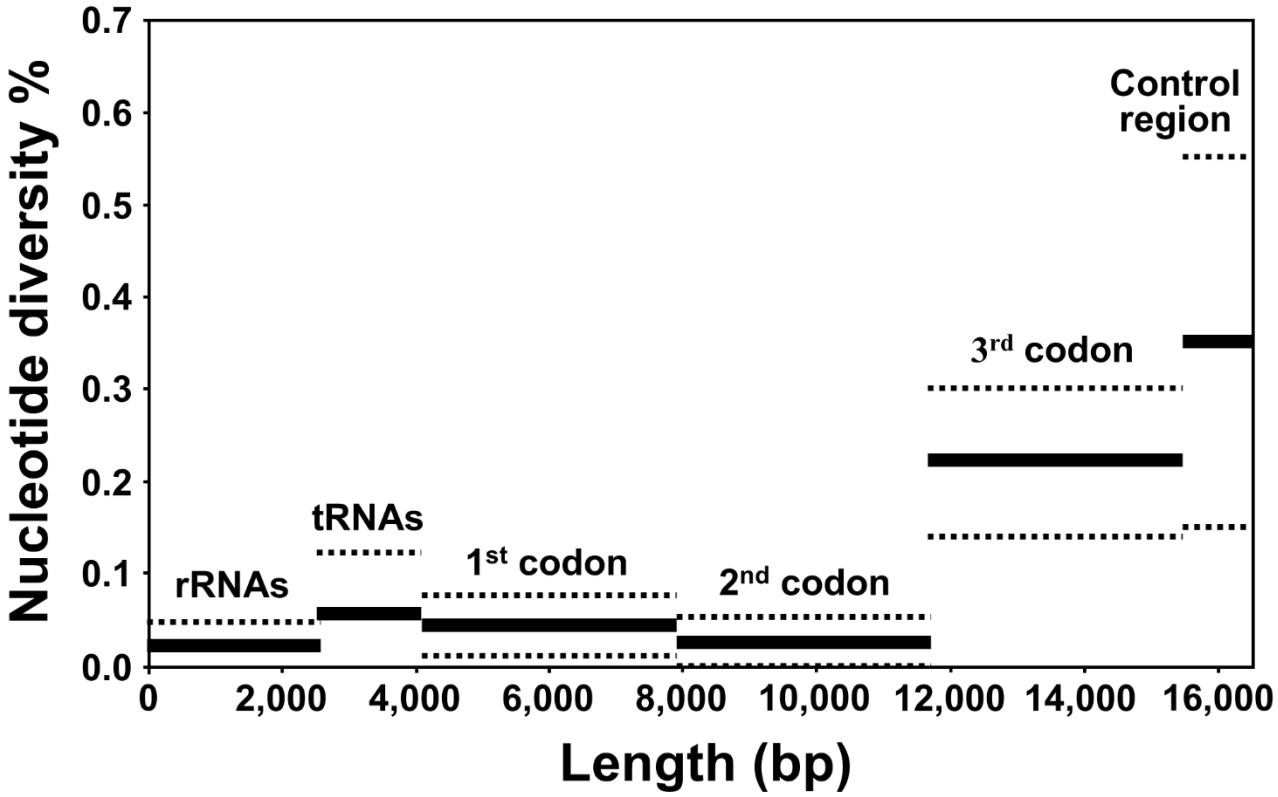
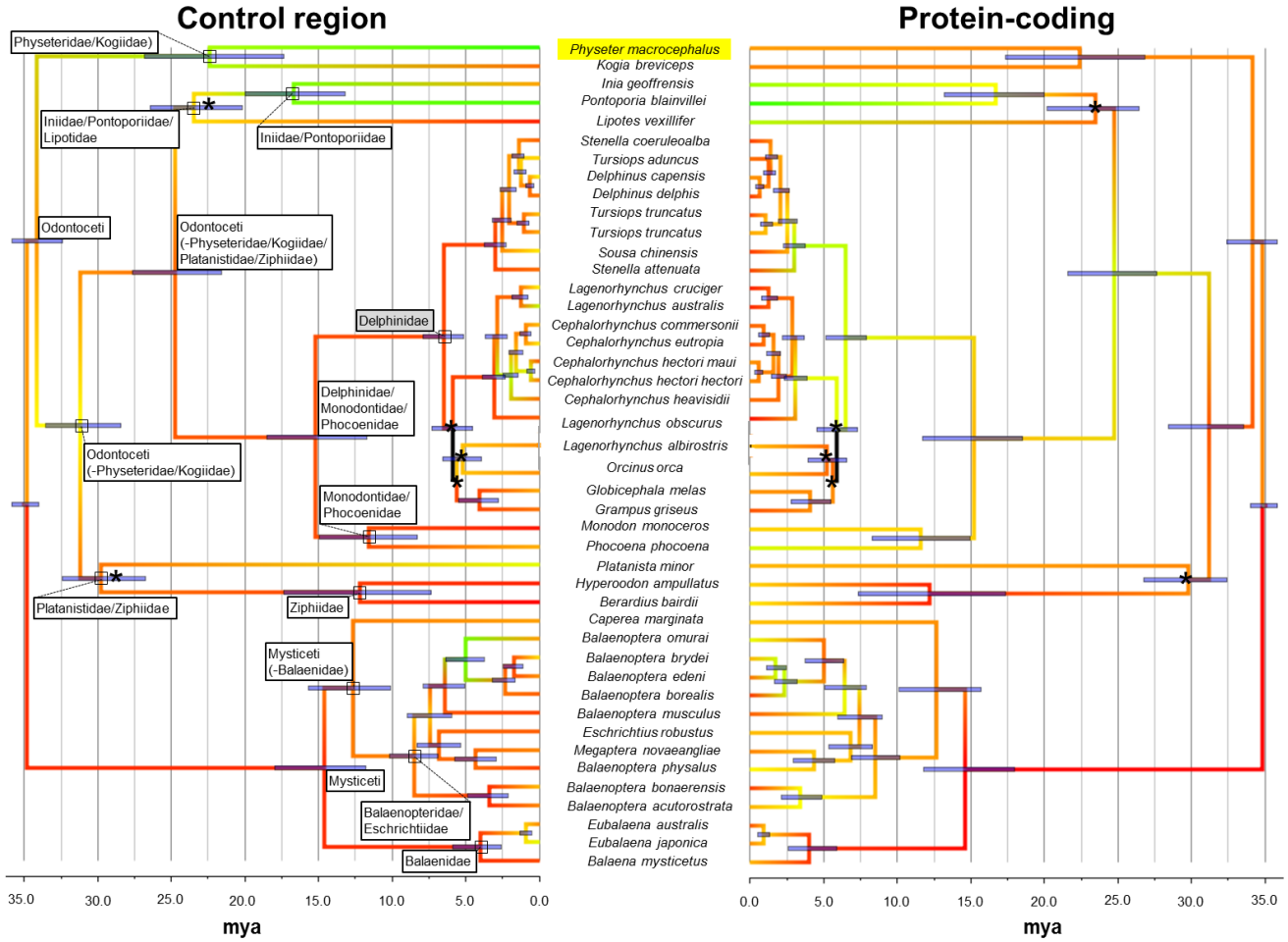


Figure 4.2: Nucleotide diversity (%) for the concatenated rRNA regions; concatenated tRNA regions; 1st, 2nd and 3rd codon positions of the protein-coding regions, and the CR of the sperm whale mitogenome (*n* = 17). 95% confidence limits (constructed with bootstrap standard error computation) are shown by the dashed lines. For the rRNAs and tRNAs, lower 95% confidence limits overlap zero.

Figure 4.3: Estimates of substitution rates for the CR and protein-coding regions using a Bayesian tree reconstruction. Tree priors for these partitions were linked, so the topology is the same. The sperm whale taxon label is highlighted in yellow. Branches are colored by rate, ranging from slow in red (0.41%/mya for the CR, 0.59%/mya for the protein-coding regions), to fast in green (2.60%/mya for the CR, 2.12%/mya for the protein-coding regions). 95% higher posterior distributions for the age of nodes are indicated by the blue bars. All nodes had greater than 95% posterior probability support with the exception of the clades indicated by asterisks. With the exception of some of the relationships within Delphinidae (clade label shaded in gray), all clades labeled on the tree were supported by the previous cetacean mitogenome studies of Ho and Lanfear (2010) and Dornburg et al. (2012). The Mysticeti and Balaenopteridae/Eschrichtiidae clades were also supported by a Mysticeti-only study, Jackson et al. (2009). A Delphinidae-only study, Vilstrup et al. (2011) also partially supported relationships in Delphinidae. Discrepancies within Delphinidae are mentioned in greater detail in the text. Clade nomenclature from Perrin (2009).

Figure 4.3 (Continued)



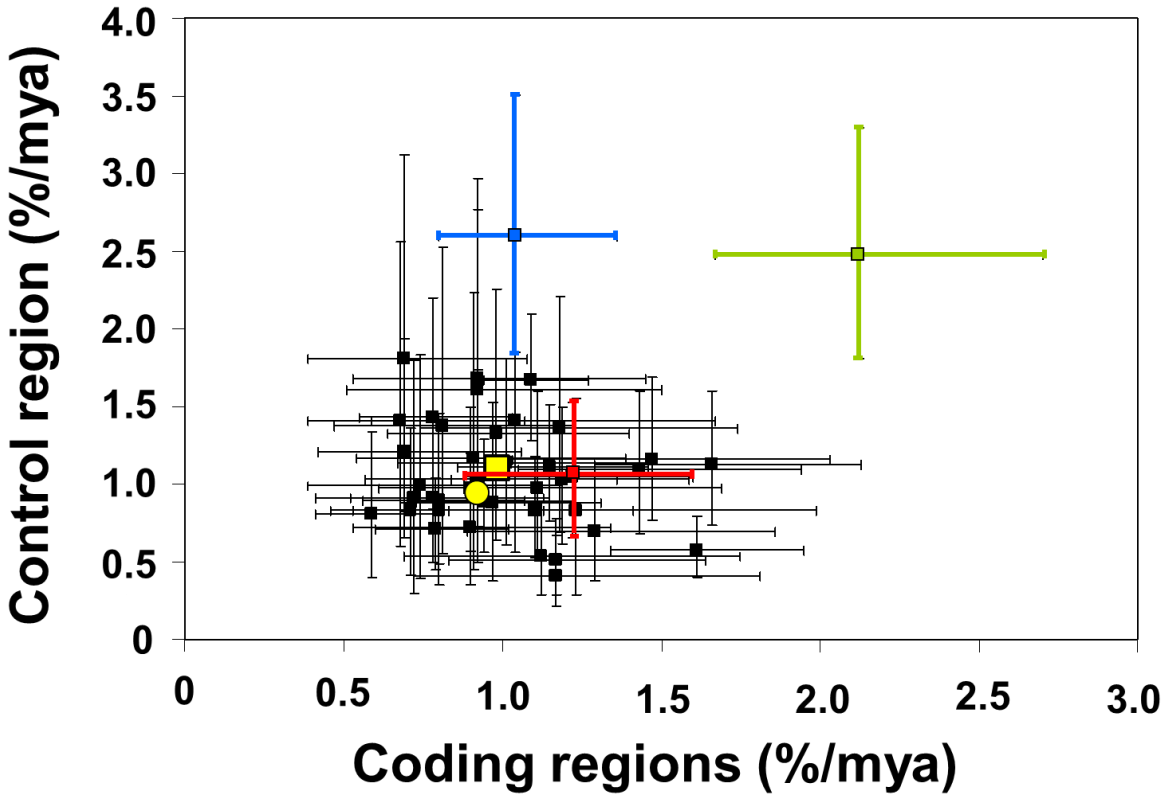


Figure 4.4: CR and protein-coding region lineage specific substitution rates for the 42 cetacean species included in this study. Error bars associated with points indicate the 95% HPD associated with each species. The sperm whale is shown in blue, the Franciscana (*Pontoporia blainvillei*) in green and the killer whale (*Orcinus orca*) in red. The median odontocete rates are shown in the yellow square, the median mysticete rates in the yellow circle.

Table 4.1: Estimates of nucleotide diversity and number of mtDNA CR haplotypes for 26 cetacean species (limited to studies with sample sizes above 100). For species with studies from multiple geographic regions, the studies with lowest and highest nucleotide diversity are presented. If multiple studies were available from the same geographic region, only the study with the lowest nucleotide diversity was retained). Species are ordered by nucleotide diversity from low to high (species with multiple estimates are ordered by greatest estimate of nucleotide diversity), with dark blue backgrounds indicating low nucleotide diversity/numbers of haplotypes, through to white indicating higher estimates of nucleotide diversity/numbers of haplotypes. The sperm whale is highlighted in yellow. Species that have worldwide ranges are indicated by an asterisk. References: [1] (Oremus et al. 2009); [2] (Chivers et al. 2007); [3] (Pimper et al. 2010); [4] (Li et al. 2011); [5] (Yang et al. 2008); [6] Studies listed in Table 4.2; [7] (O'Corry-Crowe et al. 1997); [8] (Hoelzel et al. 2002); [9] (Banguera-Hinestroza et al. 2010); [10] (Hamner et al. 2012); [11] (Pastene et al. 2007); [12] (Kanda et al. 2007); [13] (Hayano et al. 2003); [14] (Bérubé et al. 1998); [15] (Escorza-Treviño et al. 2005); [16] (Sremba et al. 2012); [17] (Rosel et al. 1999); [18] (Chivers et al. 2002); [19] (Carroll et al. 2011); [20] (Cassens et al. 2005); [21] (Chivers et al. 2005); [22] (Adams and Rosel 2006); [23] (Quérouil et al. 2010); [24] (Mirimin et al. 2009); [25] (Natoli et al. 2005); [26] (Martien et al. 2012); [27] (Krützen et al. 2004); [28] (Rosenbaum et al. 2009); [29] (Palsbøll et al. 1995).

Species	Sampling location	Sample size	CR haplotype length (bp)	π (%)	No. of haplotypes	Reference
Long-finned pilot whale* *(Globicephala melas)*	Pacific/Atlantic Ocean	643	620	0.35	13	*1*
False killer whale* *(Pseudorca crassidens)*	Worldwide	124	945	0.37	24	*2*
Commerson's dolphin (*Cephalorhynchus commersonii)*	S. America	196	466	0.40	20	*3*
Finless porpoise	W. Pacific	386	345	0.27	16	*4*
(Neophocaena phocaenoides)	China and Japan	417	345	0.44	18	*5*
Sperm whale* *(Physeter macrocephalus)*	Worldwide	1,167	399	0.51	28	*6*
Beluga *(Delphinapterus leucas)*	Arctic	324	410	0.51	29	*7*
Killer whale* *(Orcinus orca)*	Worldwide	102	995	0.52	13	*8*
White-beaked dolphin *(Lagenorhynchus albirostris)*	N. Atlantic	122	601	0.56	18	*9*
Hector's dolphin *(Cephalorhynchus hectori hectori)*	New Zealand	318	650	0.79	21	*10*
Short-finned pilot whale *(Globicephala macrorhynchus)*	S. Hemisphere and Atlantic	150	620	0.87	14	*1*
Common minke whale*	N. Atlantic	102	500	0.60	26	*11*
(*Balaenoptera acutorostrata)*	Western N. Pacific	127	500	1.00	34	*11*
Bryde's whale* *(Balaenoptera brydei)*	Pacific/Indian Ocean	472	299	1.00	51	*12*
Dall's porpoise *(Phocoenoides dalli)*	N. Pacific	103	479	1.06	49	*13*
Fin whale* *(Balaenoptera physalus)*	N. Atlantic	341	288	1.13	48	*14*
Pantropical spotted dolphin* *(Stenella attenuata)*	E. Pacific	225	455	1.36	112	*15*

Table 4.1 (Continued)

Species	Sampling location	Sample size	CR haplotype length (bp)	π (%)	No. of haplotypes	Reference
Antarctic blue whale *(Balaenoptera musculus intermedia)*	S. Ocean	183	410	1.40	52	*16*
Harbor porpoise	NE Atlantic	194	344	0.47	37	*17*
(Phocoena phocoena)	N. Pacific	249	402	1.40	88	*18*
Southern right whale *(Eubalaena australis)*	New Zealand	585	500	1.43	12	*19*
Antarctic minke whale *(Balaenoptera bonaerensis)*	S. Hemisphere	119	500	1.50	83	*11*
Dusky dolphin *(Lagenorhynchus obscurus)*	S. Hemisphere	153	591	1.63	62	*20*
Pygmy sperm whale* *(Kogia breviceps)*	Worldwide	108	406	1.65	74	*21*
Atlantic spotted dolphin	N. Atlantic	196	329	1.47	34	*22*
(Stenella frontalis)	Azores	144	611	1.80	76	*23*
Short-beaked common dolphin* *(Delphinus delphis)*	N. Atlantic	297	360	1.80	77	*24*
Bottlenose dolphin*	Eastern N. Atlantic	123	630	1.6	41	*25*
Tursiops truncatus	Hawaii	130	400	2.20	25	*26*
Western Australia bottlenose dolphin *(Tursiops sp.)*	W. Australia	220	351	2.21	8	*27*
Humpback whale* *(Megaptera novaeangliae)*	Mozambique and Eastern S. Africa	151	486	1.95	65	*28*
	N. Atlantic and Antarctic	136	288	2.60	31	*29*

Table 4.2: The diversity and oceanic distribution of published mtDNA CR haplotypes from the sperm whale, defined by DQ512921 (Haplotype A) and reflecting a consensus length of 400 bp. Haplotypes of mitogenomes sequenced in this study are shaded in gray. The numbers of samples with each haplotype were taken from previous studies by ocean basin: the North Atlantic - including the Gulf of Mexico and Mediterranean (Engelhaupt et al. 2009, Lyrholm and Gyllensten 1998, Ortega-Ortiz et al. 2012), South Indian/Pacific Oceans (Lyrholm and Gyllensten 1998) and Pacific Ocean (Lyrholm and Gyllensten 1998, Mesnick et al. 2011, Rendell et al. 2012, Richard et al. 1996a, Whitehead et al. 1998). Haplotype EE was first characterized in this study.

																									Ocean			
Haplotypes	58	62	105	107	121	150	184	200	207	208	211	235	243	260	272	273	286	287	288	305	308	319	324	350	N. ATL	S. IND/ PAC	PAC	Grand Total
A	T	C	C	A	C	C	T	T	A	A	C	A	G	A	A	C	A	A	A	C	A	G	C	C	66	12	258	336
B	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	55	12	187	254
C	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	115		117	232
D	.	.	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.			11	11
E	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		1	28	29
F	C	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.			6	6
G	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T			5	5
H	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	G	.	.	.	.	.	.			8	8
I	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.			10	10
J	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.		9	29	38
K	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	.	T	.			2	2
L	.	T	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	G	.	.	.	.	.			1	1
M	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.			7	7
N	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	3	2	13	18
O	.	T	.	.	.	.	C	.	.	.	.	.	.	G	.	.	.	.	.	.	.	A	T	.		2	12	14
P	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.			2	2
Q	.	T	.	.	.	.	C	.	.	.	.	.	.	G	G	.	.	.	.	.	.	A	T	.			1	1
R	.	T	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	A	T	.		1		1
S	.	T	T	.	.	.	C	.	.	.	.	.	.	G	G	.	.	.	.	.	.	A	T	.		1	5	6
T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.			3	3
U	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.			1	1
X	.	T	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	157			157
Y	.	T	.	.	.	.	.	.	G	.	.	.	.	.	G	.	.	.	G	.	.	.	.	.	19			19

Table 4.2 (Continued)

Haplotypes	58	62	105	107	121	150	184	200	207	208	211	235	243	260	272	273	286	287	288	305	308	319	324	350	N. ATL	S. IND/ PAC	PAC	Grand Total
Z	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	.	.			1	1
AA	C	.	.	.	.	T	.	.	.	.	.	G	.	.	.	.	.	.	.	T	.	.	.	.			2	2
BB	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2			2
EE	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.				
II	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.		1		1
																									417	41	709	1167

Table 4.3: Substitutions characterized in terms of their mitogenome region for the alignment of 17 sperm whales, including codon position for the protein-coding regions, type (Ts: transition; Tv: transversion) and whether they were non-synonymous (NS) or synonymous (S).

Mitogenome region	NS	S	Ts total	Tv total	Ts/Tv	Total
Control region		15	14	1	14	15
rRNA regions			1	1	0.5	2
tRNA regions			4			4
Total protein coding	15	46	57	4	14.25	61
Protein coding 1st codon position	10	1	11			11
Protein coding 2nd codon position	5		4	1		5
Protein coding 3rd codon position		45	42	3		45
Overall mitogenome			76	6	12.5	82

Table 4.4: Estimated substitution rates for mitogenomes of cetaceans. Rates are in % divergence/million years (%/mya), and ranges show 95% HPD intervals. References: (a) Hoelzel et al. (1991); (b) Duchêne et al. (2011); (c) Baker et al. (1993); (d) Rooney et al. (2001); (e) Jackson et al. (2009); (f) Ho & Lanfear (2010); (g) Vilstrup et al. (2011); (h) Dornburg et al (2012); (i) Nabholz et al. (2008) and (j) Alter et al. (2007).

Mitogenome partition	Taxa	Estimate from this study	Previous estimates
Control region	Cetaceans	0.99-1.26	0.28-1.04 [a]
	Odontocetes	1.09	0.23-0.76 [b]
	Mysticetes	0.99	0.70-1.00 [c], 1.20-3.70 [d], 1.67-9.32 [e]
Protein-coding region	Cetaceans	0.99-1.18	1.80-2.20 [f]
	Odontocetes	1.00	0.15-0.48 [b]*, 0.85-1.13 [g]^, 1.23-1.54 [h]
	Mysticetes	0.92	0.45-0.67 [h]
Protein-coding 1st codon position	Cetaceans	0.49	0.35 [f]
Protein-coding 2nd codon position	Cetaceans	0.14	0.11 [f]
Protein-coding 3rd codon position	Cetaceans	2.62	1.00 [i#], 2.40 [f]
	Mysticetes		0.08-2.46 [e], 0.40 [j#], 0.70-0.80 [i#]

* entire mitogenome used instead of protein-coding region

^ protein-coding genes plus tRNA genes used

estimates available for *CytB* only

5. A bottleneck for *Moby Dick*? Low mtDNA and nuclear diversity in the sperm whale

5.1. Abstract

A number of hypotheses have been proposed to explain the low mtDNA control region (CR) diversity of the sperm whale, including slow mutation rates, constraints restricting the sites free to vary in the CR, selective sweeps linked to maternal cultural innovations or functional substitutions in the protein-coding genes, lineage extinctions of social groups due to mass strandings, and historical population expansions or bottlenecks. Lineage extinction events and selective sweeps due to maternal cultural innovations were previously discounted as causes through genetic characterization of social groups and an examination of the sperm whale mass stranding record (Chapter 3). In addition, the slow mutation rate and constraints hypotheses were discounted through sequencing of mitogenomes (Chapter 4). However, to distinguish between a selective sweep of a functional substitution, a population bottleneck, and a population expansion, estimates of nuclear diversity are needed. Here, we sequenced and phased ~12,000 bp of nuclear DNA as well as sequencing mitogenomes for 22 sperm whales and 10 pygmy sperm whales (*Kogia breviceps).* The pygmy sperm whale was included as a control due to relatively high levels of mtDNA diversity, suggesting it has not been subject to a recent sweep, bottleneck, or expansion. The pygmy sperm whale is also a sister-species of the sperm whale, which could help control for phylogenetic constraints in patterns of evolution. We characterized nuclear and mitogenomic diversity of the two species, and reconstructed their demographic histories using extended Bayesian skyline plots. Under the selective

sweep hypothesis, we expected the sperm whale to show relatively high levels of nuclear diversity, with low diversity restricted to the mitogenome. Under the bottleneck and population expansion scenarios, both nuclear and mitogenomic diversity were expected to be low. We found significantly lower mtDNA diversity in the sperm whale vs. the pygmy sperm whale and other cetacean species based on comparisons including the proportion of variable sites (sperm whale CR 0.015 vs pygmy sperm whale 0.073, protein-coding 0.006 vs 0.020, chi-squared $p < 0.0167$). Sperm whales also had lower nuclear diversity than the pygmy sperm whale and other cetacean species (t-test, $df = 7$ based on 8 nuclear loci) based on observed heterozygosity (sperm whale 0.475 vs pygmy sperm whale 0.774, $p = 0.0115$), haplotype diversity (0.462 vs 0.832, $p = 0.0035$), number of alleles (6.5 vs 13, $p = 0.0082$), proportion of variable sites (0.006 vs 0.023, $p = 0.0167$), nucleotide diversity (0.072% vs 0.425%, $p = 0.0144$), and per locus theta (1.21 vs 5.63, $p = 0.0012$). Based on demographic reconstructions the sperm whale showed a markedly low, yet stable, effective population size of ~28,000 for the last two million years. Overall, these results are most consistent with a population expansion from a small long-term effective population size to the estimated pre-whaling population size of 1,100,000. The inferred recent timing of the population expansion (~20,000 years before present) is coincident with expansions in various squid species: primary prey of the sperm whale.

5.2. Introduction

The sperm whale (*Physeter macrocephalus*) is one of the most derived mammalian species: large in size, long-lived, extremely sexually dimorphic, and with a unique social

organization and capacity for deep foraging dives that few other marine mammals can match (Best 1979, Watkins et al. 1993, Watwood et al. 2006, Whitehead 2003). Despite a history of exploitation that removed hundreds of thousands of whales, the global sperm whale population remains relatively abundant (360,000 individuals) in comparison with other large whale species (Whitehead 1998; 2002). Given this abundance, it is surprising that sperm whale mitochondrial DNA control region diversity (mtDNA CR) is extremely low (Chapter 2; Lyrholm et al. 1996; Whitehead 1998). Several hypotheses have been proposed to explain the low mtDNA CR diversity including: selective sweeps of mtDNA linked with beneficial maternal cultural innovations in matrilineal social groups (Whitehead 1998; 2005), lineage extinctions of social groups and their mtDNA haplotypes (Alexander 2006, Siemann 1994, Tiedemann and Milinkovitch 1999), slow mtDNA mutation rates (Lyrholm et al. 1996, Whitehead 1998); constraints acting to restrict substitutions at most sites in the mtDNA CR (Alexander et al. 2013, Lyrholm et al. 1996); selective sweeps of mtDNA linked with beneficial mutations in the mtDNA protein-coding genes (Janik 2001); and an historical population expansion or population bottleneck (Lyrholm and Gyllensten 1998, Lyrholm et al. 1996). These hypotheses provide several predictions that could be tested with additional molecular and ecological evidence.

Lineage extinction events, such as mass stranding events, could lead to reduction in mtDNA diversity through the removal of social groups containing unique mtDNA haplotypes (Alexander 2006, Siemann 1994). However, mass stranding rates in the sperm whale were found to be too low for this to be the mechanism reducing mtDNA CR

diversity (Chapter 3, Appendix III). Slow mutation rates, as observed in some other cetacean species (Jackson et al. 2009), would mean fewer new mtDNA haplotypes evolved over time in the sperm whale, leading to lower overall mtDNA diversity. The constraints hypothesis is superficially similar to the slow mutation rate hypothesis in that it would limit the creation of new mtDNA CR haplotypes. However, under the constraints hypothesis, the mutation rate could remain high, yet substitutions would be restricted to only a few sites in the mtDNA CR. This would lead to high levels of homoplasy among haplotypes, but low overall mtDNA diversity. Similar phenomena are thought to be responsible for the low levels of mtDNA CR diversity in fishers (*Martes pennant)* and killer whales (*Orcinus orca)* (Knaus et al. 2011, Morin et al. 2010a). Alexander et al. (2013) sequenced 17 sperm whale mitogenomes and confirmed low overall diversity but found that the sperm whale did not have a slow mtDNA substitution rate, consistent with previous research (Lyrholm et al. 1996). Additionally, the mtDNA CR accurately reflected patterns of evolution shown by the protein-coding regions of the mtDNA, which was inconsistent with mtDNA CR-specific constraints. In light of these findings, an explanation is needed for the low diversity found throughout the entire sperm whale mitogenome, as well as the relatively recent coalescent date for sperm whale mtDNA in comparison with other cetacean species (Table 5.1). This leaves the selective sweep, population bottleneck and expansion hypotheses as the most likely explanations for low mitogenomic diversity in the sperm whale (Alexander et al. 2013).

Selection on a gene region can reduce diversity over a linked non-selected region of DNA, in a process known as a ‘selective sweep’. During a ‘selective sweep’ positive

selection for a gene region leads to that variant increasing in frequency in a population, with linked DNA regions 'hitch-hiking' (Guttman and Dykhuizen 1994, Nielsen 2001). During this process of the beneficial mutation rising in frequency, genetic variation is reduced at both the gene under selection, and at the 'hitch-hiking' DNA region. Previously, a selective sweep acting on the mtDNA CR linked to maternally-transmitted behavioral innovations was discounted due to observed levels of matrilineality in the sperm whale being too low (Chapter 3). However, strong positive selection on mtDNA-encoded proteins could lead to a sweep that reduces diversity in the mtDNA CR (Kim and Stephan 2003, Nielsen 2001). A plausible influence resulting in a selective sweep in sperm whales could be their prolonged foraging dives (Watkins et al. 1993, Watwood et al. 2006). Such behavior requires unique physiological adaptations that could include selection on mtDNA-encoded proteins, due to their role in oxidative phosphorylation (Ballard and Dean 2001, Janik 2001).

Alternatively, a population bottleneck is expected to reduce diversity, due to the enhanced genetic drift the population experiences as alleles are lost through the smallest 'bottlenecked' point (Nei et al. 1975). Previously, researchers have discounted the possibility of a whaling-induced bottleneck causing a significant loss of mtDNA CR diversity due to the sperm whale's high global abundance, long life-span and worldwide distribution (Whitehead 1998; 2002). These factors suggest that a whaling-induced bottleneck would have been too weak and brief to lead to the low mtDNA diversity observed in the sperm whale. Another scenario is that of a population bottleneck or expansion on an evolutionary time scale, pre-dating human influence. Under the

population expansion hypothesis, the historical population size is expected to be low with the population only recently expanding. This should be associated with low diversity, because the previously small population will have been subject to more genetic drift and therefore greater loss of genetic diversity than a stable, large population (Allendorf 1986, Wright 1931). Following an increase in abundance after a bottleneck or expansion event, there will be a lag before mutation-drift equilibrium is re-established, depending on mutation rate and population size (Birky-Jr et al. 1989, Reich et al. 1999, Storz et al. 2004). A bottleneck and/or expansion associated with climatic processes during the Pleistocene, particularly glaciation, is consistent with the sperm whale mtDNA TMRCA (time to most recent common ancestor) estimate of 24,000 to 137,400 years ago (Alexander et al. 2013, Lyrholm and Gyllensten 1998).

Selective sweeps, population bottlenecks, and expansions are all expected to reduce mtDNA diversity. However, only the population bottleneck and/or expansion scenarios are expected to reduce nuclear diversity (Kreitman 2000, Moyer et al. 2005, Rokas et al. 2001). Disentangling whether low nuclear diversity is due to a population bottleneck of a previously abundant population versus population expansion from a historically low abundance can be more difficult (Ramírez-Soriano et al. 2008, Schlötterer 2003). However, the use of extended Bayesian skyline plots to visualize the demographic reconstruction of species, and tests based on nuclear allelic diversity, offer potential methods (Heled and Drummond 2008, Maruyama and Fuerst 1985). These methods have not been previously applied to nuclear DNA sequence of sperm whales.

Here, we examine relative mitogenome and nuclear diversity levels using a suite of nuclear markers (including introns and coding regions) generated for both the sperm whale and the pygmy sperm whale (*Kogia breviceps),* a sister-taxon of the sperm whale. We included the pygmy sperm whale for comparative purposes, due to its close evolutionary relationship to the sperm whale. This is intended to help control for phylogenetic constraints in patterns of molecular evolution. In addition, the pygmy sperm whale has previously been found to have high mtDNA CR diversity in comparison with the sperm whale, consistent with levels found in other cetaceans (Alexander et al. 2013, Chivers et al. 2005). This suggests that the pygmy sperm whale has not been subject to a recent population bottleneck, expansion, or selective sweep. We use a variety of methods to contrast patterns of diversity in the mitogenome and nuclear markers of the pygmy sperm and sperm whale, including direct comparisons of genetic diversity, tests of diversity disequilibrium (Fu 1997, Tajima 1989), estimates of TMRCA, and extended Bayesian skyline plots to reconstruct the demographic history of the two species (Heled and Drummond 2008). In addition, we extend the geographic sampling of the sperm whale mitogenomes beyond the Pacific (represented by New Zealand and Oregon individuals), to the Atlantic, as represented by Gulf of Mexico individuals. Under the selective sweep hypothesis, we would expect low diversity and a young TMRCA to be restricted to the sperm whale mitogenome (Table 5.2). We would also expect demographic reconstructions to show large stable population sizes. Under the bottleneck hypothesis, we would expect low nuclear and mitogenomic diversity in the sperm whale, and for diversity disequilibrium tests to be significant for both the mitogenome and nuclear markers in the sperm whale (Table 5.2). We would also expect demographic

reconstructions to show signatures of the bottleneck event. In comparison, a population expansion would also be expected to result in low nuclear and mitogenomic diversity in the sperm whale, but demographic reconstructions should show a recent expansion from a previously small population size in the sperm whale. Recent TMRCA estimates would be expected for both nuclear loci and mitogenomes (Table 5.2).

5.3. Materials and Methods

5.3.1. Sample collection

Samples of New Zealand sperm whales (n = 14) and pygmy sperm whales (n = 17) , originating from stranded individuals, were obtained from the New Zealand Cetacean Tissue Archive (CeTA) at the University of Auckland (Thompson et al. 2013). One Oregon sperm whale stranding sample was obtained from the Oregon Marine Mammal Stranding Network (OMMSN) Tissue Archive. Samples of Gulf of Mexico sperm whales (n = 10) were collected using a small stainless steel biopsy dart deployed from a crossbow (Lambertsen 1987) under NMFS permit No. 369-1757. The New Zealand and Oregon sperm whale samples had been previously extracted and sequenced for mitogenomes as part of Alexander et al. (2013). All other samples had DNA extracted following a standard phenol/chloroform method (Sambrook et al. 1989), modified for smaller samples (Baker et al. 1994).

5.3.2. Short-range nuclear intron amplification and Sanger sequencing

Eleven nuclear short-range (<1,500 bp) introns (*Actin, Amel-Y, BGN, CAT, CHRNA1, ESD, FGG, G6PD, GBA, PTH, RHO)* were screened on the basis of success in other

cetacean species (Alter et al. 2007, Jackson et al. 2009, Macé and Crouau-Roy 2008, Ruegg et al. 2010). PCR conditions were modified from Jackson et al. (2009). Each reaction consisted of 15-30 ng of sample DNA, and a final concentration of 0.94× Platinum *Taq* buffer (Invitrogen), 0.38 μM of each primer, 2.34 mM $MgCl_2$, 0.19 mM dNTP and 0.1U of Platinum *Taq* polymerase (Invitrogen), with ddH_2O to 16 μL total volume. The temperature profile consisted of an initial denaturing step of 3 min at 94°C, followed by 35 cycles of 94°C for 30 s, locus-specific annealing temperature (Appendix V, Table V.1) for 45 s and 72°C for 1 min 30 s, followed by a final extension step of 72°C for 10 min. PCR amplifications were attempted for all 17 pygmy sperm whale samples, the 14 New Zealand sperm whale samples, the 10 Gulf of Mexico sperm whale samples, and the sperm whale sample originating from Oregon. Where amplifications were successful, sequencing was attempted in the forward and reverse direction for each individual for each locus. Excess nucleotides and primers were removed using Sap/Exo (GE Healthcare). Cycle sequencing was carried out using a BigDye (Applied Biosystems Inc.) protocol, and a final cleaning step performed using the standard CleanSEQ (Agencourt) protocol. The products were then run on an ABI3730xl DNA Analyzer.

Sequences were aligned in *Sequencher v. 4.6* (Gene Codes) and trimmed to primers and/or when sequence quality began to decline at the end of a fragment. For quality control purposes, sequences were required to have >90% of bases exceeding a PHRED score of 20 to be retained in the dataset. Sequences with PHRED scores reduced by indels and sequencing artifacts (e.g. dye blobs) were reviewed and edited by eye. Heterozygote sites were first identified in *Sequencher* using a threshold for secondary peak height

greater than or equal to 25% of the primary peak, and then confirmed visually. After calling heterozygous sites, the forward and reverse sequences were compared to create a consensus sequence for each individual. For each species, an alignment of all individuals for each locus was constructed by eye, and within-species variable sites were visually confirmed in each individual. A total alignment consisting of both pygmy sperm and sperm whales was then constructed. This alignment was trimmed to the shortest consensus that still covered all variable sites observed within each species. Any individuals with sequences shorter than this length were removed from the dataset. By species, individuals were then phased for each locus using *Phase v. 2.1* (Stephens and Scheet 2005, Stephens et al. 2001).

5.3.3. Long-range mitogenome and nuclear gene amplification and Illumina Miseq sequencing

Mitogenome sequence for 12 of the New Zealand sperm whale samples and the single Oregon sperm whale sample were obtained from Alexander et al. (2013). In this current study, mitogenomes were generated for 9 of the Gulf of Mexico sperm whale samples, and 10 of the pygmy sperm whale samples, through long-range overlapping PCR fragments, following a modified version of the protocol described in Alexander et al. (2013) (see Appendix V, Table V.1 for details). Several primers were redesigned to ensure amplification in the pygmy sperm whale (Appendix V, Table V.1). The protocol for generating sperm whale mitogenomic sequence used PCR fragments long enough (>1,500 bp) to be prepared for next-generation sequencing using whole genome library construction methods (Alexander et al. 2013). To include nuclear genes using the same methodology, nuclear alignments that exceeded 1,500 bp in length, with sequence

information available for both the pygmy sperm whale and sperm whale, were identified on GenBank. Loci were excluded if they consisted of concatenated exons, leaving four candidate loci: *SRY* (Nishida et al. 2007), *ApoB, BRCA1* and *TTN* (Meredith et al. 2011). Where possible, primers from previous studies were used (Appendix V, Table V.1), otherwise primers were designed to ensure amplification of a 1,500 bp – 2,000 bp nuclear insert. Loci were amplified in overlapping fragments with the following reagent concentrations: 1× Phusion® HF Buffer (New England Biolabs, USA); 0.5 μM of each primer (Integrated DNA Technologies); 2 % DMSO (NEB, USA); 15-30 ng of template DNA, with dNTP (Promega, USA) and Phusion® Polymerase (NEB, USA) concentration varied by fragment as described in Appendix V, Table V.1, and ddH2O to 20 μL. Thermoprofiles consisted of an initial denaturation step of 98 °C for 30 s followed by 35 cycles of 98 °C for 8 s, the specific annealing temperature for each fragment for 30 s (Appendix V, Table V.1), and 72 °C for 1 min 15 s; followed by a final extension of 72 °C for 10 min. Long-range nuclear genes were generated for the 22 sperm whales and the 10 pygmy sperm whales that also had their mitogenomes sequenced.

Gel electrophoresis was used to quantify the DNA present for each long-range fragment, and fragments were combined by individual in equimolar amounts, compensating for differences in ploidy. The combined fragments for each individual were then cleaned of excess primers and nucleotides using a Qiagen PCR purification kit (Qiagen), and individually-barcoded and prepared for sequencing using a Nextera XT DNA Sample Preparation Kit (Illumina). Individuals were pooled and run on three Illumina Miseq runs (two of these runs were 250 bp paired end runs, one was 75 bp paired end). Reads were

assembled to mitogenome and nuclear gene references (GenBank accession numbers: AF284016, JN414151, JN413980-JN413981, JN414999-JN415000, KC312603, NC_005272; Alexander et al. 2013, Arnason et al. 2004, Madsen et al. 2001, Meredith et al. 2011) using *bwa v. 0.7.4* (Li and Durbin 2009), after using the default settings (PHRED cut-off of 20) in *Trim Galore! v. 0.2.8* (Babraham Bioinformatics 2013) to remove adaptor sequence and trim poor quality sequence from reads. The consensus sequence (>50%) from the *bwa* assembly, first removing duplicate reads, was calculated using *Samtools v. 0.1.19* (Li et al. 2009a). Any site that had a sequencing depth of less than 15× was screened from analysis following Alexander et al. (2013). Given the large sequencing depth for samples in this study, the impact of down-sampling to 100,000 reads and increasing the PHRED cut-off to 30 in *Trim Galore!* was explored for a subset of the samples (Pma11GMX04, 05, 06, 08, 09).

For the mitogenomes, base calls supported by fewer than 70% of reads were reviewed for possible heteroplasmy/indels/pseudogene incorporation, following Alexander et al. (2013). If these occurred on sites variable across the entire alignment, they were converted to an ambiguity code. Assembly coverage plots were examined for gaps indicative of the incorporation of nuclear mitochondrial DNA (numt) pseudogenes. The potential for numts was also assessed by examining divergence by gene to detect any incorporation of divergent numts, and by examining coding sequences for frame-shifts/premature stop codons. If numts were detected, that long-range fragment of the mitogenome was masked from analysis. The mitogenome was represented by the control region and concatenated protein-coding regions in downstream analyses, excluding *ND6*

due to its location on the opposing strand and distinct patterns of evolution (following Alexander et al. 2013 and Ho and Lanfear 2010). Overlapping regions of protein-coding genes were represented only once in the alignment.

For nuclear genes, all bases supported by fewer than 75% of reads were reviewed as possible heterozygote sites/indels. This threshold was considered comparable to the 25% secondary peak threshold used for the Sanger-sequenced nuclear introns. After heterozygote sites had been identified, the phasing of sites was determined using *PEppyphase* (a custom-written script described in Appendix V, Figure V.1). *PEppyphase* uses *bwa*, *samtools*, *BEDTools* (Quinlan and Hall 2010) and *R* (R Core Team 2013) to haplotype samples, relying on the physical linkage of next-generation sequencing reads across multiple variable sites. *PEppyphase* contrasts with other software that infer haplotype phase (e.g. *HapCUT*, *HapCompass* - Aguiar and Istrail 2012, Bansal and Bafna 2008), as these programs use probabilistic graphical theory. Instead, in *PEppyphase*, the phase is based on the most common combinations of nucleotides across SNPs. *PEppyphase* also takes a custom input file 'variable_sites' instead of a *.vcf file, which can allow researchers to screen sites, such as those occurring on primer regions, from the analysis. For samples that could not be completely phased, haplotypes were imputed from the homozygote/phased heterozygote alleles of other samples. After assembly and phasing, the coverage and base call at all within-species variable sites was confirmed for each individual for both the mitogenomes and nuclear genes. For the nuclear genes, primer regions were trimmed from the beginning and end of the alignment. Variable sites

in primer regions in the middle of the alignment were masked in both the mitogenome and nuclear fragments.

5.3.4. Diversity metrics

Observed heterozygosity was calculated on a per-locus basis, dividing the total number of sampled heterozygotes by the total number of individuals sequenced. Haplotype diversity, number of alleles, proportion of variable sites (total, and by transitions, transversions, and indels), nucleotide diversity and theta (*S*: Watterson 1975) were calculated in *Arlequin v. 3.5* (Excoffier et al. 2005) using the Tamura and Nei (1993) correction. Additionally, nucleotide diversity by codon position was calculated for the protein-coding regions of the mitogenome and long-range nuclear genes using *MEGA v. 6.0* (Tamura et al. 2013).

Where available for loci sequenced in this study, published estimates of diversity for other cetacean species were also compared to the sperm whale. To test significance of nuclear genetic diversity differences between the sperm whale and other species, paired t-tests were carried out for each measurement of diversity, using the loci as independent samples (Nei 1987). For the CR and protein-coding regions of the mtDNA, differences in diversity between the pygmy sperm whale and sperm whale were evaluated by t-tests (Nei 1987) using the standard deviations output by *Arlequin* (with degrees of freedom equal to the combined sample size minus two). Tajima's *D* (Tajima 1989) and Fu's F_S test (Fu 1997) were performed with *Arlequin* (Excoffier et al. 2005), with 10,000 samples simulated for each test to provide an estimate of variance. Tajima's *D* was considered

significant at $\alpha < 0.05$, with Fu's Fs considered significant at $\alpha < 0.02$ following Fu (1997) and Excoffier et al. (2005).

5.3.5. Bayesian estimation of substitution rates

With the exception of the mitogenome, which was previously analyzed by Alexander et al. (2013), a multispecies alignment was constructed for each locus using cetacean species available on GenBank (Figure 5.1). If multiple sequences of equivalent length were available for a given species, one sequence was randomly chosen, otherwise the sequence with the maximum length was used to represent that species. To maintain consistency with Alexander et al. (2013), the sperm whale was represented by sequences generated from the individual PmaNZ005, and the pygmy sperm whale was represented by a randomly chosen individual, KbrNZ090. The long-range protein-coding nuclear genes were partitioned by codon, and all multispecies nuclear analyses used ambiguous states in the calculation of likelihood, as GenBank data was often unphased. Due to the low species divergence, alignments were constructed by eye. Species-specific substitution rates were then estimated in *BEAST v. 2.1* (Bouckaert et al. 2014) using the same parameters as Alexander et al. (2013), including: the same fossil calibration dates and prior shapes, a GTR + I + Γ nucleotide substitution model, estimated base frequencies, an uncorrelated lognormal relaxed clock (fixing the substitution rate, and allowing the clock rate to vary: Drummond et al. 2006) and a Yule tree prior (fossil-calibrated nodes shown in Figure 5.1). Also following Alexander et al. (2013), two independent runs of 90,000,000 states, sampling every 3,000 states were completed, with convergence of tree topologies (following annotation using *TreeAnnotator v. 2.1*,

Bouckaert et al. 2014) and posterior values confirmed in *FigTree v 1.4.1* (Rambaut 2014) and *Tracer v 1.6* (Rambaut et al. 2013), respectively. After confirming that each independent run had effective sample sizes (ESS) of more than 200, and showed convergence, the runs were combined using *LogCombiner v. 2.1* (Bouckaert et al. 2014). If intron runs failed to converge, or reach ESS values of 200, the nucleotide substitution model was simplified to HKY + Γ for that locus. Further failure to converge led to the use of a strict clock instead of uncorrelated lognormal relaxed clock for that locus. For coding genes, the analysis was first simplified to a 1+2, 3 codon partitioning scheme, and then further failure to converge led to the same model simplification steps as for the introns.

5.3.6. Demographic reconstructions and effective population size estimates

To reconstruct the demographic histories of the sperm whale and pygmy sperm whale, the extended Bayesian skyline plot approach was used (Heled and Drummond 2008), as implemented in *BEAST*. This approach was selected as it can incorporate multiple unlinked markers, adjusting for differences in ploidy level (e.g. mtDNA vs nuclear genes). The multispecies *BEAST* runs were used to obtain the median substitution rates and substitution model for each locus for the pygmy sperm and sperm whale (Figure 5.2). The substitution model and median substitution rates for each species for the control region and concatenated protein-coding regions of the mitogenome were obtained from Alexander et al. (2013). Using these rates, an extended Bayesian skyline plot (EBSP: Heled and Drummond, 2008, implemented in *BEAST*) was used to create a demographic reconstruction for each species across the multiple loci sequenced in this study, with loci

scaled for ploidy level. Concurrently, the TMRCA for each locus was estimated. All loci were unlinked in terms of substitution, clock, and tree models with the exception of the concatenated protein-coding regions and control region of the mitogenome. Empirical base frequencies were used and the EBSP operators weighting changed to 100 *sensu* the recommendations of the EBSP tutorial (Heled 2010). To account for potential bias from patterns of selection, these 'full' runs were repeated excluding the mitogenome and 1st and 2nd codon positions of the nuclear protein-coding regions to create a 'neutral' analysis. In addition, the per-locus species-specific substitution rates were used with the theta estimates generated in ***5.3.4 Diversity metrics***, and estimates of species-specific generation times (listed in Appendix V, Table V.2) from Taylor et al. (2007) to approximate effective population size by locus for each species (accounting for ploidy level).

5.4. Results

5.4.1. Sequencing results

Of the 11 nuclear short-range introns screened, 5 provided good quality sequence for a minimum of 17 sperm whale individuals and 9 pygmy sperm whales (Table 5.3). Of the discarded loci, *BGN*, *Amel-Y, GBA, CAT,* and *PTH* failed to amplify and/or sequence cleanly in the sperm whale. *Actin* contained a large insertion in the sperm whale and pygmy sperm whale sequences relative to other cetacean species, and was not reliably amplified due to its length (> 1,000 bp). Of the long-range genes, three (*ApoB*, *BRCA1*, *TTN*) reliably amplified, with *SRY* discarded due to infrequent and inconsistent amplification among the male samples included in this study. For the long-range genes

and mitogenomes, 22 sperm whale and 10 pygmy sperm whale individuals had sequence available. High coverage was achieved for all fragments: average sequencing depth for the mitogenomes was 5851×, for *ApoB* = 6626×, *BRCA1* = 5904×, and *TTN* = 5828× (average sequence depth by species and individual sample is given in Table 5.4).

To investigate whether the high coverage observed could lead to an increase in assembly errors, the impact of down-sampling to 100,000 reads and increasing the PHRED cut-off to 30 in *Trim Galore!* was explored for a subset of the samples (Pma11GMX04, 05, 06, 08, 09). Down-sampling to 100,000 reads represented between 4.05% of the original number of reads (Pma11GMX09) to 14.7% (Pma11GMX04). Average sequencing depth of the down-sampled assemblies for mitogenomes was 975×, for *ApoB* = 319×, *BRCA1* = 128×, and *TTN* = 111×. The minimum average 'down-sampled' sequencing depth observed was 48.6× for Pma11GMX08's *BRCA1*, and the maximum 1110× for Pma11GMX09's mitogenome. One difference was observed between the 'full' and 'down-sampled' mtDNA control regions of Pma11GMX06 and Pma11GMX09. However, these do not present high quality differences as they were characterized by mixed reads in the down-sampled assemblies (<70% support for the majority base). One heterozygote site was not identified in the down-sampled Pma11GMX08's *BRCA1* in comparison with the 'full' assembly (as it had >75% support for the majority base). This appeared to be a legitimate heterozygote site in the 'full' assembly, as it was successfully phased using *PEppyphase*. Based on these results, it appears the quality control steps taken with the 'full' assemblies (e.g. removing PCR duplicates, trimming to a PHRED of 20) led to accurate assemblies even in the presence of very high sequencing depth. In

fact, down-sampling appeared to increase assembly errors in comparison with the ‘full’ assemblies.

When reviewing the mitogenome long-range PCR fragments for the presence of numts, Fragment 3 had areas of lower coverage for several pygmy sperm whale individuals (KbrNZ021, KbrNZ056), potentially indicative of incorporation of a linear numt. Review of the protein-coding genes covered by this fragment (*ND4, ND4L, ND5*) revealed the presence of frame-shift indels in multiple samples as well as greater levels of divergence between samples than in other genes. Therefore, these genes were excluded from the pygmy sperm whale mitogenome coding region alignment. In addition, KbrNZ056 had a frame-shift indel in *Cytb* (Fragment 4) and was excluded from the alignment for this gene. Finally, KbrNZ102 had multiple other genes where frame-shift indels were detected, so was excluded completely from the mitogenome analyses. No indications of numts were found for the sperm whales sequenced in this study, consistent with previous studies using the same methodology (Alexander et al. 2013). Further evidence for low-level background amplification of numts in the pygmy sperm whale came from a large percentage of samples that had nucleotide calls supported by fewer than 70% of reads (converted to ambiguity codes for downstream analyses) on variable sites in comparison with the sperm whale. The genotype for each sample at each locus, as well as the sequence for each unique allele is archived as described in Appendix V.

5.4.2. Diversity metrics

For the mtDNA (Table 5.3), the pygmy sperm whale had a larger proportion of variable sites for both the CR and protein-coding regions (CR 0.073 vs 0.015, protein-coding 0.020 vs 0.006, chi-squared $p < 0.0167$). Additionally, for the protein-coding regions, the pygmy sperm whale had significantly higher haplotype diversity (1.00 vs 0.957, t-test $df = 29$, $p = 0.0035$) and for the CR, significantly higher theta (22.1 vs 3.84, t-test $df = 29$, $p = 0.0341$). The sperm whale showed a significant excess of transitions at variable sites compared to the pygmy for the mtDNA CR (100% vs 72.4%, chi-squared test $p = 0.0324$). No significant difference in the proportions of transitions, transversions, and indels on variable sites was observed for the protein-coding regions. No significant differences were detected for mtDNA CR haplotype diversity or protein-coding theta. No significant differences in nucleotide diversity were observed between pygmy sperm vs sperm whale, but the observed values were consistent with the pygmy sperm whale having greater mtDNA diversity than the sperm whale (CR 0.802% vs 0.318%, protein-coding 0.510% vs 0.121%).

The pygmy sperm whale equaled or exceeded the level of genetic diversity shown by the sperm whale across a number of metrics based on the 8 nuclear loci (>12,000 bp combined, Table 5.3). For the nuclear loci (using a paired t-test across loci, $df = 7$), the pygmy sperm whale had significantly higher average observed heterozygosity (0.774 vs 0.475, $p = 0.0115$), expected heterozygosity (0.832 vs 0.462, $p = 0.0035$), number of alleles (13 vs 6.5, $p = 0.0082$), proportion of variable sites (0.023 vs 0.006, $p = 0.0167$), nucleotide diversity (0.425% vs 0.072%, $p = 0.0144$), and theta (average locus-wide

estimate: 5.63 vs 1.21, $p = 0.0012$). Estimates of theta calculated for the sperm whale nuclear loci were consistent with previous estimates based on SNP markers (Morin et al. 2007). The only comparisons between the two species that were not significant were the proportions of transitions, transversions, and indels on variable sites.

The sperm whale also had relatively low nuclear genetic diversity in comparisons with other cetacean species based on nuclear introns (Figure 5.3, Table 5.5). Sperm whales had a lower percentage of variable sites than the humpback and Antarctic minke (Figure 5.3). Despite this lower overall diversity in comparison with other cetacean species, rarer mutation classes such as transversions and indels were found in the sperm whale. Sperm whales also appeared to have lower numbers of *CHRNA1, ESD, G6PD* and *RHO* alleles compared to other cetacean species, and their number of *FGG* alleles was low compared to the Antarctic minke (*Balaenoptera acutorostrata)* (Table 5.5). Sperm whales had lower observed heterozygosity for *CHRNA1, ESD, G6PD,* and *RHO* compared with other cetacean species (Table 5.5). Levels of nucleotide diversity and theta in the sperm whale were low compared to other species (Table 5.5), even the southern right whale (*Eubalaena australis*) and the humpback whale (*Megaptera novaeangliae*): species subjected to extreme and extended whaling-mediated bottlenecks (Perry et al. 1999). Although diversity was strikingly lower in the sperm whale than the pygmy sperm whale and other species, the diversity disequilibrium tests conducted (Tajima's *D* and Fu's F_S) showed no consistent differences between the sperm whale (6/20 tests significantly negative) and pygmy sperm whale (8/20 tests significantly negative, Table 5.3).

5.4.3. Bayesian estimation of substitution rates

All multispecies runs for obtaining species-specific substitution rates for each locus reached ESS values of over 200 and showed convergence for the following model specifications. For *CHRNA1*, the GTR + I + Γ nuclear substitution model was implemented. For all other nuclear introns (*ESD, FGG, G6PD, RHO*), adequate ESS values were not reached under the GTR model, so HKY + Γ was used as the nucleotide substitution model. For the nuclear coding genes, *ApoB* and *TTN*, the 1st and 2nd codon positions were combined into a single partition, with a separate partition for the 3rd codon position, implementing the GTR + I + Γ nuclear substitution model. The GTR + I + Γ nuclear substitution model was also implemented for *BRCA1*, but this locus was completely split into 1st, 2nd and 3rd codon partitions. The uncorrelated relaxed lognormal clock was used for all loci. Rates across nuclear loci were broadly similar to those seen in previous publications (Figure 5.2: approximately 0.05%/myr, Jackson et al. 2009; Alter et al. 2007), and were considerably slower (~10-30%) than rates previously characterized for the mitogenome (Figure 4.4), consistent with previous research (Jackson et al. 2009). Given the similar substitution rates for the loci that overlap between this and previous studies (Figure 5.2), estimates of N_e based on loci not sequenced in this study were also considered in section 5.4.4.

5.4.4. Demographic reconstructions and effective population size estimates

All pairs of EBSP *BEAST* runs had ESS values that exceeded 200, and converged on the same parameter values and tree topologies, with the exception of the 'neutral' demographic reconstruction for the sperm whale (excluding the mitogenome and 1st and

2nd codon positions of nuclear genes). This is likely due to the low overall diversity of the sperm whale nuclear gene partitions. Given the consistency of the 'neutral' and 'full' runs for the pygmy sperm whale, the 1st and 2nd codon positions for the nuclear genes were added back into the sperm whale analysis in an attempt to reach convergence. ESS values remained low, but acceptable, for this run (100 < ESS < 200). For the pygmy sperm whale, the 'neutral' runs produced very similar results to the runs based on the 'full' dataset (no median went 'out of bounds' of the 95% highest posterior density interval of another run). For the sperm whale, 'nuclear' only runs had an overall effective population size approximately 10× that of the 'full' dataset, but 95% higher probability intervals between runs still overlapped over the majority of the time interval, and the population size remained fairly constant over time (a slight non-significant increase was observed). Given the overall similarity between the 'full' and 'neutral/nuclear' EBSP, only results for the 'full' datasets are presented here.

The EBSP demographic reconstructions for the two species showed that the pygmy sperm whale underwent an expansion approximately two million years ago, reaching an effective population size of ~1,350,000 by about 500,000 years ago (Figure 5.4). In contrast, the median effective population size of the sperm whale remained constant, and low, at around 28,000 for the last two million years (Figure 5.4). Very recent (< 20,000 years) widening of the 95% highest posterior density intervals occurred in the sperm whale, but no obvious recent trend was observed for the median effective population size (Figure 5.4). An overall difference in the long-term effective population size of the two species was also reflected in the much longer times to coalescence for the pygmy sperm

whale compared with the sperm whale (Figure 5.5). However, for both species, the TMRCA was more recent for the mtDNA in comparison with the nuclear markers (Figure 5.5).

Substitution rates obtained from the fossil-calibrated cetacean-wide phylogenies for each gene (Figure 5.2), and generation times (summarized in Appendix V, Table V.2) obtained from Taylor et al. (2007) were used to calculate effective population size on a per-locus basis (Figure 5.6). The estimates of effective population size were significantly lower in the sperm whale (average N_e = 14,000) than in the pygmy sperm whale (167,000), Antarctic minke (179,000), gray whale (34,000) and humpback whale (18,000 - 94,000 depending on ocean basin), based on t-tests using the locus-specific estimates of effective population size as the independent sample, $p < 0.0354$).

5.5. Discussion

The finding of low nuclear and mitogenomic diversity in the sperm whale relative to the pygmy sperm whale and other cetacean species is strongly indicative that a recent population expansion from a previously small population size is responsible for the low overall genetic diversity found in the sperm whale. This is supported by findings of low effective population sizes across the markers surveyed in this study, and the EBSP that showed a consistently small, yet stable, population for as far back in time as coalescent events were informative. We discount the likelihood of a recent population bottleneck, as coalescent times for all genes from this study were recent and overall diversity was low, indicating that no ‘divergent’ lineages escaped through a relatively recent bottleneck

event. Based on the EBSP, if a bottleneck event did reduce genetic diversity in the sperm whale, it would have to have occurred in the distant past (more than one million years ago) with population abundance remaining low since this time.

5.5.1. Detection and timing of the expansion

The lack of expansion observed in the median of the EBSP, and the small effective population size estimates obtained for the sperm whale, suggests that any expansion in this population to its current abundance must be very recent. Given that we infer a population expansion occurred recently in the sperm whale, it is perhaps surprising the Tajima's *D* and Fu's F_S tests did not show consistent indications of a population expansion in the sperm whale. However, diversity disequilibrium tests are notorious for being relatively weak in their ability to detect population expansions (Simonsen et al. 1995), especially when based on sequences with low variability (Hendrickson et al. 2003, Simonsen et al. 1995). Therefore a null (or inconsistent) result for these tests cannot be inferred to mean that an expansion is not responsible for the lower than expected levels of diversity in the sperm whale (Kreitman 2000, Wayne and Simonsen 1998, Whitehead 1998). In contrast, the pygmy sperm whale had a number of significant Fu's F_S tests, indicating an excess of singleton mutations (Ramos-Onsins and Rozas 2002). We believe this is likely driven by the limited sampling of the pygmy sperm whale: genetic diversity appears to be high in this species, and our sample size was not large enough to adequately characterize the population.

Despite a lack of expansion indicated in the median data of the sperm whale EBSP, the 95% highest posterior density interval widened around 20,000 years ago, potentially indicating the expansion to current population sizes. The use of phylogenetic substitution rates to calibrate the timing of demographic events occurring within a lineage is likely to systematically overestimate the timing, so this expansion could be even more recent (Ho et al. 2005, Ho et al. 2008). Based on the exponential growth equation:

$$P_0 = P_t e^{rt}$$

where P_0 is the initial population size, P_t the population size at time *t*, and *r* the intrinsic rate of increase for the population, an expansion from $N = 1$ (i.e. a single pregnant female), to the estimated pre-whaling abundance of 1,100,000 whales (Whitehead 2002), could occur in as little as 1,450 years based on an intrinsic rate of increase of 0.096% (Chiquet et al. 2013). Therefore, a recent expansion is consistent with sperm whale demographic parameters.

5.5.2. Potential explanations for the timing

Sperm whales and one of their prey species, the giant squid (*Architeuthis* sp.), show an apparent parallel in terms of low mitogenomic diversity and relatively recent coalescent times (Chapter 2, Winkelmann et al. 2013). As with the sperm whale, the giant squid appears to be globally distributed and abundant, at odds with its low mitogenomic diversity (Winkelmann et al. 2013). A sudden expansion occurring 32,000+ years ago from an historically smaller population size has been proposed as an explanation for the

low mitogenomic diversity observed in the giant squid (Winkelmann et al. 2013). Similarly, we inferred the expansion of sperm whales to have occurred very recently (<20,000 years) due to the large current population size, and lack of obvious trend of increase in our EBSP. This would be consistent with sperm whale abundance increasing in response to an increase in the abundance of their prey, and also with previous mtDNA-based research indicating a potential bottleneck in the sperm whale between 24,000 and 92,000 years ago (Lyrholm and Gyllensten 1998). However, in order to accurately gauge when the inferred population expansion occurred in the sperm whale, additional loci will be needed. The low level of variability and relatively slow substitution rates of the markers used in this study limited clear identification of the timing of the population expansion (Schlötterer 2003).

Giant squid currently form only a portion of the global sperm whale prey intake (Whitehead 2003). However, inferred recent population expansions (~25,000 years ago) have occurred in other squid species that likely form the basis of sperm whale diets (Jaquet et al. 2003), including the jumbo squid (*Dosidicus gigas*) (Ibáñez et al. 2011) and Patagonian squid (*Doryteuthis gahi*) (Ibáñez et al. 2012) (Table 5.1). Squid abundance and distribution appears to be strongly tied to water temperature and oceanographic conditions (Dawe et al. 2007). Historical population expansions in squid species have been attributed to warming after the last glacial-interglacial period, which likely led to an increased latitudinal range in these species (Ibáñez et al. 2011). If the population expansions in squid, and therefore sperm whales, were caused by these climatic effects, this also has ramifications for how the sperm whale might respond to on-going

anthropogenically mediated climate change. Further increases in the population size of the sperm whale could be expected, but this would depend on how squid will respond to further climate change. This is the subject of considerable uncertainty (Pecl and Jackson 2008).

In addition to the sperm whale's inferred population expansion, the pygmy sperm whale appeared to go through its own expansion much earlier (>500,000 years ago). The Pliocene and Pleistocene were characterized by cooling and sea-level fluctuations that could have impacted on abundance (Steeman et al. 2009). In addition, productivity of oceans is thought to have greatly increased around 2 mya with an increase in upwelling (Lawrence et al. 2006), correlating with the beginning of the inferred pygmy sperm whale expansion. However, as for the sperm whale, additional nuclear loci would help to more accurately characterize when this species expanded in population size.

5.5.3. A synthesis of the processes shaping patterns of genetic diversity and differentiation in the sperm whale

In Chapter 2, the sperm whale was characterized as having low nucleotide diversity based on the mtDNA CR. Sequencing of mitogenomes in Chapter 4 showed that low nucleotide diversity was a pervasive feature of the entire mitogenome, and not just restricted to the control region. Finally, sequencing of nuclear loci within this current chapter illustrated low nucleotide diversity within the nuclear genome as well. Fossil-calibrated phylogenies constructed on mitogenomes (Chapter 4) and nuclear loci (this chapter) suggest a recent TMRCA for the sperm whale across nuclear and mtDNA markers. Along with

demographic reconstructions indicating a low effective population size in the sperm whale, we have interpreted this data as being supportive of a recent population expansion explaining the low diversity in the sperm whale.

Despite the broad patterns of low genetic diversity due to this putative population expansion, striking levels of genetic differentiation have been observed in the sperm whale, particularly in the maternally-inherited mtDNA (Chapter 2). However, the patterns of differentiation observed at the oceanic level showed a potentially conflicting pattern: strong patterns of mtDNA differentiation were observed between and within oceans, yet mtDNA haplotypes were shared across oceans. A population expansion helps explain this apparent disparity: the sharing of haplotypes between oceans likely reflects the recent expansion of the sperm whale. The time since the expansion has been too limited for ocean and region-specific lineages to have evolved, also explaining the lack of phylogeographic structure detected in Chapter 2.

The recent global expansion of a single maternal lineage, followed by strong local drift due to maternal social organization was characterized as a ‘maternally-mediated big bang’ in Chapter 2. In this dissertation, female philopatry has been identified at various scales, ranging from social group (demonstrated by lenient matrilineality in Chapter 3) through to higher geographic levels such as region and ocean (Chapter 2), supporting previous behavioral evidence of philopatry at these levels summarized in Chapters 2 and 3. These patterns of female philopatry have been a powerful force in structuring mtDNA differentiation at various hierarchical levels in sperm whale populations since the

population expansion. In contrast, male-mediated gene flow appears to have been a powerful force in preventing marked nuclear differentiation of regions and oceans. Differences in the levels of differentiation based on mtDNA and microsatellites can be driven by differences in the mode of inheritance and ploidy levels of these markers (Birky-Jr et al. 1989, Palumbi et al. 2001). However, the extreme disparity in levels of structure among these markers in the sperm whale (Chapter 2) is strongly suggestive of the importance of male-mediated gene flow.

5.5.4. Alternative explanations for the low genetic diversity observed in the sperm whale

It is important to note that effective population size estimates are generally lower than the census population size, both in cetaceans and other species (Frankham 1995, Whitehead 1998). For example, using the N_e estimate of 28,000 derived in this study for the sperm whale from the *BEAST* population trajectory, and the estimated pre-whaling abundance of 1,100,000 (Whitehead 2002), the sperm whale shows an N_e/N_c ratio of 0.025. Deviations from an N_e/N_c ratio of 1 occur because real biological populations differ from the Fisher-Wright model of an idealized population (Wright 1931, 1938) due to factors including fluctuating population size, non-random mating, reproductive variance, age-structure, spatial structure, and overlapping generations (Charlesworth 2009). An alternative to a population expansion, which reduces N_e due to the fluctuations in population size, is that the sperm whale possesses an extremely skewed N_e/N_c ratio due to other factors.

The N_e/N_c ratio observed in the sperm whale (0.025) is low in comparison to other cetacean species, which generally have a N_e/N_c ratio of ~0.1 (Whitehead 1998). Several of the previously mentioned factors that can influence the N_e/N_c ratio are not unique to sperm whales in comparison with other cetaceans, including the potential for non-random mating, age-structure, spatial structure, and overlapping generations. Therefore, it seems unlikely these factors on their own can explain the lower N_e/N_c ratio observed in sperm whales compared with other species. However, another characteristic of the sperm whale that could impact effective population size is its polygynous social structure, likely to lead to extreme variance in reproductive success of males (Best 1979). If male reproductive variance was responsible for the low N_e/N_c ratio in sperm whales, after accounting for differences in ploidy and inheritance mode, this should mean effective population size estimates based on the maternally-inherited mtDNA would be markedly higher than those based on the bi-parentally inherited nuclear markers. We did not observe this: mtDNA gave effective population size estimates on the low side of the combined estimates across all genes for the sperm whale (Figure 5.6).

Even though the extreme sexual dimorphism shown by the sperm whale often leads to a focus on male reproductive variance (Best 1979), behavioral differences between female social groups located within different acoustic clans (Marcoux et al. 2007, Whitehead and Rendell 2004) suggest this sex could also experience variation in reproductive success among clans, social groups, and individuals. Observations of apparently maternally-transmitted behaviors led to the formation of the beneficial maternal cultural innovations selective sweep hypothesis for reducing mtDNA CR diversity (Whitehead 1998; 2005).

The low levels of nuclear DNA diversity described in this current study suggest that a selective sweep (either culturally or molecular-based) is not solely responsible for the low mtDNA diversity observed in the sperm whale. However, variation in female reproductive success due to beneficial maternal cultural innovations could be a mechanism contributing to the low N_e/N_c ratio observed in the sperm whale. As a low effective population size should be associated with greater levels of genetic drift (Allendorf 1986, Wright 1931), one issue with this hypothesis is the sharing of high-frequency common haplotypes across ocean basins (haplotypes A, B, C: Chapter 2). If the levels of inter-ocean dispersal of females have been restricted on a historical scale (as suggested by analyses of mtDNA differentiation: Chapter 2, Lyrholm and Gyllensten 1998, as well as analyses of individual movements summarized in Dufault et al. 1999), then it would be expected that different mtDNA haplotypes would have risen to fixation in different ocean basins through drift (under the assumption of a constant, but small, effective population size). In contrast, under the population expansion hypothesis, these high-frequency common haplotypes can be interpreted as evidence of common ancestors during the expansion.

To further investigate the influence of reproductive variance, long-term behavioral studies of female social groups and genetic investigation of patterns of paternity is warranted. Alternatively, approximate Bayesian computational models implemented with a large number of loci have been shown to be effective at detecting changes in population abundance (Robinson et al. 2014). If such an approach was carried out for the sperm whale and revealed no evidence of a population expansion, this would strengthen the

inference that reproductive variance could be responsible for the low observed levels of genetic diversity in the sperm whale.

Alternatively, given the relatively recent TMRCA of the sperm whale mitogenome relative to nuclear loci, it might be tempting to conclude that a selective sweep *has* occurred in the sperm whale mitogenome. However, the pygmy sperm whale's mitogenome TMRCA is also more recent than that of its nuclear genes. On average the coalescence times for nuclear genes are expected to be approximately four times greater than those of the mtDNA (Birky-Jr et al. 1989, Palumbi et al. 2001). This is consistent with our data (Figure 5.5), so it seems that the smaller effective population size of mtDNA is driving the more recent coalescence times seen in this marker. However, it is important to point out that the population expansion and mtDNA selective sweep hypotheses are not mutually exclusive. Although a population expansion (or reproductive variance) is required as an explanation for the low nuclear diversity observed in the sperm whale, a selective sweep could also act in tandem to further reduce mtDNA diversity. Individuals with a competitive advantage, either through cultural innovations or a beneficial mtDNA mutation, under certain circumstances, could increase in frequency at the edge of a range expansion (Excoffier et al. 2009, Lehe et al. 2012).

5.5.5. Methodological considerations

To avoid biasing comparisons of diversity, we verified variable sites, required a minimum coverage and trimmed reads for quality in our next-generation sequence, and identified and removed sequence affected by numts. Although no numts were detected in

the sperm whale mitogenomic sequence, they appeared to be prevalent in the pygmy sperm whale assemblies. We used primers designed mostly for use in sperm whales, or designed to work across all cetaceans, with no primer pairs designed specifically for use in the pygmy sperm whale (even if single primers were redesigned). The finding of numt contamination, even with the use of long-range PCR to exceed the typical length of mammalian numts (Richly and Leister 2004), indicates that it is more efficient to redesign species-specific primers when attempting to amplify mitogenomes. The unintentional incorporation of numts can lead to anomalous results, including artificially inflating diversity (Dunshea et al. 2008, Sorenson and Quinn 1998, Zhang and Hewitt 1996). However, given our conservative approach of masking variable sites in samples with fewer than 70% of reads supporting a base, it is unlikely that incorporation of numts has artificially increased observed pygmy sperm whale mitogenomic diversity.

For nuclear DNA we also took care to ensure phasing. Computational methods such as *PHASE* can have high degrees of uncertainty as sequence diversity increases. This can complicate downstream analyses of population parameters such as theta (Ruegg et al. 2010). In this study, we implemented a new method of phasing in *PEppyphase,* relying on the 'linkage' of next-generation sequencing reads across multiple variable sites to infer haplotypes. This allowed unambiguous phasing of the majority of the long-range genes sequenced in this study. Other advantages of using next-generation sequencing to characterize nuclear loci include the greater ease of phasing indels in comparison with traditional Sanger sequencing.

5.5.6. Conclusion

The ultimate cause of low mtDNA CR diversity in the sperm whale has remained debatable (Lyrholm et al. 1996, Whitehead 1998), with the additional nuclear loci needed to distinguish alternative hypotheses previously unavailable. Here, we sequenced ~12,000 bp of nuclear DNA for the sperm whale and pygmy sperm whale in order to distinguish between an mtDNA selective sweep, population bottleneck, or population expansion. Sperm whales appear to have had a markedly low, yet stable, effective population size as far back in time as the coalescent events of the genes included in this study can inform. Therefore, it seems unlikely that a recent population bottleneck has affected this species. Instead, the low nuclear and mitogenome diversity observed seems to have been driven by the small long-term effective population size of this species. The inferred recent timing of the population expansion correlates well with expansions in various squid species: primary prey of the sperm whale. In addition, the recent expansion of sperm whales explains the seemingly contradicting results from Chapter 2: strong patterns of mtDNA differentiation within oceans, yet sharing of haplotypes between oceans. The sharing of haplotypes between oceans likely reflects the recent expansion of sperm whales. In the intervening time, high levels of female philopatry have led to the large degrees of mtDNA differentiation currently observed.

5.6. Acknowledgements

The authors thank R. Constantine and K. Thompson for curation of the CeTA database, and New Zealand Department of Conservation staff for collecting NZ samples used in this study; J. Rice for curation of the OMMSN database that supplied the Oregon sample;

B. Mate for access to the Gulf of Mexico samples, the National Oceanic and Atmospheric Administration Southwest Fisheries Science Center for discussions about cetacean next-generation sequencing, M. Dasenko and C. Sullivan (Oregon State University Center for Genomics Research and Biocomputing) for Illumina sequencing and computational support; E. Carroll, A. Liston and K. O'Malley for providing valuable comments on this manuscript, B. Slikas for primer development; J. Jackson for providing additional sequence information from previous publications, and the OSU Cetacean Conservation and Genomics Laboratory for general laboratory support. This work was supported by a Mamie Markham Award and a Lylian Brucefield Reynolds Award from the Hatfield Marine Science Center; and a 2008–2011 International Fulbright Science & Technology award to A.A; and co-funded by the ASSURE program of the Department of Defense in partnership with the National Science Foundation REU Site program to K.H and C.S.B. [grant number NSF OCE-1004947].

Figure 5.1: Species represented in the multispecies analysis of cetacean substitution rates for the loci sequenced in this study.
(a) gives major clades in this study, based on the *CHRNA1* phylogeny. Colored blocks to the right of the phylogeny in (a), correspond to the colored blocks to the left of the species listed in (b). Species are listed in the same order in (a) and (b). Gray branches in (a) correspond to species not represented in the *CHRNA1* phylogeny but represented for other loci. Clades are collapsed or shown as polytomies where they had posterior probabilities less than 0.95. Fossil-calibrated nodes are shown by boxes, with nodes constrained to monophyly shown in gray. Other nodes not fossil-calibrated or constrained to monophyly are shown in ovals.
(b) gives the specific species used in the phylogenies for each locus, with the GenBank accession code for the sequence used to represent each species. For *BRCA1*, *Sotalia fluviatilis* represented "*Sotalia* sp.". References for GenBank sequences used in multispecies phylogeny: (Caballero et al. 2008); (McGowen 2011); (Jackson et al. 2009); (Li et al. 2009b); (Dalebout et al. 2008); (Alter et al. 2007); (Hare and Palumbi 2003); (Vollmer and Rosel 2012); (Meredith et al. 2013); (Dalebout et al. 2014); (Murphy et al. 2001).

Figure 5.1 (Continued)

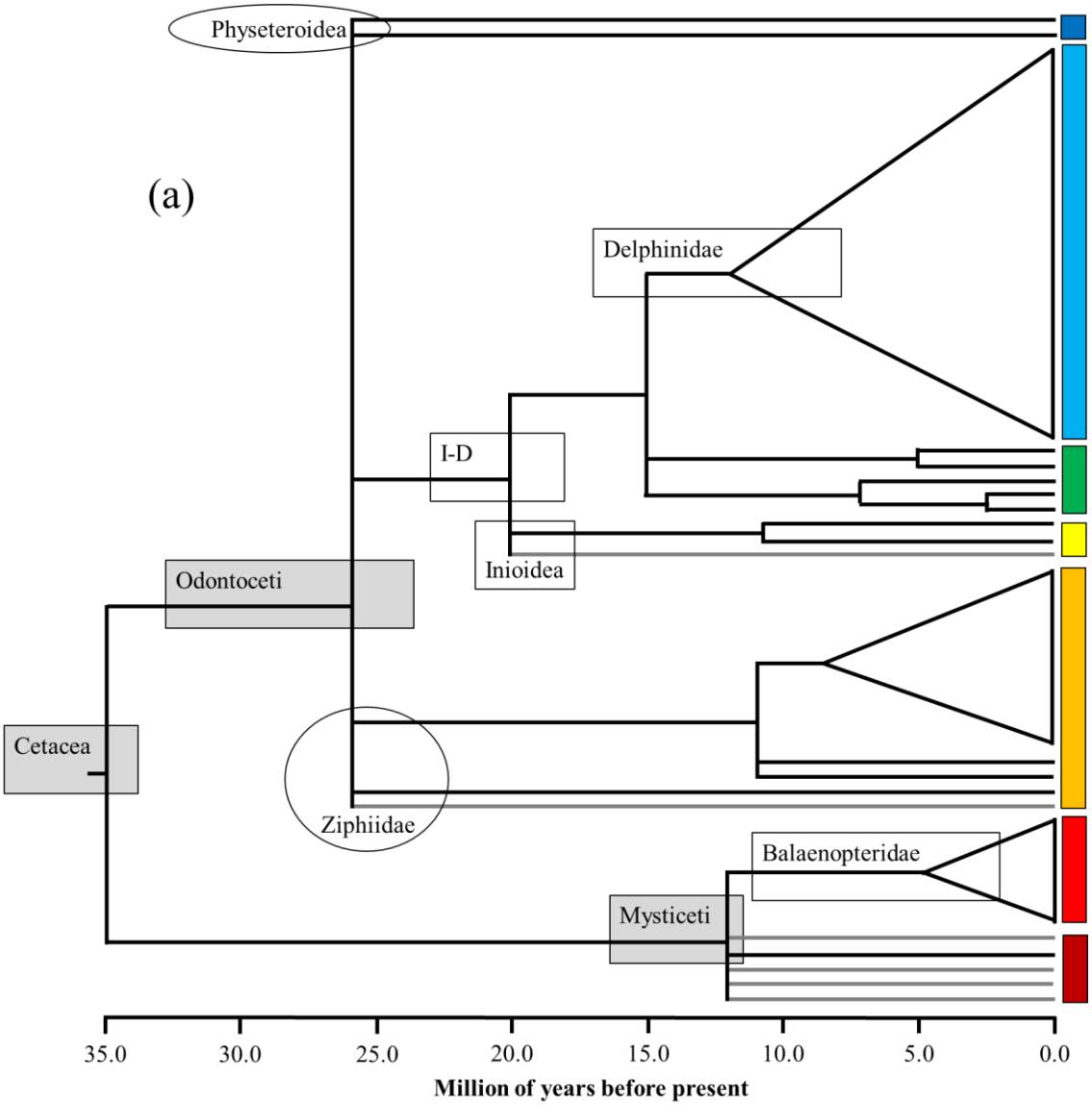


Figure 5.1 (Continued) (b)

Species	*CHRNA1*	*ESD*	*FGG*	*G6PD*	*RHO*	*APOB*	*BRCA1*	*TTN*
Physeter macrocephalus	PmaNZ005	PmaNZ005	PmaNZ005	PmaNZ005	PmaNZ005	PmaNZ005	PmaNZ005	PmaNZ005
Kogia breviceps	KbrNZ090	KbrNZ090	KbrNZ090	KbrNZ090	KbrNZ090	KbrNZ090	KbrNZ090	KbrNZ090
Cephalorhynchus commersonii	EU121141							
Delphinus capensis	JF504823							
Delphinus delphis	EU121135							
Feresa attenuata	JF504827							
Globicephala melas	GQ407769							
Globicephala macrorhynchus	GQ407770							
Grampus griseus	EU121148							
Lagenodelphis hosei	EU121139							
Lagenorhynchus acutus	JF504832							
Lagenorhynchus albirostris	JF504831							
Lagenorhynchus australis	EU121140							
Lagenorhynchus obliquidens	JF504830							
Lagenorhynchus obscurus	JF504829							
Lissodelphis borealis	JF504828							
Orcaella brevirostris	EU121147							
Orcinus orca	EU121149							
Peponocephala electra	EU121145							
Pseudorca crassidens	JF504833							
Sotalia sp.	EF027057*						AY057825	
Sousa chinensis	EF027058							
Stenella coeruleoalba	JF504825							
Stenella attenuata	JF504826							
Stenella frontalis	EU121136							
Stenella longirostris	EU121137							
Steno bredanensis	EF027059							
Tursiops aduncus	JF504824							
Tursiops truncatus	EU121138			JQ002773		JN413972	JN414143	JN415003
Delphinapterus leucas	EU121155					JN413978	JN414149	JN414997
Monodon monoceros	JF504835							
Neophocaena phocaenoides	FJ176338			FJ176352	KC676932			
Phocoena phocoena	EU121153				KC676933	JN413979	JN414150	JN414998
Phocoenoides dalli	EU121151				KC676934			
Inia geoffrensis	EU121156				KC676929	JN413977	JN414148	JN414996
Pontoporia blainvillei	JF504836				KC676937	JN413983	JN414153	JN415002
Platanista gangetica minor					KC676936	JN413982	JN414152	JN415001
Mesoplodon bidens	EU476136				EU476106	JN413976	JN414147	JN414995
Mesoplodon bowdoini	EU476137				EU476107			
Mesoplodon densirostris	EU476139				EU476109			
Mesoplodon europaeus	EU476140				EU476110			
Mesoplodon ginkgodens	EU476141				EU476111			
Mesoplodon grayi	EU476142				EU476112			
Mesoplodon hotaula					KF027317			
Mesoplodon hectori	EU476143				EU476113			
Mesoplodon layardii	EU476144				EU476114			
Mesoplodon mirus	EU476145				EU476115			
Mesoplodon perrini	EU476146				EU476116			
Mesoplodon peruvianus	EU476147				EU476117			
Mesoplodon stejnegeri	EU476148				EU476118			
Mesoplodon carlhubbsi	EU476138				EU476108			
Ziphius cavirostris	GQ407775				EU476119			
Tasmacetus shepherdi	EU476150				EU476120			
Berardius bairdii					KC676925			
Balaenoptera acutorostrata					KC676922			
Balaenoptera bonaerensis	GQ407731	EF043305	GQ407462	GQ407362	GQ408360			
Balaenoptera borealis			EF043312	EF043315				
Balaenoptera edeni				EF043318				
Balaenoptera musculus	GQ407714	EF043308	GQ407458	GQ407364	GQ408357			
Balaenoptera physalus	GQ407713	EF043306	GQ407461	GQ407363	KC676924			
Eschrichtius robustus		EF043303	EF043311	EF043322		JN413975	JN414146	JN414994
Megaptera novaeangliae	GQ407691	EF043304	GQ407365	GQ407320	GQ408259	JN413971	JN414142	JN632909
Caperea marginata					KC676926	JN413973	JN414144	JN414992
Eubalaena australis	GQ407712	GQ407582	GQ407436	GQ407339	GQ408337	JN413974	JN414145	JN414993
Eubalaena glacialis		EF043307	AY196814	AY196816				
Eubalaena japonica			EF043313	EF043321				
Balaena mysticetus				EF043320	KC676921			

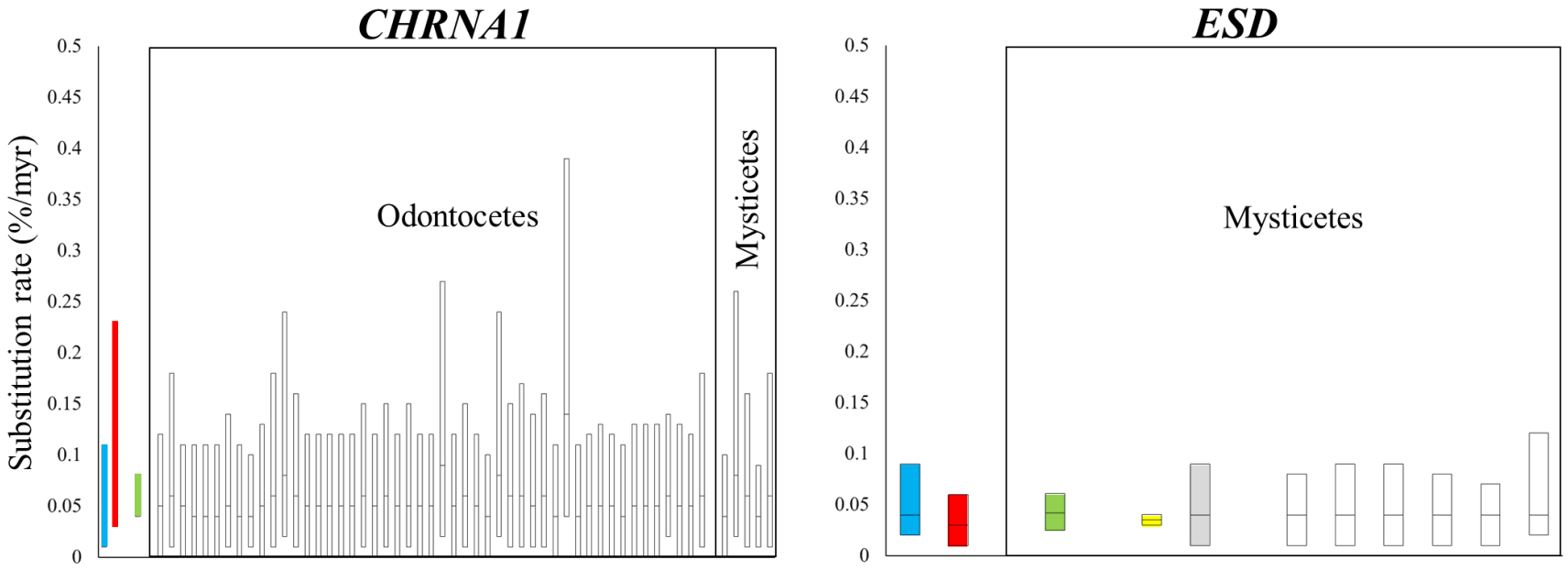


Figure 5.2: Substitution rates for the nuclear loci sequenced in this study obtained from fossil-calibrated phylogenies, showing median and upper and lower 95% highest posterior density (HPD) intervals. The sperm whale is shown in blue and the pygmy sperm whale shown in red. Average rates from Jackson et al. (2009) are shown in green. Where these occur outside of the boxes labelling odontocetes and mysticetes, they apply across all cetaceans. Also given are the gray-whale specific rates from Alter et al. (2007) (shown in yellow) in comparison with rates derived in this study (shown in gray). The Alter et al. (2007) intervals do not reflect 95% HPDs, but 95% confidence intervals constructed using standard errors that makes them likely to differ in range from a fully Bayesian 95% HPD. Rates for the mtDNA CR and protein-coding genes are given in Figure 4.4.

Figure 5.2 (Continued)

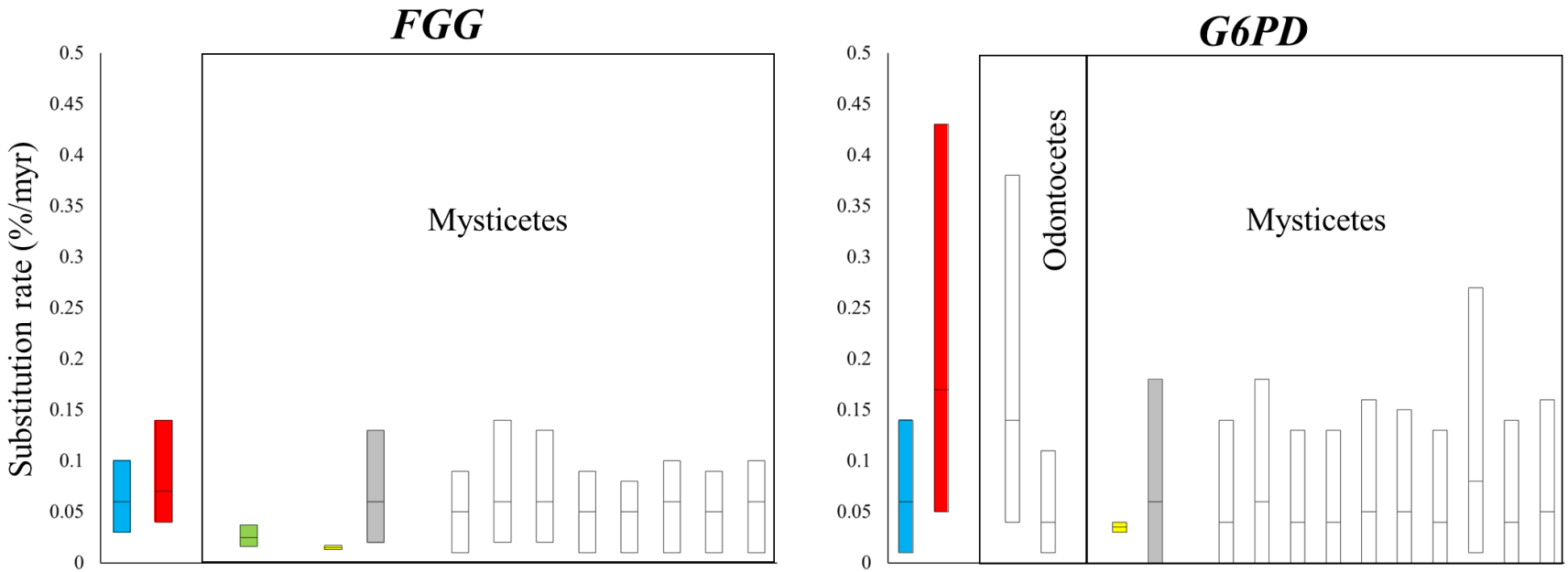


Figure 5.2 (Continued)

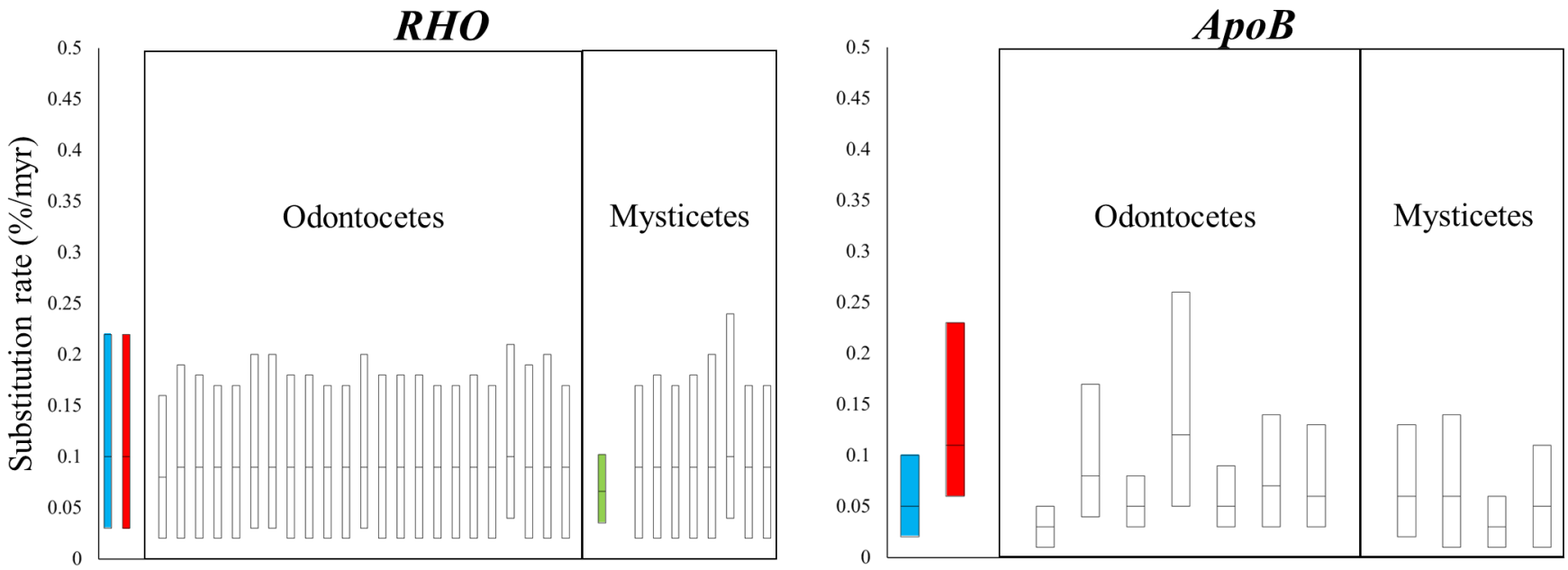
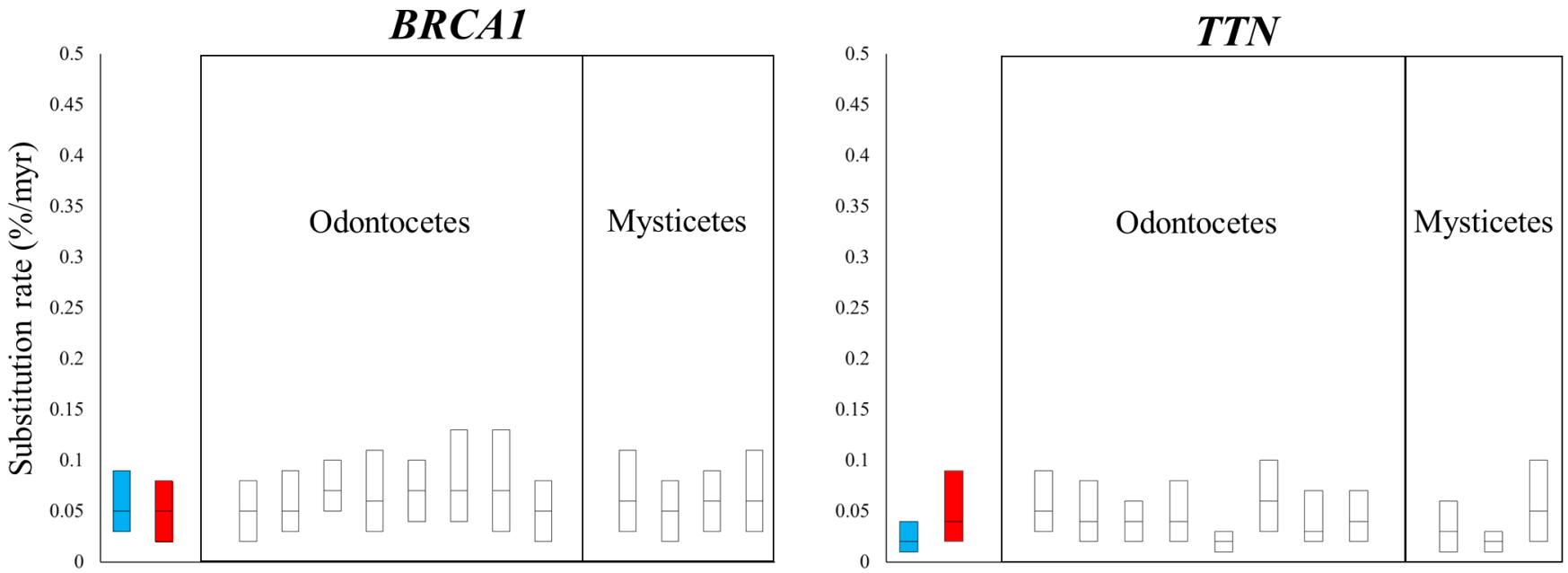


Figure 5.2 (Continued)



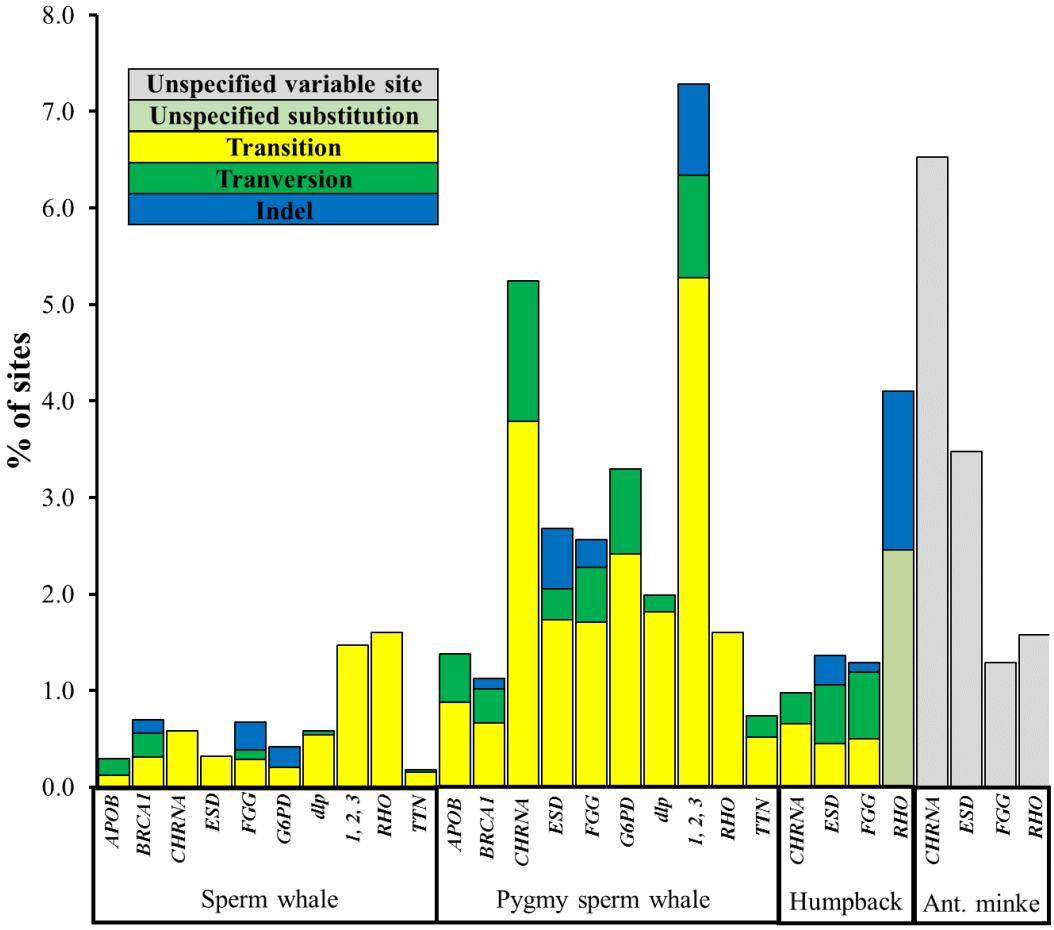


Figure 5.3: Percentage of variable sites by locus for the sperm whale, pygmy sperm whale, humpback, and Antarctic minke. Where available, percentages are broken down into those representing transitions, transversions, and indels. References: Antarctic minke (Ruegg et al. 2010); humpback whale (Jackson et al. 2014). Note, the humpback and Antarctic minke have larger sample sizes (Table 5.5) than the sperm whale and pygmy sperm whale that could increase the likelihood of finding variable sites.

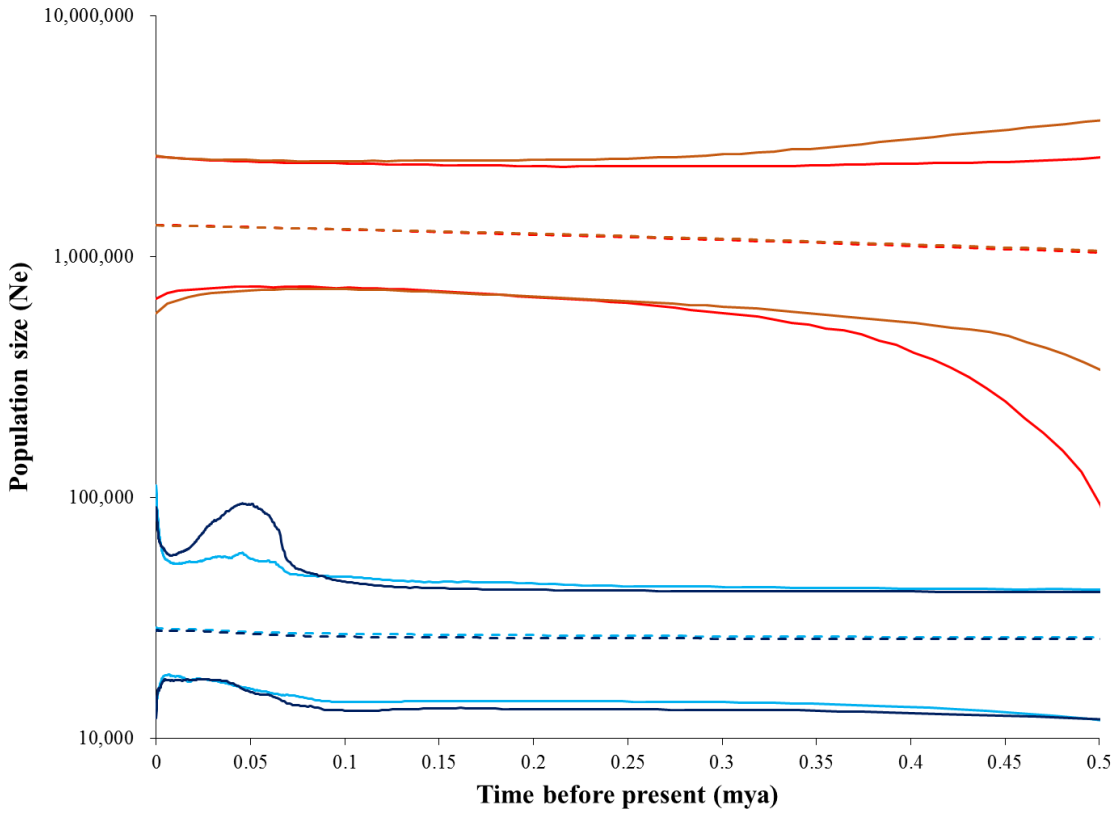


Figure 5.4: Extended Bayesian skyline plots for the sperm whale (blue) and pygmy sperm whale (red/orange) based on nuclear and mtDNA loci. Two runs are presented for each species, shown in slightly different shades of colors. The median demographic trajectory is shown by dotted lines, the 95% highest posterior density (HPD) interval by solid lines. Time before present is restricted to 0.5 mya, as HPD intervals are considerably wider before this time, rendering the plot uninformative. Note, effective population size is displayed on a log scale.

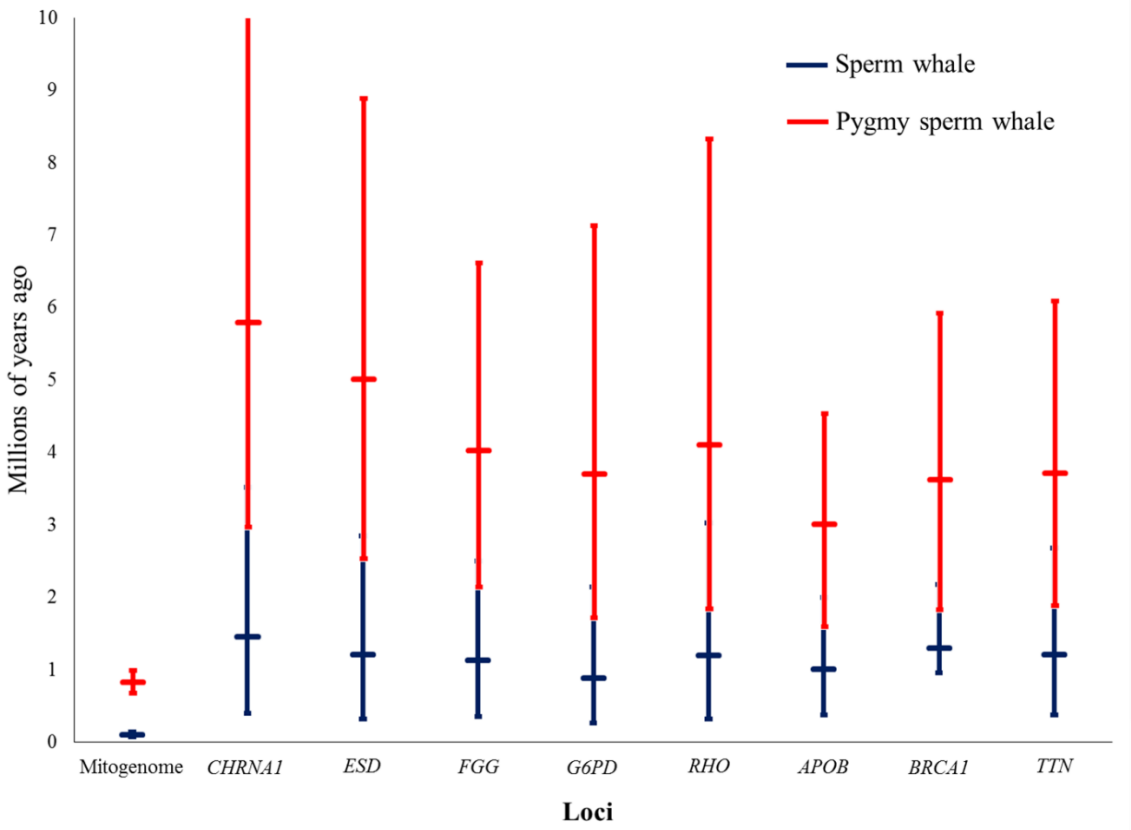


Figure 5.5: Time to most recent common ancestor (or most recent female common ancestor in the case of the mitogenome) for the sperm whale (blue) and pygmy sperm whale (red), based on the extended Bayesian skyline analyses plot implemented with substitution rates obtained from the fossil-calibrated cetacean phylogenies for each locus. The median and 95% highest posterior density intervals are presented.

Figure 5.6: Estimates of effective population size for cetacean species, restricted to species that have estimates of theta available for multiple nuclear loci (and mtDNA where available). Independent estimates of effective population size from these loci (shown in white circles for nuclear DNA, and gray triangles for mtDNA) were used to calculate the lower quartile, median, and upper quartile, with error bars giving the minimum and maximum locus-specific estimates for effective population size. The sperm whale is shown in blue (n = 10 loci, including the mtDNA CR and protein-coding regions), the pygmy sperm whale in red (same loci as for the sperm whale), and the Antarctic minke in purple (n = 11 nuclear loci). For the gray whale, three estimates are given based on: A) N_e calculated using theta estimates for *ESD, FGG, G6PD* and mtDNA *Cytb* from Alter et al. (2007) and substitution rates estimated in this study or by Alexander et al. (2013); B) effective population sizes based on the same four loci taken directly from Alter et al. (2007) and C) effective population sizes taken from Alter et al. (2007) for all 10 loci. For the humpback, estimates are given for the Southern Hemisphere (SH), North Pacific (NP) and North Atlantic (NA) (n = 9 loci, including the mtDNA CR). For loci available for the humpback and Antarctic minke that did not have substitution rates estimated in this study, an average of the intron rates that were estimated was used. References: Antarctic minke (Ruegg et al. 2010), humpback whale (Jackson et al. 2014), eastern gray (Alter et al. 2007).

Figure 5.6 (Continued)

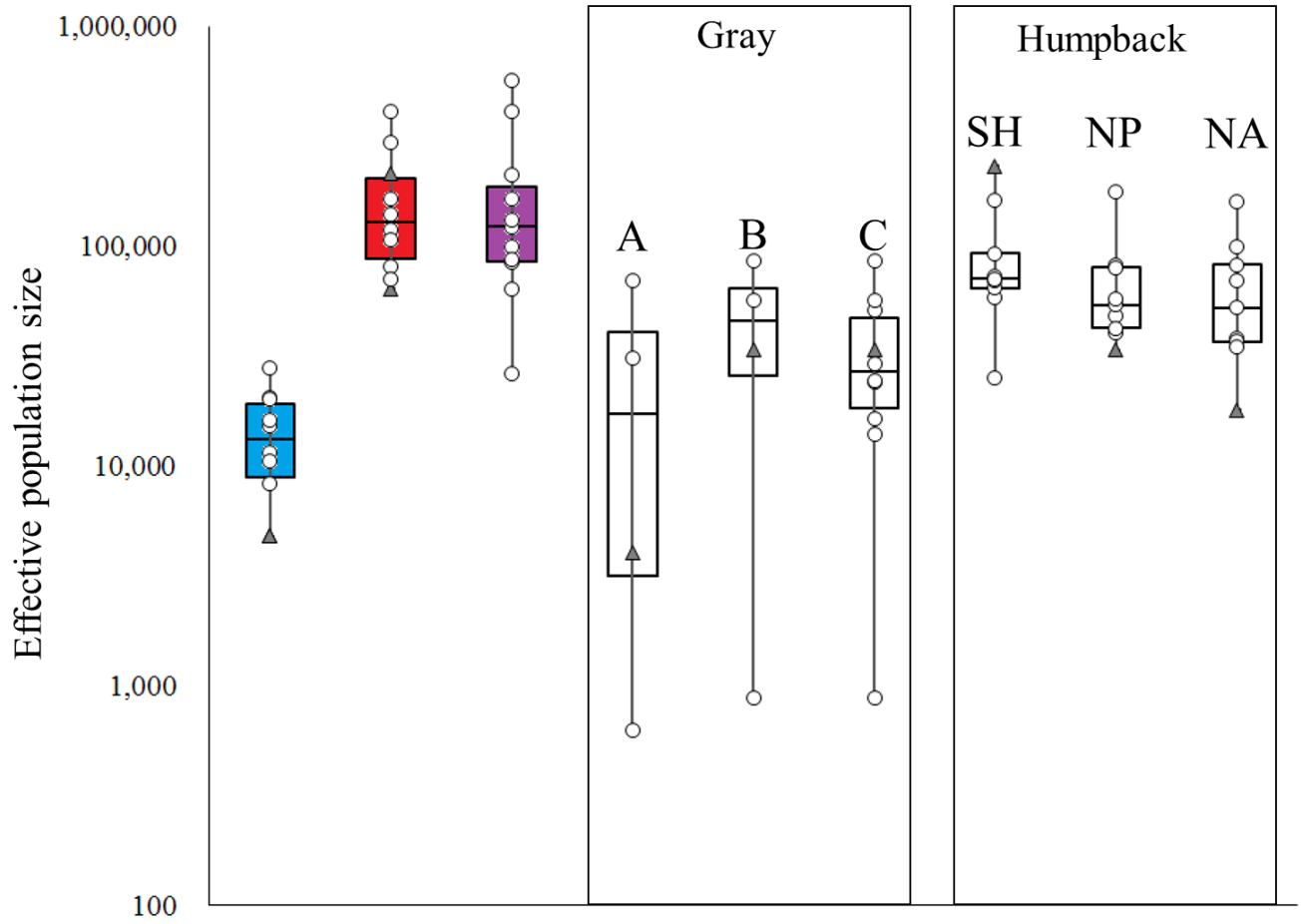


Table 5.1: Time to most recent common ancestor (TMRCA) for mtDNA of different cetacean and squid species, ordered by TMRCA (95% highest posterior density intervals given in parentheses). * might represent species complex rather than individual species. ^ mutation rate, not phylogenetically derived substitution rate, used.

Species	mtDNA regions used	TMRCA (HPD)	Reference
Jumbo squid (*Dosidicus gigas*)	Cytochrome *c* oxidase (CO1) gene	40,000 years ago (32,000 - 48,000)	(Ibáñez et al. 2011)
Patagonian squid (*Doryteuthis gahi*)	Cytochrome *c* oxidase (CO1) gene	100,000 years ago (80,000 - 119,000)	(Ibáñez et al. 2012)
Sperm whale (*Physeter macrocephalus*)	Protein-coding + control region	103,000 years ago (73,000 - 137,000)	(Alexander et al. 2013)
Bowhead whale (*Balaena mysticetus*)^	Control region	153,000 years ago (50,000 - 294,000)	(Ho et al. 2008, Rooney et al. 2001)
Giant squid (*Architeuthis dux*)	Protein-coding	381,000 years ago (32,000 - 730,000)	(Winkelmann et al. 2013)
Killer whale (*Orcinus orca)**	Whole mitogenome	702,000 years ago (489,000 - 956,000)	(Morin et al. 2010a)
Pygmy sperm whale (*Kogia breviceps*)	Protein-coding + control region	830,000 years ago (683,000 - 995,000)	This study
Fin whale (*Balaenoptera physalus*)*	Protein coding	2,080,000 years ago (1,210,000 - 3,040,000)	(Archer et al. 2013)

Table 5.2: The effect of selective sweep, population bottleneck, and population expansion hypotheses on genetic diversity, tests of diversity disequilibrium, demographic reconstructions, and time to most recent common ancestor (TMRCA).

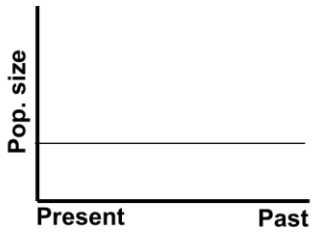
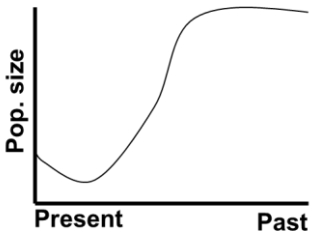
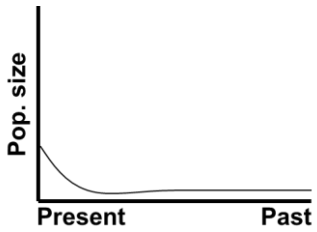
	Loci and metrics	Selective sweep	Population bottleneck	Population expansion
mtDNA	**Diversity**	Low	Low	Low
	Diversity disequilibrium tests	Significantly negative	Significantly negative	Significantly negative
	TMRCA	Young	Young/old (depending on timing and severity of bottleneck event)	Young
Nuclear DNA	**Diversity**	Higher	Low	Low
	Diversity disequilibrium tests	Not significant	Significantly positive or negative (depending on timing of event)	Significantly negative
	TMRCA	Older	Young/old (depending on timing and severity of bottleneck event)	Young
	Demographic reconstructions			

Table 5.3: Measures of genetic diversity for (a) the sperm whale and (b) the pygmy sperm whale. Observed heterozygosity is given at the per-locus level, theta at the per-bp level. Nucleotide diversity is broken down by codon for the protein-coding genes included in this study. For the Tajima's *D* and Fu's F_S tests, * is significant at $p < 0.05$ for Tajima's D, and $p < 0.02$ for Fu's Fs. ** significant at p < 0.001. Significant values bolded and italicized.

(a)	**Mito control region**	**Mito protein-coding regions**	***CHRNA1***	***ESD***	***FGG***	***G6PD***	***RHO***	***APOB***	***BRCA1***	***TTN***
No. of individuals	22	22	24	17	22	22	25	22	22	22
Length (bp)	954	10831	343	631	1046	483	125	2388	2858	4445
Obs. Heterozygosity	--	--	0.5	0.353	0.682	0.136	0.08	0.864	0.909	0.273
Haplotype diversity (s. d.)	0.879 (0.055)	0.957 (0.029)	0.443 (0.066)	0.314 (0.096)	0.714 (0.043)	0.132 (0.068)	0.079 (0.052)	0.776 (0.041)	0.903 (0.26)	0.332 (0.091)
No. of alleles	12	16	3	3	7	3	3	9	17	7
No. of variable sites	14	63	2	2	7	2	2	7	20	8
No. of transitions	14	59	2	2	3	1	2	3	9	7
No. of transversions	0	4	0	0	1	0	2	4	7	1
No. of indels	0	0	0	0	3	1	0	0	4	0
Nucleotide diversity (s. d.)	0.318 (0.192)	1st: 0.070 (0.024) 2nd: 0.050 (0.021) 3rd: 0.224 (0.037) Total: 0.121 (0.063)	0.136 (0.135)	0.052 (0.060)	0.069 (0.058)	0.010 (0.027)	0.065 (0.14)	1st: 0.120 (0.075) 2nd: 0.081 (0.059) 3rd: 0.006 (0.006) Total: 0.069 (0.046)	1st: 0.065 (0.052) 2nd: 0.265 (0.105) 3rd: 0.171 (0.095) Total: 0.167 (0.092)	1st: 0.015 (0.009) 2nd: 0.000 (0.000) 3rd: 0.018 (0.009) Total: 0.011 (0.011)
Theta (per bp×10^4) (s. d.)	40.3 (16.7)	16.0 (5.57)	13.2 (9.65)	7.75 (5.72)	8.80 (4.89)	4.76 (4.76)	35.7 (26.2)	6.74 (3.10)	12.9 (4.76)	4.14 (1.82)
Tajima's *D*	-0.77	-0.972	0.056	-0.651	-0.518	-1.12	***-1.46****	0.05	0.93	***-2.05****
Fu's F_S	***-4.62****	-1.8	0.2	-0.664	-1.09	***-2.15****	***-3.03****	-2.17	-1.78	***-4.94*** **

Table 5.3 (Continued) (b)

(b)	Mito control region	Mito protein-coding regions	*CHRNA1*	*ESD*	*FGG*	*G6PD*	*RHO*	*APOB*	*BRCA1*	*TTN*
No. of individuals	9	9	16	9	14	13	17	10	10	10
Length (bp)	947	7341	343	634	1054	455	125	2388	2849	4445
Obs. Heterozygosity	--	--	0.813	0.889	0.929	0.539	0.118	0.9	1	1
Haplotype diversity (s. d.)	0.833 (0.127)	1.00 (0.052)	0.883 (0.045)	0.941 (0.033)	0.923 (0.032)	0.797 (0.068)	0.219 (0.090)	0.990 (0.019)	0.963 (0.028)	0.938 (0.025)
No. of alleles	6	9	16	11	16	9	3	18	15	16
No. of variable sites	69	146	18	17	26	15	2	32	32	33
No. of transitions	50	133	14	11	18	11	2	21	19	23
No. of transversions	10	13	5	2	6	4	0	12	10	10
No. of indels	9	0	0	4	3	0	0	0	3	0
Nucleotide diversity (s. d.)	0.802 (0.469)	1st: 0.203 (0.279) 2nd: 0.083 (0.027) 3rd: 1.29 (0.137) Total: 0.510 (0.278)	1.11 (0.640)	0.579 (0.344)	0.416 (0.236)	0.492 (0.312)	0.183 (0.244)	1st: 0.286 (0.100) 2nd: 0. 112 (0.059) 3rd: 0.482 (0.131) Total: 0.293 (0.160)	1st: 0.123 (0.054) 2nd: 0.250 (0.085) 3rd: 0.187 (0.070) Total: 0.186 (0.105)	1st: 0.180 (0.054) 2nd: 0.020 (0.012) 3rd: 0.223 (0.062) Total: 0.141 (0.078)
Theta (per bp×10^4) (s. d.)	233 (100)	73.2 (30.8)	130.3 (49.0)	59.6 (25.7)	56.1 (20.6)	86.4 (34.5)	39.1 (28.9)	37.8 (14.1)	28.7 (10.8)	20.9 (7.81)
Tajima's *D*	***-3.48****	***-1.75****	1.68	-0.157	-0.942	-1.51	-1.07	-0.895	-1.38	-1.3
Fu's F_S	1.49	-0.442	***-5.85****	-1.73	***-5.25****	2.12	***-1.35****	***-9.88****	***-5.56****	***-6.74****

Table 5.4: Average sequencing depth by individual sample for the nuclear genes (*ApoB, BRCA1, TTN*) and mitogenomes sequenced with Illumina Miseq technology in this study. Sample codes: sperm whale (Pma), pygmy sperm whale (Kbr), Gulf of Mexico (GMX), Oregon (OR), New Zealand (NZ).

Sample	Mitogenome	*ApoB*	*BRCA1*	*TTN*
Pma11GMX04	6872	7064	2617	4196
Pma11GMX05	7241	7344	7155	6150
Pma11GMX06	7002	7235	6063	5613
Pma11GMX08	6204	7451	6972	6514
Pma11GMX09	6529	7561	7633	7682
Pma11GMX10	7382	7220	6314	5603
Pma11GMX11	4678	4717	3889	1865
Pma11GMX13	7753	7445	7469	7647
Pma11GMX15	4349	5315	4592	2386
PmaNZ005	--	7462	6305	7205
PmaNZ010	--	7449	5648	7521
PmaNZ013	--	7393	7360	7527
PmaNZ015	--	7369	6597	7183
PmaNZ016	--	7515	7425	7653
PmaNZ034	--	6159	6439	6581
PmaNZ036	--	6012	4296	4753
PmaNZ039	--	4687	6636	3636
PmaNZ056	--	6005	6765	2764
PmaNZ058	--	6269	4287	6048
PmaNZ076	--	5264	4518	4870
PmaNZ082	--	6401	4439	4335
PmaOR001	--	7517	7491	7585
Sperm whale average	6446	6675	5950	5696
KbrNZ016	6810	6874	5112	3993
KbrNZ021	7323	7418	7508	5465
KbrNZ060	7171	7524	7365	7623
KbrNZ090	5915	7457	7347	7614
KbrNZ099	7088	7439	6626	6432
KbrNZ107	7363	7252	7422	7546
KbrNZ056	2998	5535	3255	5971
KbrNZ086	2303	4434	5016	6059
KbrNZ093	3134	5415	3389	4562
KbrNZ102	3542	5916	5082	5646
Pygmy sperm whale average	5365	6526	5812	6091
Total average	5851	6626	5904	5828

Table 5.5: A comparison of diversity between the sperm whale, pygmy sperm whale and other cetacean species previously sequenced for the short-range introns characterized in this study. For each species by locus: number of individuals sampled, length of intron (bp), number of alleles (*k*), observed heterozygosity (on a per locus basis), nucleotide diversity (%) and standard error, theta (per bp × 10^4) and standard error. Number of alleles, observed heterozygosity, nucleotide diversity, and theta are shaded from light gray (low diversity) through to dark gray (high diversity). Table sorted by π. Sperm whale highlighted in yellow. References: [1] This study, [2] Jackson et al. (2014); [3] Ruegg et al. (2010); [4] Jackson et al. (2009); [5] Amaral et al. (2012); [6] Alter et al. (2007). Note, common dolphin estimate is based on genus-wide, rather than species-specific sampling. The gray whale is represented by the eastern North Pacific population only (ENP). Other location abbreviations: NZ (New Zealand), GMX (Gulf of Mexico), OR (Oregon), SO (Southern Ocean). For humpback whale theta, the subpopulation with the lowest value (Southern Hemisphere, North Pacific, North Atlantic) is presented. Theta SE approximated for the Antarctic minke and humpback based on ranges presented by Ruegg et al. (2010) and Jackson et al. (2014), respectively.

Species	Location	Locus	*n*	bp	*k*	Observed het	π (%)	π (% SE)	Theta (per bp×10^4)	Theta (SE)	Reference
Humpback	Worldwide	*CHRNA1*	79	306	3	0.190	0.100	0.110	19.0	11.2	[2]
Sperm	NZ, GMX, OR	*CHRNA1*	24	343	3	0.500	0.136	0.135	13.2	9.6	[1]
Pygmy sperm	NZ	*CHRNA1*	16	343	16	0.813	1.113	0.640	130.3	48.9	[1]
Antarctic minke	SO	*CHRNA1*	49	250	16	--	1.352	--	201.0	76.0	[3]
Southern right	Unknown	*CHRNA1*	42	331	5	--	--	--	--	--	[4]
Common*	Worldwide	*CHRNA1*	90	379	26	--	--	--	--	--	[5]
Sperm	NZ, GMX, OR	*ESD*	17	631	3	0.353	0.052	0.060	7.8	5.7	[1]
Humpback	Worldwide	*ESD*	70	661	8	0.670	0.230	0.150	20.0	11.7	[2]
Gray	ENP	*ESD*	72	827	--	--	0.310	--	25.6	--	[6]
Antarctic minke	SO	*ESD*	47	670	25	--	0.572	--	145.0	44.4	[3]
Pygmy sperm	NZ	*ESD*	9	634	11	0.889	0.579	0.344	59.6	25.7	[1]
Southern right	Unknown	*ESD*	46	782	5	--	--	--	--	--	[4]
Sperm	NZ, GMX, OR	*FGG*	22	1046	7	0.682	0.069	0.058	8.8	4.8	[1]
Humpback	Worldwide	*FGG*	71	1008	7	0.590	0.080	0.060	13.0	9.2	[2]
Gray	ENP	*FGG*	72	658	--	--	0.081	--	17.0	--	[6]
Antarctic minke	SO	*FGG*	41	1000	12	--	0.084	--	44.0	16.3	[3]
Pygmy sperm	NZ	*FGG*	14	1054	16	0.929	0.416	0.236	56.1	20.6	[1]

Table 5.5 (Continued)

Species	Location	Locus	*n*	bp	*k*	Observed het	π (%)	π (% SE)	Theta (per bp×10^4)	Theta (SE)	Reference
Southern right	Unknown	*FGG*	44	1008	4	--	--	--	--	--	[4]
Sperm	NZ, GMX, OR	*G6PD*	22	483	3	0.136	0.010	0.027	4.8	4.8	[1]
Pygmy sperm	NZ	*G6PD*	13	455	9	0.539	0.492	0.312	86.4	34.5	[1]
Southern right	Unknown	*G6PD*	19	257	2	--	--	--	--	--	[4]
Humpback	Unknown	*G6PD*	50	257	3	--	--	--	--	--	[4]
Sperm	NZ, GMX, OR	*RHO*	25	125	3	0.080	0.065	0.138	35.7	26.2	[1]
Antarctic minke	SO	*RHO*	52	180	4	--	0.074	--	51.0	34.2	[3]
Pygmy sperm	NZ	*RHO*	17	125	3	0.118	0.183	0.244	39.1	28.9	[1]
Humpback	Worldwide	*RHO*	78	122	10	0.710	1.320	0.870	55.0	30.1	[2]
Southern right	Unknown	*RHO*	40	190	4	--	--	--	--	--	[4]

6. General discussion

6.1. Patterns of genetic differentiation in the sperm whale

In this dissertation, I examined the influence of social structure and molecular evolution on genetic diversity in the sperm whale, using a large, mostly equatorial dataset from the Voyage of the *Odyssey*, supplemented by additional biopsies from the Gulf of Mexico and samples from stranding events. Using these samples I examined forces responsible for structuring genetic diversity in the sperm whale, including female philopatry and male-biased dispersal, the influence of social groups, and how these forces might interact to cause differences in the level of geographic structure by ocean basin. Here, I summarize results and describe implications and future research directions.

6.1.1. The influence of female philopatry and male-biased dispersal and gene flow on genetic structure

In Chapter 2, I confirmed the importance of female philopatry in shaping patterns of genetic differentiation in the sperm whale. The previously un-sampled equatorial regions of the Indian Ocean showed large degrees of maternal structure, similar to the degree seen in the Atlantic Ocean (Chapter 2, Engelhaupt et al. 2009). Patterns of differentiation were particularly pronounced for the Gulf of Mexico, Mediterranean, and interestingly given its lack of geographic isolation, for Sri Lanka. The presence of submarine canyons that could enhance productivity around Sri Lanka suggests female philopatry could be influenced by resource availability (e.g. oceanography). At scales ranging from oceanic to regional comparisons, I also confirmed male-biased dispersal described in previous

studies, and detected strong indications of male-biased gene flow. This was consistent with the low overall levels of geographic structure detected for the bi-parentally inherited microsatellite markers in this study. The marked pattern of female philopatry observed in comparison with males argues for management policies based on female stock-delimitation.

6.1.2. The influence of social groups on genetic structure

In this study, I confirmed strong genetic differentiation among social groups (Chapter 3). As found in previous research, the groups I characterized were mostly not strictly matrilineal. However, I confirmed the importance of matrifocal social structure in influencing mtDNA differentiation through the development of the 'standardized matrilineality index'. This showed that 69% of social groups were more matrilineal than expected by chance. In contrast, elevated relatedness did not appear to explain the level of microsatellite differentiation seen between social groups. Instead, a likely mechanism is different paternal contributions of microsatellite alleles by social group. To further investigate this, long-term genetic-recapture studies on multiple social groups would be needed. Given the long dependence of calves on mothers, the chance of sampling both the calf and mother would be reasonably high. This would increase the likelihood of accurately inferring the paternal genotype of calves, allowing an assessment of whether calves between social groups are more likely to have different fathers than calves within social groups.

The benefits for group-living among female social groups include predator defense (Arnbom et al. 1987), alloparenting/baby-sitting - particularly given the greater susceptibility of calves to predation (Arnbom and Whitehead 1989, Best et al. 1984, Gero et al. 2013a, Gordon 1991, Whitehead 1996), and potentially foraging (Ortega-Ortiz et al. 2012). However, given the presence of related and unrelated individuals within female social groups, including some from different matrilines, it is unclear to what extent sperm whale social groups function by inclusive fitness (Lyrholm et al. 1999), rather than alternative explanations for mutualism such as reciprocal altruism. The results of Chapter 3 showed a minority of groups with high relatedness, yet many groups that were partly matrilineal. This is consistent with previous research. While the close documented associations of individuals (Christal and Whitehead 2001) appears to correlate well with relatedness within some social groups (Gero et al. 2008), in other areas there is a lack of association between kinship and spatial proximity (Ortega-Ortiz et al. 2012).

Gaining an understanding of the mechanisms underlying social group formation is important, as the roles social groups play obviously contribute to the success and survival of individual female sperm whales (Gordon 1991). This means that disruption of social structure could have ramifications at the population level (Gordon 1991, Wade et al. 2012). One potential tool for investigating the mechanisms involved in social group formation is the standardized matrilineality index developed in Chapter 3. This index can be calculated on a per-social group level, and therefore correlated with measures of reproductive success to investigate the potential for inclusive fitness to explain the lenient matrilineality seen in sperm whales.

6.1.3. How does geographic structure vary by ocean in the sperm whale?

In this study, I demonstrated the importance of geographic partitions in explaining genetic diversity at oceanic and within-ocean scales. However, variation in the amount of geographic structure present occurred by ocean. In contrast to the Indian and Atlantic Oceans, I found the Pacific Ocean showed lower overall levels of structure, consistent with previous research (Chapter 2, Lyrholm et al. 1998; Mesnick et al. 2011; Rendell et al. 2012; Whitehead et al. 1998). Chapter 3 also demonstrated that although social group explained more mtDNA variance than geography in the Pacific Ocean, the opposite pattern was observed in the Indian Ocean. Acoustic structure has previously been found to differ between the Pacific Ocean (sympatric acoustic clans) and Atlantic Ocean (allopatric acoustic clans, Whitehead et al. 2012). If female gene flow was constrained to occurring within an acoustic clan, this could explain the different levels of geographically-based genetic structure observed. Given the potential for 'social philopatry' of females to acoustic clan to reduce geographic structure, linked genetic and acoustic work would be a useful future research direction, particularly in the highly structured Indian Ocean. In addition to the Atlantic, the Indian Ocean could provide a contrast to the Pacific in how acoustic clan structure might vary within oceans.

However, the ultimate process driving differences in acoustic and genetic structure between oceans remains unclear. Patterns of acoustic structure might have been disrupted in the Pacific due to whaling, which was more intense in this ocean basin (Whitehead et al. 2012). However, in Chapter 3 I found group size to be significantly larger in the Pacific than in the Indian Ocean and Atlantic Ocean. This could suggest oceanography,

predator defense or culture as alternate explanations (Whitehead et al. 2012). Detecting the mechanisms underlining genetic differentiation is important because a restriction in movement between local populations could indicate that there is a real risk of long term declines in response to threats such as ongoing unregulated direct hunts; interactions with commercial/ military sonar and interactions with fisheries (Gore et al. 2007, Jaquet et al. 2000, Laist et al. 2001, Mate et al. 1994, Moore and Early 2004, Reeves 2002, Whitehead and Rendell 2004).

6.2. Low mtDNA diversity in the sperm whale

During my research into patterns of genetic differentiation in the sperm whale (Chapter 2, 3) I assembled the largest dataset to date for sperm whale mtDNA CR: 1,654 sequences. However, despite the increased number of samples and geographic coverage (including the Indian Ocean), mtDNA CR diversity remained low in the sperm whale compared with many other cetacean species. The second half of my dissertation was concerned with examining hypotheses for the low mtDNA CR diversity in the sperm whale.

6.2.1. Socially-mediated explanations for low mtDNA diversity in the sperm whale?

Although I found lenient matrilineality in the sperm whale in Chapter 3, levels of matrilineality appear too low to be consistent with social groups/units being the unit on which the cultural innovation mtDNA selective sweep or the environmental heterogeneity lineage extinction hypotheses could act. In addition, mass stranding rates appear to be too low in the sperm whale for this to explain the low mtDNA CR diversity seen. However, information on higher-order social structure, such as acoustic clan was not available for

my samples. Large and significant differences between acoustic clans in measures of reproductive success (Marcoux et al. 2007) suggest these could be the unit upon which the cultural innovation selective sweep hypothesis could act (Whitehead 2005). To date, no significant genetic differentiation has been found between acoustic clans (Rendell and Whitehead 2003), however these analyses have been limited by sample size. This is another reason why linked acoustic and genetic work in the Indian Ocean would be of great benefit as a future research direction.

6.2.2. Control-region constraints or slow substitution rates?

By sequencing 17 stranding samples from the Pacific, I established that slow substitution rates were not responsible for the low mtDNA CR diversity in the sperm whale (Alexander et al. 2013 i.e. Chapter 4). The use of Bayesian phylogenetic analyses of substitution rate represented an improvement over previous studies using relative rate tests (Lyrholm et al. 1996, Whitehead 1998), as it has been observed that relative rate tests can lack the ability to detect small rate differences (less than four-fold) between lineages (Bromham et al. 2000). The mtDNA CR showed higher levels of diversity than other regions of the mitogenome, as well as accurately reflecting genealogical relationships shown by the rest of the mitogenome. This indicates that control region constraints are not responsible for the low mtDNA diversity seen in the sperm whale.

6.2.3. A selective sweep, population bottleneck or population expansion?

Chapter 5 generated the first estimate of nuclear DNA diversity for the sperm whale. Using this dataset, I showed that low genetic diversity was not restricted to the

mitogenome in the sperm whale, but rather, is also a feature of the nuclear genome. This suggests that a selective sweep is not the primary force that has reduced diversity in this species. Based on the extended Bayesian skyline plot demographic reconstruction, the pervasively low genetic diversity appears to be due to a population expansion, rather than a population bottleneck. However, a larger number of more variable loci are needed to confirm the timing of such an event. The development of reduced-representation next-generation sequencing approaches offers a potential method for sequencing a large number of independent loci. These loci could then give more information on recent coalescent events to increase the accuracy of demographic reconstructions (Heled and Drummond 2008).

In addition, the dynamics of the sperm whale's recovery from whaling will be of interest, because it could mirror the inferred historical expansion of the sperm whale. The intrinsic rate of increase in sperm whales is extremely low in comparison to other cetacean species (r ~ 0.96%) (Chiquet et al. 2013), suggesting that a considerable amount of time will be needed for sperm whale abundance to recover from whaling (~350 years). Assuming an N_e/N_c ratio of 0.1 (Frankham 1995, Whitehead 1998), the current levels of abundance in the sperm whale could be similar to the levels seen before the historical population expansion of this species.

6.3. The 'big bang/maternal drift' hypothesis

Chapter 5 described how the synthesis of low genetic diversity and female philopatry in the sperm whale could lead to the hypothesis of a 'big bang' expansion followed by

maternally-mediated drift leading to large differences in mtDNA haplotype frequencies between social groups, regions, and oceans. This hypothesis is based on observations of low genetic diversity in the sperm whale (Chapters 2, 4, and 5), indicative of a recent population expansion (Chapter 5), yet high degrees of mtDNA differentiation at the social group level as well as at broader geographic scales (Chapters 2, 3). The 'big bang/maternal drift' hypothesis also helps to describe a puzzling result from Chapter 2: that strong maternal structure was observed within oceans, yet mtDNA haplotypes were shared across oceans suggesting ongoing movement. These shared haplotypes are a signature of a recent expansion in sperm whales, rather than on-going gene flow (Lyrholm and Gyllensten 1998, Lyrholm et al. 1999). Following the population expansion, sperm whales appear to be at an intermediary stage of differentiation where they show strong population structure, yet an absence of phylogeographic structuring of ocean- and region-specific mtDNA lineages (Chapter 2).

6.4. Additional future research questions

Among the regions characterized in this study, the Mediterranean was unique in that only one 400 bp mtDNA CR haplotype, C, was present. The low diversity in the Mediterranean is strongly suggestive of a founder event consisting of one maternal lineage. As additional haplotypes have not been resolved from the Mediterranean, this also implies that the founding event was recent. An estimate of the expected time (*t)* for an additional haplotype to evolve after colonization of the Mediterranean by a single haplotype can be made using the following equation:

$$t = \frac{1}{\mu Nbp}$$

where μ is the mutation rate/site/year of the control region, *N* is the population size, and *bp* is the length (in bp) of the control region considered. The estimated sperm whale-specific substitution rate of the mtDNA control region (2.6E-10, Alexander et al. 2013 i.e. Chapter 4), an estimated population size of 2,500 individuals (Notarbartolo di Sciara et al. 2012), and 400 bp as the length of the control region characterized, gives an estimate of time since colonization of ~3,850 years. Given the issues of time-dependency when using fossil-calibrated rates (Ho et al. 2005), this timing could potentially be more recent. However, although this exercise is useful for establishing a rough time-estimate of the colonization of the Mediterranean, it ignores uncertainty due to sampling (e.g. current mtDNA diversity estimates in the Mediterranean are based on 40 individuals out of an estimated population total of ~2,500), genetic drift affecting the frequency and therefore likelihood of additional haplotypes appearing in a sample, changes in population size through time, and uncertainty in mutation rate. To further investigate the time since colonization of the Mediterranean, a useful approach would be an individual-based model that could incorporate these sources of uncertainty. In addition, current estimates of mtDNA diversity in the Mediterranean are based on just 400 bp of the control region. Sequencing of mitogenomes from the Mediterranean might identify additional diversity that could assist in gauging the time since the colonization event.

One of the most surprising findings of this study is that the cause of the low mitogenomic diversity in the sperm whale (a population expansion) differs from the cause of low

mitogenomic diversity in killer whales (mtDNA CR constraints) (Morin et al. 2010a). Whitehead (1998) originally formulated the ‘cultural innovation’ mtDNA selective-sweep hypothesis in response to the observation that many matrifocal species have low mtDNA diversity. However, the explanation for low mtDNA diversity in these species does not seem to be parsimonious. Instead, it appears that the cause will need to be investigated independently for other matrifocal species with low diversity such as the pilot whales (*Globicephala* sp.) and narwhal (*Monodon monoceros*) (Whitehead 1998).

Bibliography

Aasen E, Medrano JF. 1990. Amplification of the ZFY and ZFX genes for sex identification in humans, cattle, sheep and goats. Nat Biotechnol. 8: 1279-1281.

Adachi J, Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol. 42: 459-468.

Adams L, Rosel P. 2006. Population differentiation of the Atlantic spotted dolphin (*Stenella frontalis*) in the western North Atlantic, including the Gulf of Mexico. Mar Biol. 148: 671-681.

Aguiar D, Istrail S. 2012. HapCompass: A fast cycle basis algorithm for accurate haplotype assembly of sequence data. Journal of Computational Biology. 19: 577-590.

Aguilar A. 1985. Further information on the movements of the sperm whale (*Physeter macrocephalus*) in the North Atlantic. Mammalia. 49: 421-424.

Alexander A et al. 2013. Low diversity in the mitogenome of sperm whales revealed by next-generation sequencing. Genome Biology and Evolution. 5: 113-129.

Alexander AM. 2006. "Oh Mother, When Art Thou?" Low mtDNA diversity in long-finned pilot whales (*Globicephala melas*). BSc(Hons). The University of Auckland, Auckland.

Allen KR. 1980. Size distribution of male sperm whales in the pelagic catches. In: editor(s). *Report of the International Whaling Commission, Special Issue 2.*

Allendorf FW. 1986. Genetic drift and the loss of alleles versus heterozygosity. Zoo Biology. 5: 181-190.

Alter SE, Rynes E, Palumbi SR. 2007. DNA evidence for historic population size and past ecosystem impacts of gray whales. P Natl Acad Sci USA. 104: 15162-15167.

Amaral AR et al. 2012. Influences of past climatic changes on historical population structure and demography of a cosmopolitan marine predator, the common dolphin (genus *Delphinus*). Molecular Ecology. 21: 4854-4871.

Amos B, Schlötterer C, Tautz D. 1993. Social structure of pilot whales revealed by analytical DNA profiling. Science. 260: 670-672.

Amos W. 1999. Culture and genetic evolution in whales. Science. 284: 2055a.

Archer FI et al. 2013. Mitogenomic phylogenetics of fin whales (*Balaenoptera physalus spp.*): genetic evidence for revision of subspecies. PLoS ONE. 8: e63396.

Archie EA, Chiyo PI. 2012. Elephant behaviour and conservation: social relationships, the effects of poaching, and genetic tools for management. Molecular Ecology. 21: 765-778.

Archie EA et al. 2008. Fine-scale population genetic structure in a fission–fusion society. Molecular Ecology. 17: 2666-2679.

Archie EA, Moss CJ, Alberts SC. 2006. The ties that bind: genetic relatedness predicts the fission and fusion of social groups in wild African elephants. Proceedings of the Royal Society B: Biological Sciences. 273: 513-522.

Arnason U, Gullberg A, Gretarsdottir S, Ursing B, Janke A. 2000. The mitochondrial genome of the sperm whale and a new molecular reference for estimating Eutherian divergence dates. J Mol Evol. 50: 569-578.

Arnason U, Gullberg A, Janke A. 2004. Mitogenomic analyses provide new insights into cetacean origin and evolution. Gene. 333: 27-34.

Arnbom T, Papastavrou V, Weilgart LS, Whitehead H. 1987. Sperm whales react to an attack by killer whales. Journal of Mammalogy. 68: 450-453.

Arnbom T, Whitehead H. 1989. Observations on the composition and behaviour of groups of female sperm whales near the Galapagos Islands. Canadian Journal of Zoology. 67: 1-7.

Babraham Bioinformatics. 2013. Trim Galore! v. 0.2.8 available from: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Baker CS et al. 1998. Population structure of nuclear and mitochondrial DNA variation among humpback whales in the North Pacific. Mol Ecol. 7: 695-707.

Baker CS et al. 1993. Abundant mitochondrial DNA variation and world-wide population structure in humpback whales. P Natl Acad Sci USA. 90: 8239-8243.

Baker CS et al. 1994. Hierarchical structure of mitochondrial DNA gene flow among humpback whales *Megaptera novaeangliae,* world-wide. Mol Ecol. 3: 313-327.

Baker CS et al. 2013. Strong maternal fidelity and natal philopatry shape genetic structure in North Pacific humpback whales. Marine Ecology Progress Series. 494: 291-306.

Ballard JWO, Dean MD. 2001. The mitochondrial genome: mutation, selection and recombination. Curr Opin Genet Dev. 11: 667-672.

Banguera-Hinestroza E, Bjørge A, Reid R, Jepson P, Hoelzel A. 2010. The influence of glacial epochs and habitat dependence on the diversity and phylogeography of a coastal dolphin species: *Lagenorhynchus albirostris*. Conserv Genet. 11: 1823-1836.

Bansal V, Bafna V. 2008. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics. 24: i153-i159.

Beale T. 1839. *The natural history of the sperm whale. To which is added, A sketch of a South-sea whaling voyage*. London: John Van Voorst.

Bérubé M et al. 1998. Population genetic structure of North Atlantic, Mediterranean Sea and Sea of Cortez fin whales, *Balaenoptera physalus* (Linnaeus 1758): analysis of mitochondrial and nuclear loci. Mol Ecol. 7: 585-599.

Bérubé M, Jørgensen H, McEwing R, Palsbøll PJ. 2000. Polymorphic di-nucleotide microsatellite loci isolated from the humpback whale, Megaptera novaeangliae. Molecular Ecology. 9: 2181-2183.

Best PB. 1979. Chapter 7: Social organization in sperm whales, *Physeter macrocephalus*. In: Winn, HE, Olla, BL, editor(s). *Behavior of Marine Animals. Volume 3: Cetaceans*. New York-London: Plenum Press.

Best PB, Canham PAS, Macleod N. 1984. Patterns of reproduction in sperm whales, *Physeter macrocephalus*. Report of the International Whaling Commission. Special Issue 6: 51-79.

Birky-Jr CW, Fuerst P, Maruyama T. 1989. Organelle gene diversity under migration, mutation, and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. Genetics. 121: 613-627.

Bouckaert R et al. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol. 10: e1003537.

Boyd IL. 1997. The behavioural and physiological ecology of diving. Trends in Ecology & Evolution. 12: 213-217.

Bromham L, Penny D, Rambaut A, Hendy MD. 2000. The power of relative rates tests depends on the data. J Mol Evol. 50: 296-301.

Brown Gladden JG, Ferguson MM, Clayton JW. 1997. Matriarchal genetic population structure of North American beluga whales *Delphinapterus leucas* (Cetacea: Mondontidae). Mol Ecol. 6: 1033-1046.

Brown SG. 1967. Whale marks recovered in the Antarctic whaling season 1966/1967. Norsk Hvalfangst-Tidende. 56: 122.

Brown SG. 1981. Movements of marked sperm whales in the Southern Hemisphere. Report of the International Whaling Commission. 31: 835-837.

Buchanan FC, Friesen MK, Littlejohn RP, Clayton JW. 1996. Microsatellites from the beluga whale *Delphinapterus leucas*. Molecular Ecology. 5: 571-575.

Caballero S et al. 2008. Molecular systematics of South American dolphins *Sotalia*: Sister taxa determination and phylogenetic relationships, with insights into a multi-locus phylogeny of the Delphinidae. Molecular Phylogenetics and Evolution. 46: 252-268.

Carraher CJF. 2004. Comparative mitogenomics of the Southern Hemisphere dolphin genus *Cephalorhynchus.* MSc Thesis. The University of Auckland, Auckland.

Carroll E et al. 2011. Population structure and individual movement of southern right whales around New Zealand and Australia. Mar Ecol-Prog Ser. 432: 257-268.

Carroll EL et al. 2012. Paternity assignment and demographic closure in the New Zealand southern right whale. Molecular Ecology. 21: 3960-3973.

Cassens I et al. 2005. Evidence for male dispersal along the coasts but no migration in pelagic waters in dusky dolphins (*Lagenorhynchus obscurus*). Mol Ecol. 14: 107-121.

Castellini MA, Somero GN. 1981. Buffering capacity of vertebrate muscle: correlations with potentials for anaerobic function. J Comp Physiol B. 143: 191-198.

Charlesworth B, Charlesworth D, Barton NH. 2003. The effects of genetic and geographic structure on neutral variation. Annu Rev Ecol Evol S. 34: 99-125.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics. 10: 195-205.

Childerhouse SJ, Dawson SM, Slooten E. 1995. Abundance and seasonal residence of sperm whales at Kaikoura, New Zealand. Canadian Journal of Zoology. 73: 723-731.

Chiquet RA, Ma B, Ackleh AS, Pal N, Sidorovskaia N. 2013. Demographic analysis of sperm whales using matrix population models. Ecological Modelling. 248: 71-79.

Chivers SJ et al. 2007. Genetic variation and evidence for population structure in eastern North Pacific false killer whales (*Pseudorca crassidens*). Can J Zoolog. 85: 783-794.

Chivers SJ, Dizon AE, Gearin PJ, Robertson KM. 2002. Small-scale population structure of eastern North Pacific harbour porpoises (*Phocoena phocoena*) indicated by molecular genetic analyses. J Cetacean Res Manage. 4: 111-122.

Chivers SJ, Leduc RG, Robertson KM, Barros NB, Dizon AE. 2005. Genetic variation of *Kogia* spp. with preliminary evidence for two species of *Kogia sima*. Mar Mammal Sci. 21: 619-634.

Christal J, Whitehead H. 1997. Aggregations of mature male sperm whales on the Galápagos Islands breeding ground. Marine Mammal Science. 13: 59-69.

Christal J, Whitehead H. 2001. Social affiliations within sperm whale (*Physeter macrocephalus*) groups. Ethology. 107: 323-340.

Christal J, Whitehead H, Letteval E. 1998. Sperm whale social units: variation and change. Canadian Journal of Zoology. 76: 1431-1440.

Christie MR, Tennessen JA, Blouin MS. 2013. Bayesian parentage analysis with systematic accountability of genotyping error, missing data and false matching. Bioinformatics. 29: 725-732.

Clarke MR. 1972. New technique for the study of sperm whale migration. Nature. 238: 405-406.

Clarke MR, Martins HR, Pascoe P. 1993. The diet of sperm whales (*Physeter macrocephalus* Linnaeus 1758) off the Azores. Philosophical Transactions: Biological Sciences. 339: 67-82.

Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. Molecular Ecology. 9: 1657-1660.

Coakes AK, Whitehead H. 2004. Social structure and mating system of sperm whales off northern Chile. Canadian Journal of Zoology. 82: 1360-1369.

Constantine R et al. 2012. Abundance of humpback whales in Oceania using photo-identification and microsatellite genotyping. Marine Ecology Progress Series. 453: 249-261.

Cronn R et al. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 36: e122.

Dalebout ML et al. 2014. Resurrection of *Mesoplodon hotaula* Deraniyagala 1963: A new species of beaked whale in the tropical Indo-Pacific. Marine Mammal Science. 30: 1081-1108.

Dalebout ML, Steel D, Baker CS. 2008. Phylogeny of the beaked whale genus *Mesoplodon* (Ziphiidae: Cetacea) revealed by nuclear introns: implications for the evolution of male tusks. Systematic Biology. 57: 857-875.

Davies RM. 2009. Custom PERL script, available from http://seqanswers.com/forums/showthread.php?s=361b236c525031c8219a97238f053b06&t=5192.

Dawe EG, Hendrickson LC, Colbourne EB, Drinkwater KF, Showell MA. 2007. Ocean climate effects on the relative abundance of short-finned (*Illex illecebrosus*) and long-finned (*Loligo pealeii*) squid in the northwest Atlantic Ocean. Fisheries Oceanography. 16: 303-316.

de Ruiter JR, Geffen E. 1998. Relatedness of matrilines, dispersing males and social groups in long-tailed macaques (*Macaca fascicularis*). Proceedings of the Royal Society of London. Series B: Biological Sciences. 265: 79-87.

de Silva S, Wittemyer G. 2012. A comparison of social organization in Asian elephants and African savannah elephants. International Journal of Primatology. 33: 1125-1141.

de Vos A et al. 2012. Cetacean sightings and acoustic detections in the offshore waters of Sri Lanka: March-June 2003. Journal of Cetacean Research and Management. 12: 185-193.

Donovan GP. 1991. A review of IWC stock boundaries. Report of the International Whaling Commission (Special Issue 13). 39-68.

Dornburg A, Brandley MC, McGowen MR, Near TJ. 2012. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). Mol Biol Evol. 29: 721-736.

Drouot V et al. 2004a. A note on genetic isolation of Mediterranean sperm whales (*Physeter macrocephalus*) suggested by mitochondrial DNA. Journal of Cetacean Research and Management. 6: 29-32.

Drouot V, Gannier A, Goold JC. 2004b. Summer social distribution of sperm whales (*Physeter macrocephalus*) in the Mediterranean Sea. Journal of the Marine Biological Association of the UK. 84: 675-680.

Drummond A, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7: 214.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4: e88.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 22: 1185-1192.

Duchêne S, Archer FI, Vilstrup J, Caballero S, Morin PA. 2011. Mitogenome phylogenetics: the impact of using single regions and partitioning schemes on topology, substitution rate and divergence time estimation. PLoS ONE. 6: e27138.

Dufault S, Whitehead H. 1998. Regional and group-level differences in fluke markings and notches of sperm whales. Journal of Mammalogy. 79: 514-520.

Dufault S, Whitehead H, Dillon M. 1999. An examination of the current knowledge on the stock structure of sperm whales (*Physeter macrocephalus*) worldwide. Journal of Cetacean Research and Management. 1: 1-10.

Dunshea G et al. 2008. Pseudogenes and DNA-based diet analyses: a cautionary tale from a relatively well sampled predator-prey system. B Entomol Res. 98: 239-248.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources. 4: 359-361.

Edgar R. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5: 113.

Eggert LS, Rasner CA, Woodruff DS. 2002. The evolution and phylogeography of the African elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers. Proceedings of the Royal Society of London. Series B: Biological Sciences. 269: 1993-2006.

Engelhaupt D et al. 2009. Female philopatry in coastal basins and male dispersion across the North Atlantic in a highly mobile marine species, the sperm whale (*Physeter macrocephalus*). Mol Ecol. 18: 4193-4205.

Engelhaupt DT. 2004. Phylogeography, kinship and molecular ecology of sperm whales (*Physeter macrocephalus*). University of Durham, Durham City.

Escorza-Treviño S, Archer F, Rosales M, Lang A, Dizon A. 2005. Genetic differentiation and intraspecific structure of Eastern Tropical Pacific spotted dolphins, *Stenella attenuata*, revealed by DNA analyses. Conserv Genet. 6: 587-600.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology. 14: 2611-2620.

Evans K, Morrice M, Hindell M, Thiele D. 2002. Three mass strandings of sperm whales (*Physeter macrocephalus*) in southern Australian waters. Marine Mammal Science. 18: 622-643.

Excoffier L, Foll M, Petit RJ. 2009. Genetic Consequences of Range Expansions. Annual Review of Ecology, Evolution, and Systematics. 40: 481-501.

Excoffier L, Laval G, Schneider S. 2005. Arlequin ver 3.0: an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online. 1: 47-50.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 164: 1567-1587.

Farris JS. 1989. The retention index and homoplasy excess. Syst Biol. 38: 406-407.

Farris JS, Källersjö M, Kluge AG, Bult C. 1995. Testing significance of incongruence. Cladistics. 10: 315-319.

Felsenstein J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Systematic Zoology. 27: 401-410.

Fernando P, Lande R. 2000. Molecular genetic and behavioral analysis of social organization in the Asian elephant (*Elephas maximus*). Behavioral Ecology and Sociobiology. 48: 84-91.

Fordyce RE, Barnes LG. 1994. The evolutionary history of whales and dolphins. Annual Review of Earth and Planetary Sciences. 22: 419-455.

Frankham R. 1995. Effective population size/adult population size ratios in wildlife: a review. Genetics Research. 66: 95-107.

Fu Y-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 147: 915-925.

Garrigue C, Dodemont R, Steel D, Baker CS. 2004. Organismal and 'gametic' capture-recapture using microsatellite genotyping confirm low abundance and reproductive autonomy of humpback whales on the wintering grounds of New Caledonia. Marine Ecology Progress Series. 274: 251-262.

Gero S, Engelhaupt D, Whitehead H. 2008. Heterogeneous social associations within a sperm whale, *Physeter macrocephalus*, unit reflect pairwise relatedness. Behavioral Ecology and Sociobiology. 63: 143-151.

Gero S, Gordon J, Whitehead H. 2013a. Calves as social hubs: dynamics of the social network within sperm whale units. Proceedings of the Royal Society B: Biological Sciences. 280: 20131113 (20131111-20131119).

Gero S et al. 2013b. Behavior and social structure of the sperm whales of Dominica, West Indies. Marine Mammal Science. n/a: n/a-n/a.

Gilbert KJ et al. 2012. Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program structure. Molecular Ecology. 21: 4925-4930.

Gilis D, Massar S, Cerf NJ, Rooman M. 2001. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biol. 2: research0049.0041–0049.0012.

Gobush KS, Wasser SK. 2009. Behavioural correlates of low relatedness in African elephant core groups of a poached population. Animal Behaviour. 78: 1079-1086.

Godard C, Clark R, Kerr I, Madsen PT, Payne R. 2003. Preliminary report on the sperm whale data collected during the *Voyage of the Odyssey*. Paper SC/55/O17 presented to the IWC Scientific Committee, May 2003, Berlin. 8pp.

Goloboff PA, Farris JS, Nixon KC. 2008. TNT, a free program for phylogenetic analysis. Cladistics. 24: 774-786.

Gordon JCD. 1987. Sperm whale groups and social behavior observed off Sri Lanka. Report of the International Whaling Commission. 37: 205-217.

Gordon JCD. 1991. The World Wildlife Fund's Indian Ocean sperm whale project: an example of cetacean research within the Indian Ocean Sanctuary. In: Leatherwood, S, Donovan, GP, editor(s). *Cetaceans and cetacean research in the Indian Ocean Sanctuary*. Nairobi, Kenya: UNEP Marine mammal technical report no. 3.

Gore MA et al. 2007. Sperm whale, *Physeter macrocephalus*, stranding on the Pakistani coast. Journal of the Marine Biological Association of the UK. 87: 363-364.

Gosho ME, Rice DW, Breiwick JM. 1984. The sperm whale, *Physeter macrocephalus*. Marine Fisheries Review. 46: 54-64.

Goto H et al. 2011. Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome Biol. 12: R59.

Goudet J. 2001. FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available from http://www2.unil.ch/popgen/softwares/fstat.htm.

Goudet J, Perrin N, Waser P. 2002. Tests for sex-biased dispersal using bi-parentally inherited genetic markers. Molecular Ecology. 11: 1103-1114.

Guttman DS, Dykhuizen DE. 1994. Detecting selective sweeps in naturally occurring *Escherichia coli*. Genetics. 138: 993-1003.

Hage P, Marck J. 2003. Matrilineality and the Melanesian origin of Polynesian Y chromosomes. Current Anthropology. 44: S121-S127.

Hamner R, Pichler F, Heimeier D, Constantine R, Baker C. 2012. Genetic differentiation and limited gene flow among fragmented populations of New Zealand endemic Hector's and Maui's dolphins. Conserv Genet. 13: 987-1002.

Hamner RM. 2014. All in a DNA's work: conservation genetics and monitoring of the New Zealand endemic Maui's and Hector's dolphins. PhD dissertation. Oregon State University, Corvallis.

Hare MP, Palumbi SR. 2003. High intron sequence conservation across three mammalian orders suggests functional constraints. Molecular Biology and Evolution. 20: 969-978.

Hayano A, Amano M, Miyazaki N. 2003. Phylogeography and population structure of the Dall's porpoise, *Phocoenoides dalli*, in Japanese waters revealed by mitochondrial DNA. Genes Genet Syst. 78: 81-91.

Hedrick PW, Allendorf FW, Baker CS. 2013. Estimation of male gene flow from measures of nuclear and female genetic differentiation. Journal of Heredity. 104: 713-717.

Heled J, Drummond A. 2008. Bayesian inference of population size history from multiple loci. BMC Evolutionary Biology. 8: 289.

Hendrickson SL et al. 2003. Low genetic variability in the geographically widespread Andean condor. The Condor. 105: 1-12.

Ho SYW, Lanfear R. 2010. Improved characterisation of among-lineage rate variation in cetacean mitogenomes using codon-partitioned relaxed clocks. Mitochondr DNA. 21: 138-146.

Ho SYW, Phillips MJ, Cooper A, Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Molecular Biology and Evolution. 22: 1561-1568.

Ho SYW, Saarma U, Barnett R, Haile J, Shapiro B. 2008. The effect of inappropriate calibration: three case studies in molecular ecology. PLoS ONE. 3: e1615.

Hoelzel AR, Hancock JM, Dover GA. 1991. Evolution of the cetacean mitochondrial D-loop region. Mol Biol Evol. 8: 475-493.

Hoelzel AR et al. 2002. Low worldwide genetic diversity in the killer whale (*Orcinus orca*): implications for demographic history. P Roy Soc B-Biol Sci. 269: 1467-1473.

Hoelzel AR et al. 2007. Evolution of population structure in a highly social top predator, the killer whale. Molecular Biology and Evolution. 24: 1407-1415.

Holekamp KE, Smith JE, Strelioff CC, Van Horn RC, Watts HE. 2012. Society, demography and genetic structure in the spotted hyena. Molecular Ecology. 21: 613-632.

Ibáñez CM et al. 2012. Spatial genetic structure and demographic inference of the Patagonian squid *Doryteuthis gahi* in the south-eastern Pacific Ocean. Journal of the Marine Biological Association of the United Kingdom. 92: 197-203.

Ibáñez CM et al. 2011. Genetic diversity and demographic history of *Dosidicus gigas* (Cephalopoda: Ommastrephidae) in the Humboldt Current System. Mar Ecol Prog Ser. 431: 163-171.

International Whaling Commission. 1980. Report of the special meeting on sperm whale assessments 27 November to 8 December 1978. Report of the International Whaling Commission, Special Issue 2. 107-136.

Ivashchenko YV, Clapham PJ, Brownell Jr. RL. 2013. Soviet catches of whales in the North Pacific: revised totals. Journal of Cetacean Research and Management. 13: 59-71.

Ivashin MV. 1981. Some results of the marking of sperm whales (*Physeter macrocephalus*) in the Southern Hemisphere under the Soviet marking programme. Report of the International Whaling Commission. 31: 707-718.

Ivashin MV, Rovnin AA. 1967. Some results of the Soviet whale marking in the waters of the North Pacific. Norsk Hvalfangst-Tidende. 56: 123-135.

Jackson JA et al. 2009. Big and slow: phylogenetic estimates of molecular evolution in baleen whales (Suborder Mysticeti). Mol Biol Evol. 26: 2427-2440.

Jackson JA et al. 2014. Global diversity and oceanic divergence of humpback whales (*Megaptera novaeangliae*). Proceedings of the Royal Society B: Biological Sciences. 281: 20133222.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 23: 1801-1806.

Janik VM. 2001. Is cetacean social learning unique? Behavioral and Brain Sciences. 24: 337-338.

Jaquet N. 1996. How spatial and temporal scales influence understanding of sperm whale distribution: a review. Mammal Review. 26: 51-65.

Jaquet N, Dawson S, Slooten E. 2000. Seasonal distribution and diving behaviour of male sperm whales off Kaikoura: foraging implications. Canadian Journal of Zoology. 78: 407-419.

Jaquet N, Gendron D, Coakes A. 2003. Sperm whales in the Gulf of California: residency, movements, behavior, and the possible influence of variation in food supply. Marine Mammal Science. 19: 545-562.

Jennings TN, Knaus BJ, Mullins TD, Haig SM, Cronn RC. 2011. Multiplexed microsatellite recovery using massively parallel sequencing. Mol Ecol Resour. 11: 1060-1067.

Jepson PD et al. 2003. Gas-bubble lesions in stranded cetaceans. Nature. 425: 575-576.

Johnson ET, Snyder-Mackler N, Beehner JC, Bergman TJ. 2014. Kinship and dominance rank influence the strength of social bonds in female geladas (*Theropithecus gelada*). International Journal of Primatology. 35: 288-304.

Jost L. 2008. G_{ST} and its relatives do not measure differentiation. Mol Ecol. 17: 4015-4026.

Kalinowski ST, Taper ML, Marshall TC. 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular Ecology. 16: 1099-1006.

Kaminski G, Brandt S, Baubet E, Baudoin C. 2005. Life-history patterns in female wild boars (*Sus scrofa*): mother-daughter postweaning associations. Canadian Journal of Zoology. 83: 474-480.

Kanda N, Goto M, Kato H, McPhee M, Pastene L. 2007. Population genetic structure of Bryde's whales (*Balaenoptera brydei*) at the inter-oceanic and trans-equatorial levels. Conserv Genet. 8: 853-864.

Karl SA, Toonen RJ, Grant WS, Bowen BW. 2012. Common misconceptions in molecular ecology: echoes of the modern synthesis. Mol Ecol. 21: 4171-4189.

Kasuya T, Miyashita T. 1988. Distribution of sperm whale stocks in the North Pacific. The Scientific reports of the Whales Research Institute. 39: 31-75.

Kerth G, Mayer F, König B. 2000. Mitochondrial DNA (mtDNA) reveals that female Bechstein's bats live in closed societies. Molecular Ecology. 9: 793-800.

Kim Y, Stephan W. 2003. Selective sweeps in the presence of interference among partially linked loci. Genetics. 164: 389-398.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16: 111-120.

Kircher M, Kelso J. 2010. High-throughput DNA sequencing – concepts and limitations. BioEssays. 32: 524-536.

Kluge AG, Farris JS. 1969. Quantitative phyletics and the evolution of Anurans. Syst Biol. 18: 1-32.

Knaus BJ, Cronn R, Liston A, Pilgrim K, Schwartz MK. 2011. Mitochondrial genome sequences illuminate maternal lineages of conservation concern in a rare carnivore. BMC Ecol. 11: 10.

Kooyman GL, Ponganis PJ. 1998. The physiological basis of diving to depth: birds and mammals. Annu Rev Physiol. 60: 19-32.

Kreitman M. 2000. Methods to detect selection in populations with applications to the human. Annual Review of Genomics and Human Genetics. 1: 539-559.

Krützen M, Sherwin WB, Berggren P, Gales N. 2004. Population structure in an inshore cetacean revealed by microsatellite and mtDNA analysis: bottlenose dolphins (*Tursiops* sp.) in Shark Bay, Western Australia. Mar Mammal Sci. 20: 28-47.

Laist DW, Knowlton AR, Mead JG, Collet AS, Podesta M. 2001. Collisions between ships and whales. Marine Mammal Science. 17: 35-75.

Lambertsen RH. 1987. A Biopsy system for large whales and its use for cytogenetics. Journal of Mammalogy. 68: 443-445.

Lawrence KT, Liu Z, Herbert TD. 2006. Evolution of the Eastern Tropical Pacific through Plio-Pleistocene glaciation. Science. 312: 79-83.

Lehe R, Hallatschek O, Peliti L. 2012. The rate of beneficial mutations surfing on the wave of a range expansion. PLoS Comput Biol. 8: e1002447.

Letteval E et al. 2002. Social structure and residency in aggregations of male sperm whales. Canadian Journal of Zoology. 80: 1189-1196.

Lewis PO, Zaykin D. 2001. Genetic Data Analysis: Computer program for the analysis of allelic data v 1.0, available from http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25: 1754-1760.

Li H et al. 2009a. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25: 2078-2079.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18: 1851-1858.

Li M et al. 2010. Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet. 87: 237-249.

Li M, Schroeder R, Ko A, Stoneking M. 2012. Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs. Nucleic Acids Res. Advanced access.

Li S, Wan H, Ji H, Zhou K, Yang G. 2009b. SNP discovery based on CATS and genotyping in the finless porpoise (*Neophocaena phocaenoides*). Conservation Genetics. 10: 2013-2019.

Li X et al. 2011. Analysis of global and local population stratification of finless porpoises *Neophocaena phocaenoides* in Chinese waters. Mar Biol. 158: 1791-1804.

Lyrholm T, Gyllensten U. 1998. Global matrilineal population structure in sperm whales as indicated by mitochondrial DNA sequences. P Roy Soc B-Biol Sci. 265: 1679-1684.

Lyrholm T, Leimar O, Gyllensten U. 1996. Low diversity and biased substitution patterns in the mitochondrial DNA control region of sperm whales: implications for estimates of time since common ancestry. Mol Biol Evol. 13: 1318-1326.

Lyrholm T, Leimar O, Johanneson B, Gyllensten U. 1999. Sex-biased dispersal in sperm whales: contrasting mitochondrial and nuclear genetic structure of global populations. P Roy Soc B-Biol Sci. 266: 347-347.

Macé M, Crouau-Roy B. 2008. A highly polymorphic insertion in the Y-chromosome amelogenin gene can be used for evolutionary biology, population genetics and sexing in *Cetacea* and *Artiodactyla*. BMC Genetics. 9: 64.

Maddison WP, Maddison DR. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75. Available from http://mesquiteproject.org.

Madsen O et al. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature. 409: 610-614.

Madsen PT et al. 2002. Sperm whale sound production studied with ultrasound time/depth-recording tags. Journal of Experimental Biology. 205: 1899-1906.

Marcoux M, Rendell L, Whitehead H. 2007. Indications of fitness differences among vocal clans of sperm whales. Behavioral Ecology and Sociobiology. 61: 1093-1098.

Martien KK et al. 2012. Population structure of island-associated dolphins: evidence from mitochondrial and microsatellite markers for common bottlenose dolphins (*Tursiops truncatus*) around the main Hawaiian Islands. Mar Mammal Sci. 28: E208-E232.

Martin AP, Palumbi SR. 1993. Body size, metabolic rate, generation time, and the molecular clock. Proc Natl Acad Sci USA. 90: 4087-4091.

Martin AR. 1980. An examination of sperm whale age and length data from the 1949-78 Icelandic catch. Report of the International Whaling Commission. 30: 227-231.

Maruyama T, Fuerst P. 1985. Population bottlenecks and nonequilibrium models on population genetics. III. Genic homozygosity in population which experience periodic bottlenecks. Genetics. 111: 691-703.

Mate BR, Stafford KM, Ljungblad DK. 1994. A change in sperm whale (*Physeter macrocephalus*) distribution correlated to seismic surveys in the Gulf of Mexico. The Journal of the Acoustical Society of America. 96: 3268-3269.

May-Collado LJ, Agnarsson I. 2006. Cytochrome *b* and Bayesian inference of whale phylogeny. Molecular Phylogenetics and Evolution. 38: 344-354.

McGowen MR. 2011. Toward the resolution of an explosive radiation - A multilocus phylogeny of oceanic dolphins (Delphinidae). Molecular Phylogenetics and Evolution. 60: 345-357.

Meirmans PG, Hedrick PW. 2011. Assessing population structure: F_{ST} and related measures. Molecular Ecology Resources. 11: 5-18.

Meirmans PG, van Tienderen PH. 2004. GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. Molecular Ecology Notes. 4: 792-794.

Melville H. 1851. *Moby-Dick; or, The Whale*. New York: Harper and Brothers.

Meredith RW, Gatesy J, Emerling CA, York VM, Springer MS. 2013. Rod monochromacy and the coevolution of cetacean retinal opsins. PLoS Genet. 9: e1003432.

Meredith RW et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. Science. 334: 521-524.

Mesnick SL. 2001. Genetic relatedness in sperm whales: evidence and cultural implications. Behavioral and Brain Sciences. 24: 346-347.

Mesnick SL, Engelhaupt D, Lyrholm T, Taylor BL, Escorza-Trevino S. 2005. Population structure and the social animal: global genetic differentiation in sperm whales, *Physeter macrocephalus*. Cachalot Assessment Research Planning (CARP) Workshop.

Mesnick SL et al. 2011. Sperm whale population structure in the eastern and central North Pacific inferred by the use of single-nucleotide polymorphisms, microsatellites and mitochondrial DNA. Mol Ecol Resour. 11: 278-298.

Mesnick SL et al. 1999. Culture and genetic evolution in whales. Science. 284: 2055a.

Milinkovitch MC. 1995. Molecular phylogeny of cetaceans prompts revision of morphological transformations. Trends in Ecology & Evolution. 10: 328-334.

Milinkovitch MC, Ortí G, Meyer A. 1993. Revised phylogeny of whales suggested by mitochondrial ribosomal DNA sequences. Nature. 346-348.

Miller MP. 2005. Alleles In Space (AIS): computer software for the joint analysis of interindividual spatial and genetic information. Journal of Heredity. 96: 722-724.

Mirimin L et al. 2009. Population structure of short-beaked common dolphins (*Delphinus delphis*) in the North Atlantic Ocean as revealed by mitochondrial and nuclear genetic markers. Mar Biol. 156: 821-834.

Mizroch SA, Rice DW. 2013. Ocean nomads: distribution and movements of sperm whales in the North Pacific shown by whaling data and Discovery marks. Marine Mammal Science. 29: E136-E165.

Møhl B, Wahlberg M, Madsen PT, Miller LA, Surlykke A. 2000. Sperm whale clicks: Directionality and source level revisited. The Journal of the Acoustical Society of America. 107: 638-648.

Möller LM. 2014. Sociogenetic structure, kin associations and bonding in delphinids. Molecular Ecology. 21: 745-764.

Monmonier MS. 1973. Maximum-difference barriers: an alternative numerical regionalization method. Geographical Analysis. 5: 245-261.

Moore MJ, Early GA. 2004. Cumulative sperm whale bone damage and the bends. Science. 306: 2215-2215.

Moors-Murphy HB. 2014. Submarine canyons as important habitat for cetaceans, with special reference to the Gully: A review. Deep Sea Research Part II: Topical Studies in Oceanography. 104: 6-19.

Morin PA, Aitken NC, Rubio-Cisneros N, Dizon AE, Mesnick S. 2007. Characterization of 18 SNP markers for sperm whale (*Physeter macrocephalus*). Molecular Ecology Notes. 7: 626-630.

Morin PA et al. 2010a. Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. Genome Res. 20: 908-916.

Morin PA et al. 2010b. Applied conservation genetics and the need for quality control and reporting of genetic data used in fisheries and wildlife management. J Hered. 101: 1-10.

Moyer GR, Winemeller KO, McPhee MV, Turner TF. 2005. Historical demography, selection, and coalescence of mitochondrial and nuclear genes in *Prochilodus* species of northern South America. Evolution. 59: 599-610.

Mulligan CJ, Kitchen A, Miyamoto MM. 2006. Comment on "Population size does not influence mitochondrial genetic diversity in animals". Science. 314: 1390.

Murphy WJ et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science. 294: 2348-2351.

Nabholz B, Glémin S, Galtier N. 2008. Strong variations of mitochondrial mutation rate across mammals - the longevity hypothesis. Mol Biol Evol. 25: 120-130.

Nabholz B, Jarvis ED, Ellegren H. 2010. Obtaining mtDNA genomes from next-generation transcriptome sequencing: A case study on the basal Passerida (Aves: Passeriformes) phylogeny. Mol Phylogenet Evol. 57: 466-470.

Natoli A, Birkun A, Aguilar A, Lopez A, Hoelzel AR. 2005. Habitat structure and the dispersal of male and female bottlenose dolphins (*Tursiops truncatus*). P Roy Soc B-Biol Sci. 272: 1217-1226.

Nei M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press.

Nei M, Maruyama T, Chakraborty R. 1975. The bottleneck effect and genetic variability in populations. Evolution. 29: 1-10.

Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. Heredity. 86: 641-647.

Nikaido M et al. 2001. Evolution of CHR-2 SINEs in cetartiodactyl genomes: possible evidence for the monophyletic origin of toothed whales. Mammalian Genome. 12: 909-915.

Nishida S, Goto M, Pastene LA, Kanda N, Koike H. 2007. Phylogenetic relationships among cetaceans revealed by Y-chromosome sequences. Zoological Science. 24: 723-732.

Noren SR, Williams TM. 2000. Body size and skeletal muscle myoglobin of cetaceans: adaptations for maximizing dive duration. Comp Biochem Phys A. 126: 181-191.

Notarbartolo di Sciara G, Frantzis A, Bearzi G, Reeves R. 2012. *Physeter macrocephalus* (Mediterranean subpopulation). IUCN 2013. IUCN Red List of Threatened Species. Version 2013.2. http://www.iucnredlist.org/details/16370739/0. Downloaded on 09 February 2014.

O'Corry-Crowe GM, Suydam RS, Rosenberg A, Frost KJ, Dizon AE. 1997. Phylogeography, population structure and dispersal patterns of the beluga whale *Delphinapterus leucas* in the western Nearctic revealed by mitochondrial DNA. Mol Ecol. 6: 955-970.

O'Leary MA, Gatesy J. 2008. Impact of increased character sampling on the phylogeny of Cetartiodactyla (Mammalia): combined analysis including fossils. Cladistics. 24: 397-442.

Oremus M et al. 2009. Worldwide mitochondrial DNA diversity and phylogeography of pilot whales (*Globicephala* spp.). Biol J Linn Soc. 98: 729-744.

Oremus M, Gales R, Kettles H, Baker CS. 2013. Genetic evidence of multiple matrilines and spatial disruption of kinship bonds in mass strandings of long-finned pilot whales, Globicephala melas. Journal of Heredity. 104: 301-311.

Oremus M, Poole MM, Steel D, Baker CS. 2007. Isolation and interchange among insular spinner dolphin communities in the South Pacific revealed by individual identification and genetic diversity. Marine Ecology Progress Series. 336: 275-289.

Ortega-Ortiz JG, Engelhaupt D, Winsor M, Mate BR, Hoelzel AR. 2012. Kinship of long-term associates in the highly social sperm whale. Mol Ecol. 21: 732-744.

Palsbøll PJ, Bérubé M, Allendorf FW. 2007. Identification of management units using population genetic data. Trends in Ecology & Evolution. 22: 11-16.

Palsbøll PJ, Bérubé M, Larsen AH, Jørgensen H. 1997. Primers for the amplification of tri-and tetramer microsatellite loci in baleen whales. Molecular Ecology. 6: 893-895.

Palsbøll PJ et al. 1995. Distribution of mtDNA haplotypes in North Atlantic humpback whales: the influence of behaviour on population structure. Mar Ecol-Prog Ser. 116: 1-10.

Palumbi SR, Cipriano F, Hare MP. 2001. Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. Evolution. 55: 859-868.

Pascoe PL, Mickiewicz MC, Castello HP. 1990. Cephalopod remains from the stomach of a sperm whale stranded off Patagonia. Marine Biology. 104: 1-4.

Pastene LA et al. 2007. Radiation and speciation of pelagic organisms during periods of global warming: the case of the common minke whale, *Balaenoptera acutorostrata*. Mol Ecol. 16: 1481-1495.

Patenaude NJ et al. 2007. Mitochondrial DNA diversity and population structure among southern right whales (*Eubalaena australis*). Journal of Heredity. 98: 147-157.

Patriquin KJ, Palstra F, Leonard ML, Broders HG. 2013. Female northern myotis (*Myotis septentrionalis*) that roost together are related. Behavioral Ecology. 24: 949-954.

Pavan G et al. 2000. Time patterns of sperm whale codas recorded in the Mediterranean Sea 1985-1996. The Journal of the Acoustical Society of America. 107: 3487-3495.

Peakall R, Smouse PE. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes. 6: 288-295.

Peakall R, Smouse PE. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research - an update. Bioinformatics. 28: 2537-2539.

Pecl GT, Jackson GD. 2008. The potential impacts of climate change on inshore squid: biology, ecology and fisheries. Reviews in Fish Biology and Fisheries. 18: 373-385.

Perrin WF. 2009. World Cetacea Database. Available online at http://www.marinespecies.org/cetacea. Consulted on 14-Jun-2012.

Perry SL, DeMaster DP, Silber GK. 1999. The great whales: history and status of six species listed as endangered under the U.S. Endangered Species Act of 1973. Marine Fisheries Review. 61: 1-74.

Petit R. 2010. PERMUT v 2.0. Available from http://www.pierroton.inra.fr/genetics/labo/Software/PermutCpSSR/index.html.

Pichler FB, Baker CS. 2000. Loss of genetic diversity in the endemic Hector's dolphin due to fisheries-related mortality. Proceedings of the Royal Society B: Biological Sciences. 267: 97-102.

Pichler FB, Dawson SM, Slooten E, Baker CS. 1998. Geographic isolation of Hector's dolphin populations described by mitochondrial DNA sequences. Conserv Biol. 12: 676-682.

Pimper L, Baker C, Goodall R, Olavarría C, Remis M. 2010. Mitochondrial DNA variation and population structure of Commerson's dolphins (*Cephalorhynchus commersonii*) in their southernmost distribution. Conserv Genet. 11: 2157-2168.

Pinela AM et al. 2009. Population genetics and social organization of the sperm whale (*Physeter macrocephalus*) in the Azores inferred by microsatellite analyses. Canadian Journal of Zoology. 87: 802-813.

Pompanon F, Bonin A, Bellemain E, Taberlet P. 2005. Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics. 6: 847-859.

Pons O, Petit RJ. 1996. Measuring and Testing Genetic Differentiation With Ordered Versus Unordered Alleles. Genetics. 144: 1237-1245.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155: 945-959.

Queller DC, Goodnight KF. 1989. Estimating relatedness using genetic markers. Evolution. 43: 258-275.

Quérouil S et al. 2010. Molecular insight into the population structure of common and spotted dolphins inhabiting the pelagic waters of the Northeast Atlantic. Mar Biol. 157: 2567-2580.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26: 841-842.

R Core Team. 2013. R: A language and environment for statistical computing. URL http://www.R-project.org/.

Rambaut A. 2009. FigTree, available from http://tree.bio.ed.ac.uk/software/figtree/.

Rambaut A. 2014. FigTree v. 1.4.1, available from http://tree.bio.ed.ac.uk/software/figtree/.

Rambaut A, Drummond A. 2007. Tracer v1.4, available from http://beast.bio.ed.ac.uk/Tracer.

Rambaut A, Suchard M, Drummond A. 2013. Tracer v 1.6, available from http://tree.bio.ed.ac.uk/software/tracer/.

Ramírez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A. 2008. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics. 179: 555-567.

Ramos-Onsins SE, Rozas J. 2002. Statistical properties of new neutrality tests against population growth. Molecular Biology and Evolution. 19: 2092-2100.

Raymond M, Rousset F. 1995a. An exact test for population differentiation. Evolution. 49: 1280-1283.

Raymond M, Rousset F. 1995b. GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. Journal of Heredity. 86: 248-249.

Reeves RR. 2002. The origins and character of 'aboriginal subsistence' whaling: a global review. Mammal Review. 32: 71-106.

Reich DE, Feldman MW, Goldstein DB. 1999. Statistical properties of two tests that use multilocus data sets to detect population expansions. Molecular Biology and Evolution. 16: 453-466.

Reilly SB et al. 2013. *Eubalaena australis*. The IUCN Red List of Threatened Species. Version 2014.2. Downloaded from: http://www.iucnredlist.org/details/8153/0.

Rendell L, Mesnick S, Dalebout M, Burtenshaw J, Whitehead H. 2012. Can genetic differences explain vocal dialect variation in sperm whales, *Physeter macrocephalus?* Behav Genet. 42: 332-343.

Rendell LE, Whitehead H. 2003. Vocal clans in sperm whales (*Physeter macrocephalus*). Proceedings of the Royal Society B: Biological Sciences. 270: 225-231.

Rice DW, Wolman AA, Mate BR, Harvey JT. 1986. A mass stranding of sperm whales in Oregon: sex and age composition of the school. Marine Mammal Science. 2: 64-69.

Richard KR, Dillon MC, Whitehead H, Wright JM. 1996a. Patterns of kinship in groups of free-living sperm whales (*Physeter macrocephalus*) revealed by multiple molecular genetic analyses. P Natl Acad Sci USA. 93: 8792-8795.

Richard KR, McCarrey SW, Wright JM. 1994. DNA sequence from the *SRY* gene of the sperm whale (*Physeter macrocephalus*) for use in molecular sexing. Can J Zoolog. 72: 873-877.

Richard KR, Whitehead H, Wright JM. 1996b. Polymorphic microsatellites from sperm whales and their use in the genetic identification of individuals from naturally sloughed pieces of skin. Molecular Ecology. 5: 313-315.

Richly E, Leister D. 2004. NUMTs in sequenced eukaryotic genomes. Mol Biol Evol. 21: 1081-1084.

Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ. 2014. ABC inference of multi-population divergence with admixture from unphased population genomic data. Molecular Ecology. 23: 4458-4471.

Rokas A, Atkinson RJ, Brown GS, West SA, Stone GN. 2001. Understanding patterns of genetic diversity in the oak gallwasp *Biorhiza pallida*: demographic history or a *Wolbachia* selective sweep? Heredity. 87: 294-304.

Rooney AP, Honeycutt RL, Derr JN. 2001. Historical population size change of bowhead whales inferred from DNA sequence polymorphism data. Evolution. 55: 1678-1685.

Rosel PE, Dizon AE, Heyning JE. 1994. Genetic analysis of sympatric morphotypes of common dolphins (genus *Delphinus*). Marine Biology. 119: 159-167.

Rosel PE, France SC, Wang JY, Kocher TD. 1999. Genetic structure of harbour porpoise *Phocoena phocoena* populations in the northwest Atlantic based on mitochondrial and nuclear markers. Mol Ecol. 8: S41-S54.

Rosenbaum HC et al. 2009. Population structure of humpback whales from their breeding grounds in the South Atlantic and Indian Oceans. PLoS ONE. 4: e7318.

Rousset F. 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics. 145: 1219-1228.

Rousset F. 2008. GENEPOP'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular Ecology Resources. 8: 103-106.

Ruegg KC et al. 2010. Are Antarctic minke whales unusually abundant because of 20th century whaling? Mol Ecol. 19: 281-291.

Sambrook J, Fritsch EF, Maniatis T. 1989. *Molecular cloning: a laboratory manual*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Savery LC et al. 2013a. Global mercury and selenium concentrations in skin from free-ranging sperm whales (*Physeter macrocephalus*). Science of The Total Environment. 450-451: 59-71.

Savery LC et al. 2013b. Global assessment of silver pollution using sperm whales (*Physeter macrocephalus*) as an indicator species. Environmental & Analytical Toxicology. 3: 169.

Schlötterer C. 2003. Hitchhiking mapping - functional genomics from the population genetics perspective. Trends in Genetics. 19: 32-38.

Schlötterer C, Amos B, Tautz D. 1991. Conservation of polymorphic simple sequence loci in cetacean species. Nature. 354: 63-65.

Schuttler SG, Philbrick JA, Jeffery KJ, Eggert LS. 2014. Fine-scale genetic structure and cryptic associations reveal evidence of kin-based sociality in the African forest elephant. PLoS ONE. 9: e88074.

Shen Y et al. 2010. A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. 20: 273-280.

Siddall ME. ild.run, available from http://tnt.insectmuseum.org/index.php/Scripts/ild.run.

Siemann LA. 1994. Mitochondrial DNA sequence variation in North Atlantic long-finned pilot whales, *Globicephala melas*. Doctor of Philosophy. Massachusetts Institute of Technology/Woods Hole Oceanographic Institution, Massachusetts, U.S.A.

Silk JB. 2007. The adaptive value of sociality in mammalian groups. Philosophical Transactions of the Royal Society B: Biological Sciences. 362: 539-559.

Silk JB. 2009. Nepotistic cooperation in non-human primate groups. Philosophical Transactions of the Royal Society B: Biological Sciences. 364: 3243-3254.

Simonsen KL, Churchill GA, Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics. 141: 413-429.

Smith DR et al. 2008. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res. 18: 1638-1642.

Sorenson MD, Quinn TW. 1998. Numts: A challenge for avian systematics and population biology. Auk. 115: 214-221.

Sosa MX et al. 2012. Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS Comput Biol. 8: e1002737.

Sremba AL, Hancock-Hanser B, Branch TA, LeDuc RL, Baker CS. 2012. Circumpolar diversity and geographic differentiation of mtDNA in the critically endangered Antarctic Blue whale (*Balaenoptera musculus intermedia*). PLoS ONE. 7: e32579.

Steeman ME et al. 2009. Radiation of extant cetaceans driven by restructuring of the oceans. Syst Biol. 58: 573-585.

Steiner L et al. 2012. A link between male sperm whales, *Physeter macrocephalus*, of the Azores and Norway. Journal of the Marine Biological Association of the United Kingdom. 92: 1751-1756.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. The American Journal of Human Genetics. 76: 449-462.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. The American Journal of Human Genetics. 68: 978-989.

Stephenson AB. 1975. Sperm whales stranded at Muriwai Beach, New Zealand. New Zealand Journal of Marine and Freshwater Research. 9: 299-304.

Stewart JB et al. 2008. Strong purifying selection in transmission of mammalian mitochondrial DNA. PLoS Biol. 6: e10.

Storz JF, Payseur BA, Nachman MW. 2004. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Molecular Biology and Evolution. 21: 1800-1811.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123: 585-595.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution. 10: 512-526.

Tamura K et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28: 2731-2739.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution. 30: 2725-2729.

Taylor BL, Chivers SJ, Larese J, Perrin WF. 2007. Generation length and percent mature estimates for IUCN assessments of cetaceans. Administrative Report LJ-07-01, Southwest Fisheries Science Center, 8604 La Jolla Shores Blvd., La Jolla, CA 92038, USA. 24 pp.

Thalmann O, Hebler J, Poinar HN, Pääbo S, Vigilant L. 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. Mol Ecol. 13: 321-335.

Thompson KF et al. 2013. A novel conservation approach provides insights into the management of rare cetaceans. Biol Conserv. 331-340.

Thompson KF et al. In Press. A novel conservation approach provides insights into the management of rare cetaceans. Biol Conserv.

Tiedemann R, Milinkovitch MC. 1999. Culture and genetic evolution in whales. Science. 284: 2055.

Valsecchi E, Amos W. 1996. Microsatellite markers for the study of cetacean populations. Molecular Ecology. 5: 151-156.

Valsecchi E, Hale P, Corkeron P, Amos W. 2002. Social structure in migrating humpback whales (*Megaptera novaeangliae*). Molecular Ecology. 11: 507-518.

van Helden A. 2009. New Zealand *Physeter macrocephalus* stranding record. Administered by Te Papa Tongarewa for the Department of Conservation (DOC).

Van Horn RC, Engh AL, Scribner KT, Funk SM, Holekamp KE. 2004. Behavioural structuring of relatedness in the spotted hyena (Crocuta crocuta) suggests direct fitness benefits of clan-level cooperation. Molecular Ecology. 13: 449-458.

Van Horn RC, Buchan JC, Altmann J, Alberts SC. 2007. Divided destinies: group choice by female savannah baboons during social group fission. Behavioral Ecology and Sociobiology. 61: 1823-1837.

van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P. 2004. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes. 4: 535-538.

Vidya TNC, Sukumar R. 2005. Social and reproductive behaviour in elephants. Current Science. 89: 1200-1207.

Vidya TNC, Sukumar R, Melnick DJ. 2009. Range-wide mtDNA phylogeography yields insights into the origins of Asian elephants. Proceedings of the Royal Society B: Biological Sciences. 276: 893-902.

Vilstrup J et al. 2011. Mitogenomic phylogenetic analyses of the Delphinidae with an emphasis on the Globicephalinae. BMC Evol Biol. 11: 65.

Vollmer NL, Rosel PE. 2012. Developing genomic resources for the common bottlenose dolphin (*Tursiops truncatus*): isolation and characterization of 153 single nucleotide polymorphisms and 53 genotyping assays. Molecular Ecology Resources. 12: 1124-1132.

Vollmer NL, Viricel A, Wilcox L, Moore KM, Rosel PE. 2011. The occurrence of mtDNA heteroplasmy in multiple cetacean species. Curr Genet. 57: 115-131.

Wade PR, Reeves RR, Mesnick SL. 2012. Social and behavioural factors in cetacean responses to overexploitation: are odontocetes less "resilient" than mysticetes? Journal of Marine Biology. 2012: 15.

Waldick RC, Brown MW, White BN. 1999. Characterization and isolation of microsatellite loci from the endangered North Atlantic right whale. Molecular Ecology. 8: 1763.

Wang JY, Chou LS, White BN. 1999. Mitochondrial DNA analysis of sympatric morphotypes of bottlenose dolphins (genus: *Tursiops*) in Chinese waters. Molecular Ecology. 8: 1603-1612.

Waples RS, Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology. 15: 1419-1439.

Watkins WA, Daher MA, Fristrup KM, Howald TJ, Notarbartolo di Sciara G. 1993. Sperm whales tagged with transponders and tracked underwater by sonar. Mar Mammal Sci. 9: 55-67.

Watkins WA, Schevill WE. 1977. Sperm whale codas. The Journal of the Acoustical Society of America. 62: 1485-1490.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology. 7: 256-276.

Watwood SL, Miller PJO, Johnson M, Madsen PT, Tyack PL. 2006. Deep-diving foraging behaviour of sperm whales (*Physeter macrocephalus*). J Anim Ecol. 75: 814-825.

Wayne ML, Simonsen KL. 1998. Statistical tests of neutrality in the age of weak selection. Trends in Ecology & Evolution. 13: 236-240.

Weilgart L, Whitehead H. 1993. Coda communication by sperm whales (*Physeter macrocephalus*) off the Galápagos Islands. Canadian Journal of Zoology. 71: 744-752.

Weilgart L, Whitehead H. 1997. Group-specific dialects and geographical variation in coda repertoire in South Pacific sperm whales. Behavioral Ecology and Sociobiology. 40: 277-285.

Weilgart L, Whitehead H, Payne K. 1996. A colossal convergence. American Scientist. 84: 278-287.

Weinrich MT, Rosenbaum H, Baker CS, Blackmer AL, Whitehead H. 2006. The influence of maternal lineages on social affiliations among humpback whales (*Megaptera*

novaeangliae) on their feeding grounds in the southern Gulf of Maine. Journal of Heredity. 97: 226-234.

Whitehead H. 1993. The behaviour of mature male sperm whales on the Galápagos Islands breeding grounds. Canadian Journal of Zoology. 71: 689-699.

Whitehead H. 1994. Delayed competitive breeding in roving males. Journal of Theoretical Biology. 166: 127-133.

Whitehead H. 1996. Babysitting, dive synchrony, and indications of alloparental care in sperm whales. Behavioral Ecology and Sociobiology. 38: 237-244.

Whitehead H. 1998. Cultural selection and genetic diversity in matrilineal whales. Science. 282: 1708-1711.

Whitehead H. 1999. Culture and genetic evolution in whales. Science. 284: 2055a.

Whitehead H. 2001. Analysis of animal movement using opportunistic individual identifications: application to sperm whales. Ecology. 82: 1417-1432.

Whitehead H. 2002. Estimates of the current global population size and historical trajectory for sperm whales. Mar Ecol-Prog Ser. 242: 295-304.

Whitehead H. 2003. *Sperm whales: social evolution in the ocean*. Chicago: The University of Chicago Press.

Whitehead H. 2005. Genetic diversity in the matrilineal whales: models of cultural hitchhiking and group-specific non-heritable demographic variation. Mar Mammal Sci. 21: 58-79.

Whitehead H et al. 2012. Multilevel societies of female sperm whales (*Physeter macrocephalus*) in the Atlantic and Pacific: why are they so different? International Journal of Primatology. 33: 1142-1164.

Whitehead H, Arnbom T. 1987. Social organization of sperm whales off the Galapagos Islands, February-April 1985. Canadian Journal of Zoology. 65: 913-919.

Whitehead H, Brennan S, Grover D. 1992. Distribution and behaviour of male sperm whales on the Scotian Shelf, Canada. Canadian Journal of Zoology. 70: 912-918.

Whitehead H, Christal J, Dufault S. 1997. Past and distant whaling and the rapid decline of sperm whales off the Galápagos Islands. Conserv Biol. 11: 1387-1396.

Whitehead H, Dillon M, Dufault S, Weilgart L, Wright J. 1998. Non-geographically based population structure of South Pacific sperm whales: dialects, fluke-markings and genetics. J Anim Ecol. 67: 253-262.

Whitehead H, Rendell L. 2004. Movements, habitat use and feeding success of cultural clans of South Pacific sperm whales. J Anim Ecol. 73: 190-196.

Whitehead H, Waters S, Lyrholm T. 1991. Social organization of female sperm whales and their offspring: constant companions and casual acquaintances. Behavioral Ecology and Sociobiology. 29: 385-389.

Whitlock MC, McCauley DE. 1999. Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$. Heredity. 82: 117-125.

Winkelmann I et al. 2013. Mitochondrial genome diversity and population structure of the giant squid *Architeuthis*: genetics sheds new light on one of the most enigmatic marine species. Proceedings of the Royal Society B: Biological Sciences. 280: 20130273.

Wise Sr JP et al. 2009. A global assessment of chromium pollution using sperm whales (*Physeter macrocephalus*) as an indicator species. Chemosphere. 75: 1461-1467.

Wise Sr. JP et al. 2011. A global assessment of gold, titanium, strontium and barium pollution using sperm whales (*Physeter macrocephalus*) as an indicator species. Journal of Ecosystem & Ecography. 1: 101.

Wittemyer G et al. 2009. Where sociality and relatedness diverge: the genetic basis for hierarchical social organization in African elephants. Proceedings of the Royal Society B: Biological Sciences.

Wright S. 1931. Evolution in Mendelian populations. Genetics. 16: 97-159.

Wright S. 1938. Size of population and breeding structure in relation to evolution. Science. 87: 430-431.

Yang G, Guo LI, Bruford MW, Wei F, Zhou K. 2008. Mitochondrial phylogeography and population history of finless porpoises in Sino-Japanese waters. Bio J Linn Soc. 95: 193-204.

Zhang D-X, Hewitt GM. 1996. Nuclear integrations: challenges for mitochondrial DNA markers. Trends Ecol Evol. 11: 247-251.

Appendices

Appendix I
Supplementary material associated with Chapter 2: Female philopatry drives high levels of regional structure in the globally distributed sperm whale (*Physeter macrocephalus*)

The files listed below are archived with the author, as well as the Cetacean Conservation and Genomics Laboratory at Oregon State University's Hatfield Marine Science Center. Requests for data access can be submitted to Alana Alexander (alanamaryalexander@gmail.com) and C. Scott Baker (scott.baker@oregonstate.edu), and should be accompanied by a proposal describing how the data would be used. A two-year embargo applies to the data, however, requests made before two years will be considered on a case-by-case basis, depending upon the proposed use and potential for collaboration.

DNA profiles and metadata for* Odyssey *and stranding samples

File name: Sperm_whale_DNA_profiles_AA_16Aug2014.txt

Sample name, individual code, latitude, longitude, date of sampling and group code for *Odyssey* and stranding samples that passed QC. DNA profiles (394 bp mitochondrial control region haplotype, sex, and 13-microsatellite genotype) also listed.

Sperm whale mitochondrial DNA control region haplotype definitions

File name: Sperm_whale_mtDNA_CR_haplotypes_AA_16Aug2014.fas

Fasta file with the DNA sequence definitions for the haplotypes sequenced in this study, also defined in Appendix I, Table I.1.

Table I.1: Summary of the variable sites that define the sperm whale mtDNA CR haplotypes found in this study. Sequences from this study start 5 bp later than those from Mesnick et al. (2011), so both numbering schemes are provided (Mes Ref and Alx Ref). Dashes indicate no information for a sequence at that site (no indels were observed among the samples sequenced in this study). Haplotypes named with reference to identity over 394, 569 and 619 bp consensus lengths. Haplotypes CC-HH, JJ-OO are new to this study.

Haplotypes (bp)			Mes Ref	43	58	62	105	107	109	121	150	184	200	207	208	211	235	238	243	260	272	273	283	286	287	288	289	291	295	305	308	319	324	350	574	608	624
394	569	619	Alx Ref	38	53	57	100	102	104	116	145	179	195	202	203	206	230	233	238	255	267	268	278	281	282	283	284	286	290	300	303	314	319	345	569	603	619
A	'.001	'.001		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	A	G
A	'.001	'.002		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	A	A
A	'.001	'.003		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	G	G
A	'.002	'.001		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	T	A	G
B	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	A	G
B	'.001	'.002		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	A	A
B	'.002	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	T	A	G
C	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	G	T	A	G	C	A	G	C	C	G	A	A
C	'.001	'.002		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	G	T	A	G	C	A	G	C	C	G	A	G
C	'.002	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	G	T	A	G	C	A	G	C	C	T	A	G
D	'.001	'.001		T	T	C	C	A	G	C	C	T	C	A	A	T	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	A	G
E	'.001	'.001		T	C	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	A	G
F	'.NA	'.NA		T	C	C	C	A	G	C	T	T	T	A	A	C	A	T	G	A	A	T	C	A	A	A	T	A	G	C	A	G	C	C	-	-	-
G	'.NA	'.NA		T	C	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	T	-	-	-
H	'.001	'.001		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	T	C	A	G	A	T	A	G	C	A	G	C	C	G	A	G
I	'.001	'.001		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	T	C	G	A	G
J	'.001	'.001		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	G	T	A	G	C	A	G	C	C	G	A	G
J	'.002	'.001		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	G	T	A	G	C	A	G	C	C	T	A	G
K	'.NA	'.NA		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	T	C	A	A	G	T	A	G	C	A	G	T	C	-	-	-
L	'.NA	'.NA		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	A	A	A	C	C	A	A	G	T	A	G	C	A	G	C	C	-	-	-
M	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	G	A	A	T	A	G	C	A	G	C	C	G	A	G
N	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	A	C	C	G	A	G
N	'.001	'.002		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	A	C	C	G	A	A
N	'.002	'.NA		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	A	C	C	T	-	-
O	'.001	'.001		T	T	T	C	A	G	C	C	C	T	A	A	C	A	T	G	G	A	C	C	A	A	A	T	A	G	C	A	A	T	C	G	A	G
P	'.NA	'.NA		T	T	T	C	A	G	C	C	T	T	A	A	T	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	-	-	-
Q	'.001	'.001		T	T	T	C	A	G	C	C	C	T	A	A	C	A	T	G	G	G	C	C	A	A	A	T	A	G	C	A	A	T	C	G	A	G
R	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	G	A	C	C	A	A	A	T	A	G	C	A	A	T	C	G	A	G
S	'.001	'.001		T	T	T	T	A	G	C	C	C	T	A	A	C	A	T	G	G	G	C	C	A	A	A	T	A	G	C	A	A	T	C	G	A	G
T	'.NA	'.NA		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	G	C	C	A	A	A	T	A	G	C	A	G	C	C	-	-	-
U	'.NA	'.NA		T	T	C	C	A	G	C	C	T	T	A	G	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	-	-	-
V	'.NA	'.NA		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	C	A	G	C	A	G	C	C	-	-	-
W	'.001	'.NA		T	T	C	C	A	G	C	C	T	T	A	A	T	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	-	-
X	'.001	'.001		T	T	T	C	A	G	C	C	T	T	G	A	C	A	T	G	A	A	C	C	A	A	G	T	A	G	C	A	G	C	C	G	A	G
Y	'.NA	'.NA		T	T	T	C	A	G	C	C	T	T	G	A	C	A	T	G	A	G	C	C	A	A	G	T	A	G	C	A	G	C	C	-	-	-
Z	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	G	C	G	G	C	C	G	A	G	G	A	G
AA	'.NA	'.NA		T	C	C	C	A	G	C	T	T	T	A	A	C	G	T	G	A	A	C	C	A	A	A	T	A	G	T	A	G	C	C	-	-	-
BB	'.001	'.001		T	T	C	C	A	G	T	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	T	A	G

Table I.1 (Continued)

Haplotypes (bp)			Mes Ref	43	58	62	105	107	109	121	150	184	200	207	208	211	235	238	243	260	272	273	283	286	287	288	289	291	295	305	308	319	324	350	574	608	624
394	569	619	Alx Ref	38	53	57	100	102	104	116	145	179	195	202	203	206	230	233	238	255	267	268	278	281	282	283	284	286	290	300	303	314	319	345	569	603	619
CC	'.001	'.001		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	T	C	A	A	A	T	A	G	C	A	G	C	C	G	A	G
DD	'.001	'.001		T	T	T	C	A	G	C	C	C	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	A	G
EE	'.001	'.001		T	T	C	C	G	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	A	G
FF	'.001	'.001		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	A	C	A	G	C	C	T	A	G
GG	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	C	G	A	A	C	C	A	A	G	T	A	G	C	A	G	C	C	G	A	G
HH	'.001	'.001		C	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	T	C	G	A	G
II	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	T	C	A	A	A	T	A	G	C	A	G	C	C	G	A	G
JJ	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	G	T	A	G	C	A	G	T	C	T	A	G
JJ	'.002	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	G	T	A	G	C	A	G	T	C	G	A	G
KK	'.001	'.001		T	T	C	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	G	T	G	G	C	A	G	C	C	G	A	G
LL	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	A	C	T	A	A	G	T	A	G	C	A	G	C	C	G	A	G
MM	'.001	'.001		T	T	T	C	A	G	C	C	T	T	A	A	C	A	T	G	A	G	C	C	A	A	G	T	A	G	C	A	G	C	C	G	A	G
NN	'.001	'.001		T	C	C	C	A	G	C	T	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	G	C	C	G	A	G
OO	'.001	'.001		T	T	T	C	A	A	C	C	T	T	A	A	C	A	T	G	A	A	C	C	A	A	A	T	A	G	C	A	A	C	C	G	-	-

Table I.2: Summary of the sperm whale mtDNA CR haplotypes obtained for samples from previous publications, with ocean and geographic region. Haplotype definitions are listed in ***Table I.1.*** Haplotypes were also available from Drouot et al. (2004) but were not included as they were 200 bp shorter than the haplotypes listed here. No attempt has been made to identify replicate samples between studies. Abbreviations: PAC (Pacific), IO (Indian), ATL (Atlantic), NA (North Atlantic), SH (Southern Hemisphere), GPG (Galapagos), JC (Japanese Coast), NPW (western North Pacific), NPC (central North Pacific), NPE (eastern North Pacific), GMX (Gulf of Mexico), NSEA (North Sea), WNAO (western North Atlantic), MED (Mediterranean), CHI/PER (Chile/Peru), GPG/ECU (Galapagos/Ecuador), SWP (southwest Pacific), CA/OR (California/Oregon), HAW (Hawai'i), ETP (eastern tropical Pacific), CHI (Chile), PER (Peru).

Haplotype (394 bp)	569bp	619bp	Region	Ocean	Publication	No. of individuals	Notes
A	NA	NA	NA	ATL	Lyrholm et al. 1998	18	
A	NA	NA	SH	IO/PAC	Lyrholm et al. 1998	12	
A	NA	NA	GPG	PAC	Lyrholm et al. 1998	15	
A	NA	NA	JC	PAC	Lyrholm et al. 1998	19	
A	NA	NA	NPW	PAC	Lyrholm et al. 1998	11	
A	NA	NA	NPC	PAC	Lyrholm et al. 1998	5	
A	NA	NA	NPE	PAC	Lyrholm et al. 1998	15	
C	NA	NA	NA	ATL	Lyrholm et al. 1998	17	
C	NA	NA	GPG	PAC	Lyrholm et al. 1998	8	
C	NA	NA	JC	PAC	Lyrholm et al. 1998	2	
C	NA	NA	NPW	PAC	Lyrholm et al. 1998	3	
C	NA	NA	NPC	PAC	Lyrholm et al. 1998	5	
C	NA	NA	NPE	PAC	Lyrholm et al. 1998	1	
B	NA	NA	NA	ATL	Lyrholm et al. 1998	9	
B	NA	NA	SH	IO/PAC	Lyrholm et al. 1998	12	
B	NA	NA	GPG	PAC	Lyrholm et al. 1998	14	
B	NA	NA	JC	PAC	Lyrholm et al. 1998	2	
B	NA	NA	NPW	PAC	Lyrholm et al. 1998	12	
B	NA	NA	NPC	PAC	Lyrholm et al. 1998	1	
B	NA	NA	NPE	PAC	Lyrholm et al. 1998	9	
BB	NA	NA	NA	ATL	Lyrholm et al. 1998	1	
E	NA	NA	SH	IO/PAC	Lyrholm et al. 1998	1	This is also a match to haplotype G (slightly shorter than 394 bp)
E	NA	NA	JC	PAC	Lyrholm et al. 1998	1	This is also a match to haplotype G (slightly shorter than 394 bp)
E	NA	NA	NPW	PAC	Lyrholm et al. 1998	4	This is also a match to haplotype G (slightly shorter than 394 bp)
E	NA	NA	NPE	PAC	Lyrholm et al. 1998	3	This is also a match to haplotype G (slightly shorter than 394 bp)
Q	NA	NA	JC	PAC	Lyrholm et al. 1998	1	
J	NA	NA	SH	IO/PAC	Lyrholm et al. 1998	9	
N	NA	NA	NA	ATL	Lyrholm et al. 1998	2	
N	NA	NA	SH	IO/PAC	Lyrholm et al. 1998	2	
N	NA	NA	JC	PAC	Lyrholm et al. 1998	4	
N	NA	NA	NPE	PAC	Lyrholm et al. 1998	1	
O	NA	NA	SH	IO/PAC	Lyrholm et al. 1998	2	
H	NA	NA	GPG	PAC	Lyrholm et al. 1998	1	
H	NA	NA	NPW	PAC	Lyrholm et al. 1998	1	
M	NA	NA	NPW	PAC	Lyrholm et al. 1998	1	
K	NA	NA	NPW	PAC	Lyrholm et al. 1998	2	
D	NA	NA	NPE	PAC	Lyrholm et al. 1998	2	
R	NA	NA	SH	IO/PAC	Lyrholm et al. 1998	1	
S	NA	NA	SH	IO/PAC	Lyrholm et al. 1998	1	

Table I.2 (Continued)

Haplotype (394 bp)	569bp	619bp	Region	Ocean	Publication	No. of individuals	Notes
II	NA	NA	SH	IO/PAC	Lyrholm et al. 1998	1	
A	NA	NA	GMX	ATL	Engelhaupt et al. 2009	4	
A	NA	NA	NSEA	ATL	Engelhaupt et al. 2009	8	
A	NA	NA	WNAO	ATL	Engelhaupt et al. 2009	36	
B	NA	NA	GMX	ATL	Engelhaupt et al. 2009	23	
B	NA	NA	NSEA	ATL	Engelhaupt et al. 2009	2	
B	NA	NA	WNAO	ATL	Engelhaupt et al. 2009	8	
C	NA	NA	MED	ATL	Engelhaupt et al. 2009	38	
C	NA	NA	GMX	ATL	Engelhaupt et al. 2009	11	
C	NA	NA	NSEA	ATL	Engelhaupt et al. 2009	8	
C	NA	NA	WNAO	ATL	Engelhaupt et al. 2009	37	
X	NA	NA	GMX	ATL	Engelhaupt et al. 2009	105	
X	NA	NA	WNAO	ATL	Engelhaupt et al. 2009	1	
Y	NA	NA	GMX	ATL	Engelhaupt et al. 2009	10	
N	NA	NA	WNAO	ATL	Engelhaupt et al. 2009	1	
BB	NA	NA	WNAO	ATL	Engelhaupt et al. 2009	1	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	3	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	5	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	7	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	4	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	4	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	5	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	2	
B	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	1	
Y	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	1	
C	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	1	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	3	
B	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	1	
C	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	1	
B	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	5	
B	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	2	
C	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	2	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	2	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	2	
Y	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	1	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	6	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	2	
Y	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	2	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	4	
B	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	3	
B	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	1	
X	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	2	
Y	NA	NA	GMX	ATL	Ortega-Ortiz et al. 2012	5	
A	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
C	NA	NA	CHI/PER	PAC	Rendell et al. 2012	2	
Z	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
J	NA	NA	CHI/PER	PAC	Rendell et al. 2012	2	
B	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
C	NA	NA	CHI/PER	PAC	Rendell et al. 2012	2	
J	NA	NA	CHI/PER	PAC	Rendell et al. 2012	2	
A	NA	NA	CHI/PER	PAC	Rendell et al. 2012	2	
B	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
I	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
B	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
C	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
A	NA	NA	CHI/PER	PAC	Rendell et al. 2012	19	
B	NA	NA	CHI/PER	PAC	Rendell et al. 2012	5	
A	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
C	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
A	NA	NA	CHI/PER	PAC	Rendell et al. 2012	7	
B	NA	NA	CHI/PER	PAC	Rendell et al. 2012	8	
C	NA	NA	CHI/PER	PAC	Rendell et al. 2012	2	

Table I.2 (Continued)

Haplotype (394 bp)	569bp	619bp	Region	Ocean	Publication	No. of individuals	Notes
A	NA	NA	CHI/PER	PAC	Rendell et al. 2012	2	
B	NA	NA	CHI/PER	PAC	Rendell et al. 2012	2	
O	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
B	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
E	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
B	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
E	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
B	NA	NA	CHI/PER	PAC	Rendell et al. 2012	2	
E	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
D	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
A	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
B	NA	NA	CHI/PER	PAC	Rendell et al. 2012	6	
A	NA	NA	CHI/PER	PAC	Rendell et al. 2012	1	
A	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	9	
C	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	4	
G	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	1	
N	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	3	
M	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	4	
A	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	17	
H	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	1	
A	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	3	
B	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	13	
C	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	1	
F	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	1	
A	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	3	
A	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	1	
B	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	2	
C	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	17	
C	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	2	
A	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	2	
D	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	3	
A	NA	NA	GPG/ECU	PAC	Rendell et al. 2012	1	
A	NA	NA	SWP	PAC	Rendell et al. 2012	4	
A	NA	NA	SWP	PAC	Rendell et al. 2012	1	
C	NA	NA	SWP	PAC	Rendell et al. 2012	2	
O	NA	NA	SWP	PAC	Rendell et al. 2012	1	
A	NA	NA	SWP	PAC	Rendell et al. 2012	2	
C	NA	NA	SWP	PAC	Rendell et al. 2012	1	
M	NA	NA	SWP	PAC	Rendell et al. 2012	1	
A	NA	NA	SWP	PAC	Rendell et al. 2012	1	
B	NA	NA	SWP	PAC	Rendell et al. 2012	2	
G	NA	NA	SWP	PAC	Rendell et al. 2012	2	
D	NA	NA	SWP	PAC	Rendell et al. 2012	1	
G	NA	NA	SWP	PAC	Rendell et al. 2012	1	
I	NA	NA	SWP	PAC	Rendell et al. 2012	5	
B	NA	NA	SWP	PAC	Rendell et al. 2012	1	
A	NA	NA	CA/OR	PAC	Mesnick et al. 2011	15	
A	NA	NA	HAW	PAC	Mesnick et al. 2011	15	
A	NA	NA	ETP	PAC	Mesnick et al. 2011	38	
AA	NA	NA	CA/OR	PAC	Mesnick et al. 2011	2	
B	NA	NA	CA/OR	PAC	Mesnick et al. 2011	10	
B	NA	NA	HAW	PAC	Mesnick et al. 2011	6	
B	NA	NA	ETP	PAC	Mesnick et al. 2011	39	
C	NA	NA	CA/OR	PAC	Mesnick et al. 2011	6	
C	NA	NA	ETP	PAC	Mesnick et al. 2011	10	
D	NA	NA	ETP	PAC	Mesnick et al. 2011	3	
E	NA	NA	CA/OR	PAC	Mesnick et al. 2011	11	
E	NA	NA	HAW	PAC	Mesnick et al. 2011	1	
E	NA	NA	ETP	PAC	Mesnick et al. 2011	2	
F	NA	NA	CA/OR	PAC	Mesnick et al. 2011	1	
F	NA	NA	ETP	PAC	Mesnick et al. 2011	4	
G	NA	NA	CA/OR	PAC	Mesnick et al. 2011	1	

Table I.2 (Continued)

Haplotype (394 bp)	569bp	619bp	Region	Ocean	Publication	No. of individuals	Notes
H	NA	NA	CA/OR	PAC	Mesnick et al. 2011	2	
H	NA	NA	ETP	PAC	Mesnick et al. 2011	3	
I	NA	NA	CA/OR	PAC	Mesnick et al. 2011	1	
J	NA	NA	HAW	PAC	Mesnick et al. 2011	6	
J	NA	NA	ETP	PAC	Mesnick et al. 2011	11	
L	NA	NA	ETP	PAC	Mesnick et al. 2011	1	
M	NA	NA	CA/OR	PAC	Mesnick et al. 2011	1	
N	NA	NA	CA/OR	PAC	Mesnick et al. 2011	1	
N	NA	NA	ETP	PAC	Mesnick et al. 2011	1	
O	NA	NA	ETP	PAC	Mesnick et al. 2011	1	
P	NA	NA	CA/OR	PAC	Mesnick et al. 2011	1	
P	NA	NA	ETP	PAC	Mesnick et al. 2011	1	
B	NA	NA	CHI/PER	PAC	Whitehead et al. 1998	2	
A	NA	NA	CHI/PER	PAC	Whitehead et al. 1998	1	
B	NA	NA	CHI/PER	PAC	Whitehead et al. 1998	6	
A	NA	NA	CHI	PAC	Whitehead et al. 1998	1	
C	NA	NA	CHI	PAC	Whitehead et al. 1998	1	
B	NA	NA	CHI	PAC	Whitehead et al. 1998	2	
E	NA	NA	CHI	PAC	Whitehead et al. 1998	1	
I	NA	NA	CHI	PAC	Whitehead et al. 1998	1	
A	NA	NA	PER	PAC	Whitehead et al. 1998	1	
N	NA	NA	PER	PAC	Whitehead et al. 1998	2	
O	NA	NA	PER	PAC	Whitehead et al. 1998	1	
A	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	3	
B	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	1	
J	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	1	
A	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	3	
C	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	1	
B	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	13	
D	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	1	
A	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	2	
A	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	1	
C	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	17	
B	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	2	
A	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	2	
A	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	9	
C	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	4	
J	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	1	
T	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	3	
S	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	4	
B	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	6	
E	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	1	
I	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	2	
A	NA	NA	GPG/ECU	PAC	Whitehead et al. 1998	4	
A	NA	NA	SWP	PAC	Whitehead et al. 1998	1	
C	NA	NA	SWP	PAC	Whitehead et al. 1998	2	
U	NA	NA	SWP	PAC	Whitehead et al. 1998	1	
A	NA	NA	SWP	PAC	Whitehead et al. 1998	4	
O	NA	NA	SWP	PAC	Whitehead et al. 1998	5	
J	NA	NA	SWP	PAC	Whitehead et al. 1998	1	
A	NA	NA	SWP	PAC	Whitehead et al. 1998	2	
C	NA	NA	SWP	PAC	Whitehead et al. 1998	1	
S	NA	NA	SWP	PAC	Whitehead et al. 1998	1	
A	NA	NA	SWP	PAC	Whitehead et al. 1998	1	
B	NA	NA	SWP	PAC	Whitehead et al. 1998	2	
J	NA	NA	SWP	PAC	Whitehead et al. 1998	2	
E	NA	NA	SWP	PAC	Whitehead et al. 1998	1	
A	NA	NA	GPG/ECU	PAC	Richard et al. 1996	3	
C	NA	NA	GPG/ECU	PAC	Richard et al. 1996	1	
B	NA	NA	GPG/ECU	PAC	Richard et al. 1996	12	
A	NA	NA	GPG/ECU	PAC	Richard et al. 1996	1	
C	NA	NA	GPG/ECU	PAC	Richard et al. 1996	17	

Table I.2 (Continued)

Haplotype (394 bp)	569bp	619bp	Region	Ocean	Publication	No. of individuals	Notes
B	NA	NA	GPG/ECU	PAC	Richard et al. 1996	2	
A	NA	NA	GPG/ECU	PAC	Richard et al. 1996	8	
C	NA	NA	GPG/ECU	PAC	Richard et al. 1996	3	
N	NA	NA	GPG/ECU	PAC	Richard et al. 1996	1	
O	NA	NA	GPG/ECU	PAC	Richard et al. 1996	3	
J	NA	NA	GPG/ECU	PAC	Richard et al. 1996	3	

Table I.3: PCR reaction conditions, dye-labels, repeat unit, size range, and reference for the 13 microsatellite loci used in this study. * Not explicitly stated

Locus	$MgCl_2$ [mM]	Polymerase [U]	Anneal temp	Dye label	Repeat unit	Size range	Reference
EV1	3.63	0.125	50°C	NED	2	118 – 142	(Valsecchi and Amos 1996)
EV5	1.36	0.25	59°C	FAM	2	148 – 174	(Valsecchi and Amos 1996)
EV14	2.27	0.125	50°C	VIC	2	121 - 155	(Valsecchi and Amos 1996)
EV37	3.18	0.25	56°C	NED	2	177 - 250	(Valsecchi and Amos 1996)
EV94	2.27	0.25	50°C	FAM	2	193 - 225	(Valsecchi and Amos 1996)
GATA417	2.27	0.125	50°C	FAM	4	172 - 202	(Palsbøll et al. 1997)
GT23	2.27	0.125	50°C	VIC	2	75 - 99	(Bérubé et al. 2000)
GT575	1.36	0.125	50°C	FAM	2	131 - 137	(Bérubé et al. 2000)
rw4-10	2.72	0.125	50°C	VIC	2	177 - 213	(Waldick et al. 1999)
SW13	1.81	0.25	57°C	VIC	2	134 - 176	(Richard et al. 1996b)
464/465	2.27	0.125	50°C	FAM	2	141 - 145	(Schlötterer et al. 1991)
SW19	1.36	0.25	56°C	NED	2	89 - 167	(Richard et al. 1996b)
FCB1	1.36	0.25	53°C	FAM	2*	107 - 145	(Buchanan et al. 1996)

Figure I.1: R code for carrying out permutations of haplotype diversity based on global (for comparing oceans) and oceanic (for comparing regions within oceans) haplotype frequencies. Sample sizes given in Table 2.1. The script is parameterized for up to six populations, but can be extended as needed.

```
MM <- rep.int(1,1)
BB <- rep.int(2,2)
N <- rep.int(3,8)
Y <- rep.int(4,19)
B <- rep.int(5,55)
A <- rep.int(6,69)
C <- rep.int(7,128)
X <- rep.int(8,157)
# Number of each haplotype in the background haplotype frequency to be
# sampled (e.g. global frequencies for oceanic samples, ocean
#frequencies for within-ocean samples). R needs these to be coded as
#'numerical' so haplotypes MM-X have been recoded here as 1-8.

Pop_1 <- 40
Pop_2 <- 14
Pop_3 <- 87
Pop_4 <- 230
Pop_5 <- 0
Pop_6 <- 0
# Sample size for each population that haplotype diversity will be
#estimated for

haplotypes <- c(MM,BB,N,Y,B,A,C,X)
# Combining the haplotypes defined above in one array of haplotypes

no_haps <- 8
# Total number of haplotypes (not sequences) in the 'haplotypes' array

i <- NULL
j <- NULL
# Parameterizing i and j for use in loops below

diff_Pop_12 <- c("diff_Pop_12")
diff_Pop_13 <- c("diff_Pop_13")
diff_Pop_14 <- c("diff_Pop_14")
diff_Pop_15 <- c("diff_Pop_15")
diff_Pop_16 <- c("diff_Pop_16")
diff_Pop_23 <- c("diff_Pop_23")
diff_Pop_24 <- c("diff_Pop_24")
diff_Pop_25 <- c("diff_Pop_25")
diff_Pop_26 <- c("diff_Pop_26")
diff_Pop_34 <- c("diff_Pop_34")
diff_Pop_35 <- c("diff_Pop_35")
diff_Pop_36 <- c("diff_Pop_36")
diff_Pop_45 <- c("diff_Pop_45")
diff_Pop_46 <- c("diff_Pop_46")
```

Figure I.1 (Continued)

```
diff_Pop_56 <- c("diff_Pop_56")
# Naming the array for each population comparison

for (i in 1:1000) {
# This sampling process will be completed 1000 times

a <- sample(haplotypes, Pop_1, replace=TRUE, prob=NULL)
b <- sample(haplotypes, Pop_2, replace=TRUE, prob=NULL)
cc <- sample(haplotypes, Pop_3, replace=TRUE, prob=NULL)
d <- sample(haplotypes, Pop_4, replace=TRUE, prob=NULL)
e <- sample(haplotypes, Pop_5, replace=TRUE, prob=NULL)
f <- sample(haplotypes, Pop_6, replace=TRUE, prob=NULL)
# For each replicate, take a sample from the background haplotype
# frequencies equal to the observed sample size for each population

pi_a <- 0
pi_b <- 0
pi_c <- 0
pi_d <- 0
pi_e <- 0
pi_f <- 0
# Initializing the 'pi' values for each replicate

j <- 1
while (j < no_haps+1) {
# While j is less than or equal to the total number of haplotypes in
#the population

pi_a <- pi_a + ((sum(a==j))/Pop_1)^2
pi_b <- pi_b + ((sum(b==j))/Pop_2)^2
pi_c <- pi_c + ((sum(cc==j))/Pop_3)^2
pi_d <- pi_d + ((sum(d==j))/Pop_4)^2
pi_e <- pi_e + ((sum(e==j))/Pop_5)^2
pi_f <- pi_f + ((sum(f==j))/Pop_6)^2

# Count the number of haplotypes in each population that equals
# haplotype code 'j'. Divide them by the total sample size (to get the
# frequency of that haplotype in the sample) and square this
# difference. Add this to the value calculated for other haplotypes.

j <- j+1
# The counter for the while loop goes up one (we move on and do these
# calculations for the next haplotype)
}

Pop_1H <- Pop_1/(Pop_1 - 1)*(1-pi_a)
Pop_2H <- Pop_2/(Pop_2 - 1)*(1-pi_b)
Pop_3H <- Pop_3/(Pop_3 - 1)*(1-pi_c)
```

Figure I.1 (Continued)

```
Pop_4H <- Pop_4/(Pop_4 - 1)*(1-pi_d)
Pop_5H <- Pop_5/(Pop_5 - 1)*(1-pi_e)
Pop_6H <- Pop_6/(Pop_6 - 1)*(1-pi_f)
# Using the values calculated in the while loop above, we calculate
# haplotype diversity for each of our samples

diff_Pop_12 <- rbind(diff_Pop_12, abs(Pop_1H - Pop_2H))
diff_Pop_13 <- rbind(diff_Pop_13, abs(Pop_1H - Pop_3H))
diff_Pop_14 <- rbind(diff_Pop_14, abs(Pop_1H - Pop_4H))
diff_Pop_15 <- rbind(diff_Pop_15, abs(Pop_1H - Pop_5H))
diff_Pop_16 <- rbind(diff_Pop_16, abs(Pop_1H - Pop_6H))
diff_Pop_23 <- rbind(diff_Pop_23, abs(Pop_2H - Pop_3H))
diff_Pop_24 <- rbind(diff_Pop_24, abs(Pop_2H - Pop_4H))
diff_Pop_25 <- rbind(diff_Pop_25, abs(Pop_2H - Pop_5H))
diff_Pop_26 <- rbind(diff_Pop_26, abs(Pop_2H - Pop_6H))
diff_Pop_34 <- rbind(diff_Pop_34, abs(Pop_3H - Pop_4H))
diff_Pop_35 <- rbind(diff_Pop_35, abs(Pop_3H - Pop_5H))
diff_Pop_36 <- rbind(diff_Pop_36, abs(Pop_3H - Pop_6H))
diff_Pop_45 <- rbind(diff_Pop_45, abs(Pop_4H - Pop_5H))
diff_Pop_46 <- rbind(diff_Pop_46, abs(Pop_4H - Pop_6H))
diff_Pop_56 <- rbind(diff_Pop_56, abs(Pop_5H - Pop_6H))
# We then record the differences in haplotype diversity between
# populations based on this random sampling procedure. The for (i)
# loop then resets and we do this for the number of permutations (1000)
# that we defined
}

result_table <- cbind(diff_Pop_12, diff_Pop_13, diff_Pop_14,
diff_Pop_15, diff_Pop_16, diff_Pop_23, diff_Pop_24, diff_Pop_25,
diff_Pop_26, diff_Pop_34, diff_Pop_35, diff_Pop_36, diff_Pop_45,
diff_Pop_46, diff_Pop_56)
# This summarizes the differences in haplotype diversity between each
# of our populations, for each of our 1,000 permutations

write.table(result_table, "haplotype_diversity_differences.txt",
sep="\t")
# Using this table of haplotype diversity differences based on the
# random sampling procedure, we assess the significance of our observed
# differences in haplotype diversity between populations
```

Figure I.2: R code for comparing microsatellite heterozygosity/allelic richness levels between different geographic partitions. Code courtesy of Rebecca Hamner (Hamner 2014).

```
data <- read.table("Data.txt", header=T)
#Data format in "Data.txt" is as follows for 3 populations (tab
#separated):
#  Pop_1   Pop_2   Pop_3
#  L1_ho   L1_ho   L1_ho
#  L2_ho   L2_ho   L2_ho
#  L3_ho   L3_ho   L3_ho
# Population names in row one, then observed heterozygosities/allelic
# richness in the rows below this
# Separate row for each microsatellite locus
# If additional populations, code below can be expanded

attach(data)
data

#test for equal variances
var.test(data$Pop_1, data$Pop_2)
var.test(data$Pop_2, data$Pop_3)
var.test(data$Pop_1, data$Pop_3)

#test for normality of differences between pairs
shapiro.test(data$Pop_1-data$Pop_2)
shapiro.test(data$Pop_2-data$Pop_3)
shapiro.test(data$Pop_1-data$Pop_3)

#t-test (if variances equal and differences between pairs are normal)
t.test(data$Pop_1, data$Pop_2, alternative = c("two.sided"), paired =
TRUE, var.equal=T)
t.test(data$Pop_2, data$Pop_3, alternative = c("two.sided"), paired =
TRUE, var.equal=T)
t.test(data$Pop_1, data$Pop_3, alternative = c("two.sided"), paired =
TRUE, var.equal=T)

#Wilcoxon signed rank test (if differences between pairs are NOT
# normal)
wilcox.test(data$Pop_1, data$Pop_2, alternative = c("two.sided"),
paired = TRUE)
```

Figure I.3: R code for measuring the nearest geographic distance between samples located in different geographic areas

```
#Make sure geosphere and aspaces packages are loaded and your input
# file has no headers
rm(list=ls()) # This just clears off variables that were already in
# your R space

temp_file <- read.csv(file="Input.csv",sep=",",
stringsAsFactors=FALSE,header=FALSE) #read in your csv file. The
#following column format is expected (but include no headers):
#Ocean    Region    Sample_name    Lat(decimal degrees)    Long(decimal
#degrees)
#Each row should have a separate sample listed. This script is designed
#to find the nearest distance between different regions located in the
#same ocean
maxcolumn <- dim(temp_file)[1]  #just getting some dimensions for the
#for loop below
maxrow <- dim(temp_file)[2]     #just getting some dimensions for the
# for loop below
output <- c("Sample_1", "Lat1", "Long1", "Sample_2", "Lat2", "Long2",
"Dist", "Area1", "Area2", "Comparison") #defining
#how many fields are in our output table "output"

for (i in 1:maxcolumn) { #for each sample
for (w in 1:maxcolumn) { #compare it to every other sample that...
if ((temp_file[i,2])!=(temp_file[w,2])) {  #is NOT from the same region
if ((temp_file[i,1])==(temp_file[w,1])) {  #but IS from the same ocean
#(because the nearest distance between oceans is over land, we manually
#calculate the nearest distance between oceans to ensure a non-
# terrestrial path)
if (i < w) {
distance <- acos(cos(as_radians(90-temp_file[w,4]))*cos(as_radians(90-
temp_file[i,4]))+sin(as_radians(90-temp_file[w,4]))*sin(as_radians(90-
temp_file[i,4]))*cos(as_radians(temp_file[w,5]-temp_file[i,5])))*6371
#calculating the distance between samples in km
temp <-
cbind(temp_file[i,3],temp_file[i,4],temp_file[i,5],temp_file[w,3],temp_
file[w,4],temp_file[w,5], distance, temp_file[i,2],temp_file[w,2],
temp_file[i,1]) #for each sample, binding together the fields we
# mentioned in output above
output <- rbind(output, temp) } } } } }  #binding all the rows together
#to make a full table

write.csv(output,file="output.csv") #writing results out
```

Figure I.4: 619 bp mtDNA CR analyses of diversity and differentiation.
(a) Maximum parsimony network based on 619 bp haplotype length dataset. Sequences of this length were not available for the following 394 bp haplotypes: F, G, K, L, P, T, W, U, V, X, Y, AA, EE, and OO. Haplotypes are colored by ocean. Size of haplotype pie is proportional on a log scale to the total number of individuals with the haplotype. Lines represent substitutions (one or two between haplotypes as defined by the key). New haplotypes over 394 bp characterized in this study are outlined in red, variants of 394 bp haplotypes split over the longer 619 bp fragment are outlined in light orange.

(b) Number of individuals in the 'full' and 'restricted' *Odyssey* dataset successfully sequenced for 619 bp mtDNA CR, number of haplotypes (*k*), haplotype diversity (*h*) and nucleotide diversity (in %, π). Data was not used from previous studies as haplotypes from these studies were defined over the shorter 394 bp length. In comparison with the 394 bp analyses conducted in this paper, the following regions were too limited in samples to include: western North Atlantic, Gulf of Mexico, Mediterranean, and Hawaii. "=" in the 'restricted' columns indicates the values are the same as the 'full' dataset.

Figure I.4 (Continued) (a)

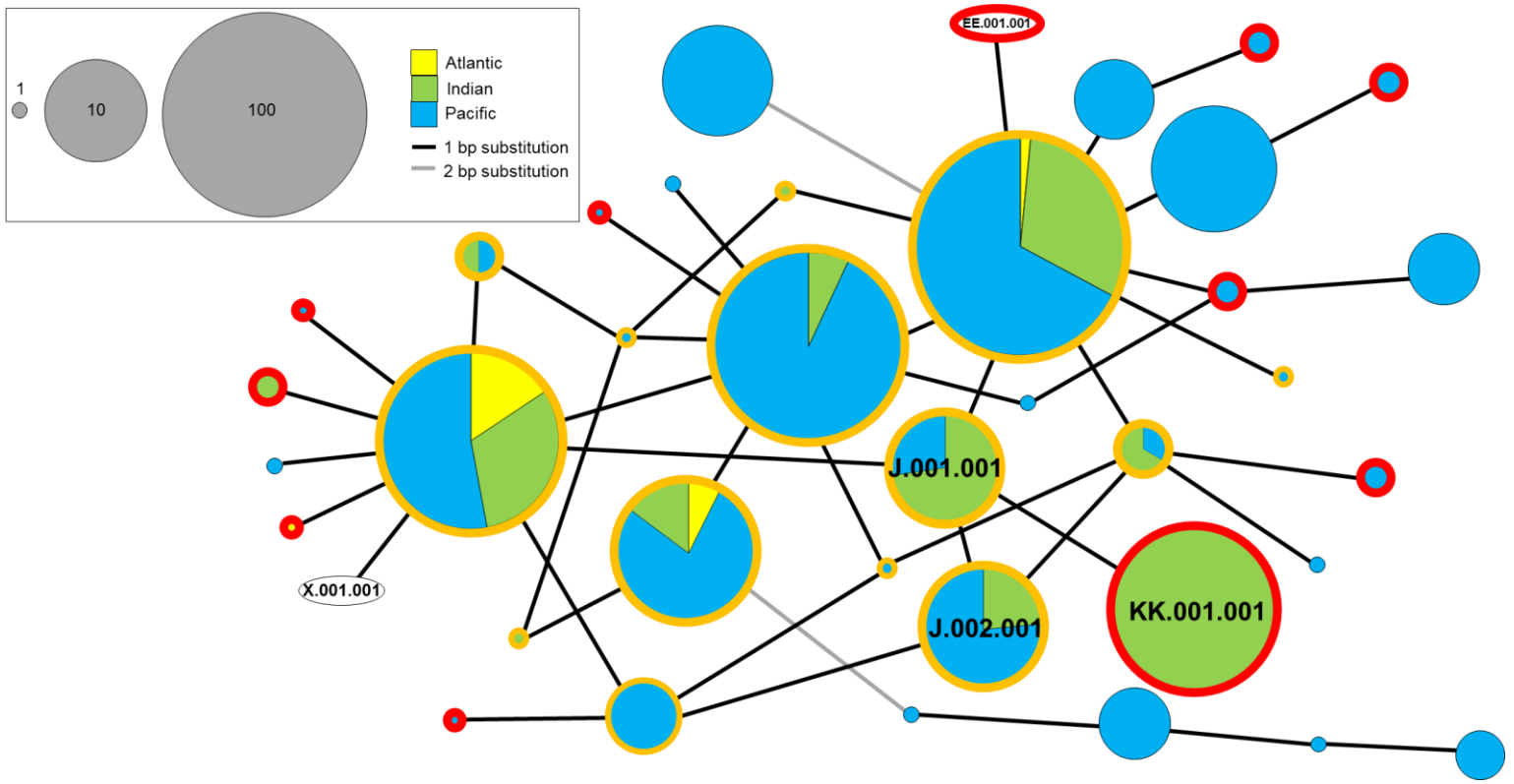


Figure I.4 (Continued) (b)

(b)	'Full' dataset				'Restricted' dataset			
Ocean/Region	**n**	***k***	***h***	**π (%)**	**n**	***k***	***h***	**π (%)**
Atlantic	13	4	0.615	0.192	=	=	=	=
Canary Islands	7	3	0.667	0.215	=	=	=	=
Indian	144	12	0.796	0.286	141	=	0.800	0.285
Aldabras	12	4	0.773	0.258	=	=	=	=
Cocos Island	18	3	0.451	0.146	=	=	=	=
Maldives/ Chagos Archipelago	29	4	0.591	0.200	28	=	0.601	0.204
Seychelles	26	7	0.745	0.328	=	=	=	=
Sri Lanka	37	2	0.315	0.0511	35	=	0.329	0.0534
Southwestern Australia	20	5	0.790	0.191	=	=	=	=
Pacific	243	23	0.807	0.257	237	=	0.808	0.256
Gulf of California	117	12	0.798	0.269	112	=	0.799	0.266
Galapagos	20	5	0.742	0.195	=	=	=	=
Kiribati	11	4	0.746	0.355	=	=	=	=
Papua New Guinea	59	9	0.760	0.219	58	=	0.764	0.221
Pacific Crossing	36	8	0.679	0.200	=	=	=	=
TOTAL	400	28	0.843	0.286	391	=	0.842	0.285

Figure I.5: Significance of comparisons of haplotype and nucleotide diversity values at various geographic levels. Significant comparisons for oceanic level given to left of table. Arrow points to ocean with highest genetic diversity for a given comparison (*h*: haplotype diversity, π: nucleotide diversity). Within-ocean comparisons are listed in table format by each ocean, with nucleotide diversity above the diagonal, haplotype diversity below the diagonal. No significant differences in nucleotide diversity were observed. Regions colored by within-ocean diversity level with dark green = highest haplotype/nucleotide diversity through yellow and red to brown-red = lowest haplotype/nucleotide diversity.

Figure I.5 (Continued)

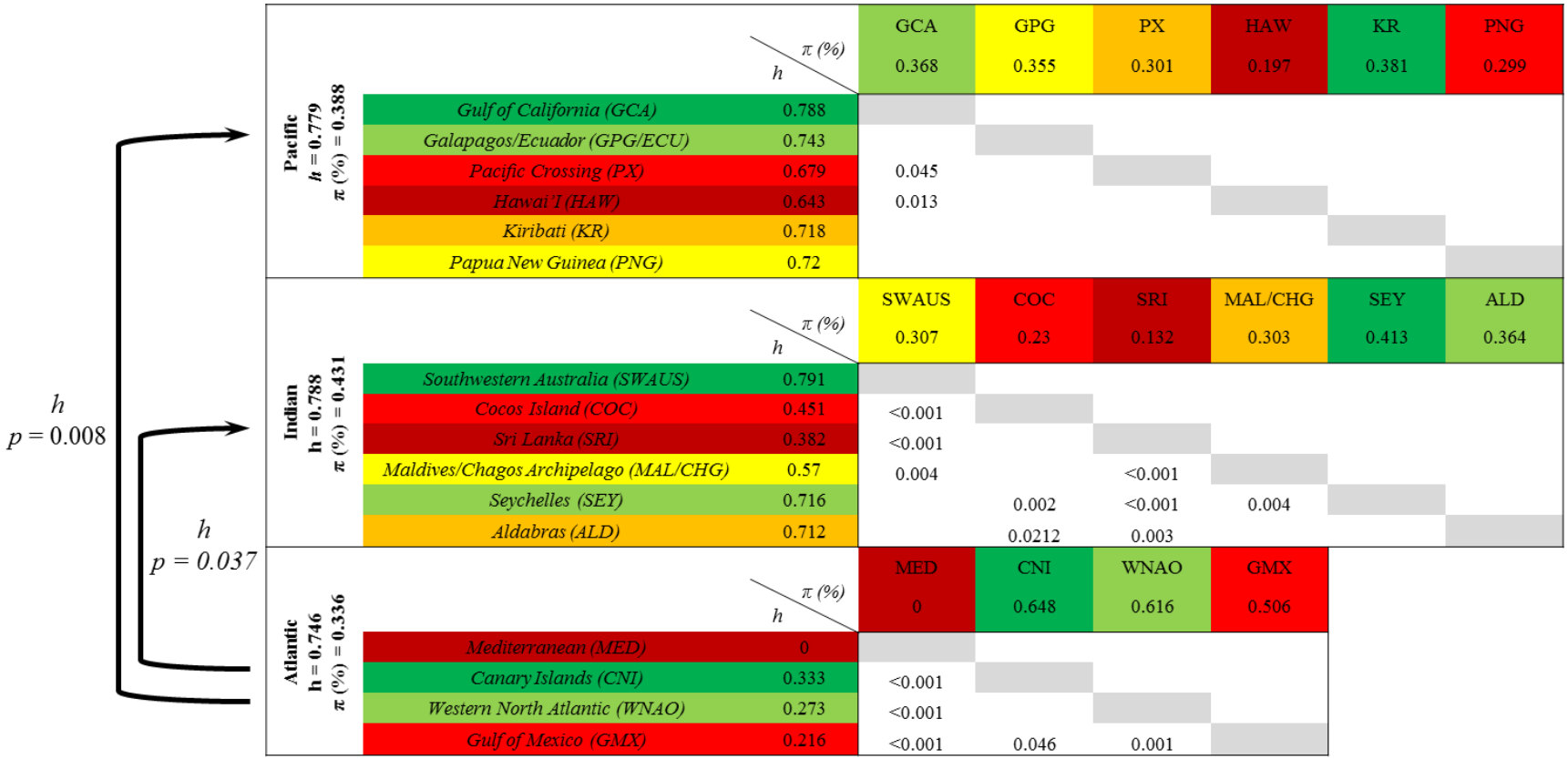


Table I.4: Regional pairwise comparisons of differentiation for mtDNA and microsatellites
(a) Pairwise mtDNA CR regional comparisons, with sample sizes given by '*n'*. F_{ST} above diagonal of matrix, Φ_{ST} underneath diagonal. Significant values italicized and shown in bold, with * <0.05, ** <0.001. Within-ocean comparisons are shown shaded in gray. Abbreviations: Gulf of California (GCA), Galapagos (GPG), 'Pacific Crossing' (PX), Hawai'i (HAW), Kiribati (KR), Papua New Guinea (PNG), SW Australia (SWAUS), Cocos Island (COC), Sri Lanka (SRI), Maldives/Chagos Archipelago (MAL), Seychelles (SEY), Aldabras (ALD), Mediterranean (MED), Canary Island (CNI), Western North Atlantic (WNAO), Gulf of Mexico (GMX).

(b) Pairwise microsatellite regional comparisons for 'full' dataset, with sample sizes given by '*n'*. F_{ST} above diagonal of matrix, G''_{ST} underneath diagonal. Significant values italicized and shown in bold, with * <0.05, ** <0.001. Within-ocean comparisons are shown shaded in gray. Abbreviations: Gulf of California (GCA), Galapagos (GPG), 'Pacific Crossing' (PX), Kiribati (KR), Papua New Guinea (PNG), SW Australia (SWAUS), Cocos Island (COC), Sri Lanka (SRI), Maldives/Chagos Archipelago (MAL), Seychelles (SEY), Aldabras (ALD), Mediterranean (MED), Canary Island (CNI).

(a)		Pacific						Indian						Atlantic			
	n	GCA	GPG	PX	HAW	KR	PNG	SWAUS	COC	SRI	MAL	SEY	ALD	MED	CNI	WNAO	GMX
GCA	122	--	***0.0513***	***0.1061***	***0.0484****	***0.1298***	***0.0673***	0.0331	***0.1850***	***0.3642***	***0.1737***	***0.2106***	***0.1524***	***0.4874***	***0.2222***	***0.1594***	***0.3264***
GPG	284	***0.0584***	--	***0.1003***	***0.0817****	***0.0755****	0.0029	0.0209	***0.1513***	***0.3815***	***0.1492***	***0.2051***	0.0515	***0.3419***	***0.1249****	***0.0357****	***0.3289***
PX	36	***0.0453****	***0.1601***	--	***0.0517****	***0.2785***	***0.0577****	***0.0783****	0.0547	***0.4713***	***0.0955****	***0.2459***	***0.1338****	***0.6534***	***0.2405***	***0.1431***	***0.4256***
HAW	28	0.0226	***0.0889****	0.03144	--	***0.2639***	***0.0655****	***0.0592****	***0.1015****	***0.4771***	***0.1239****	***0.2639***	***0.1801****	***0.7217***	***0.3013***	***0.1725***	***0.4249***
KR	13	***0.1798***	0.0501	***0.3270***	***0.2894***	--	***0.1246****	***0.1177****	***0.4117***	***0.4914***	***0.3520***	***0.2128***	***0.1147****	***0.5402***	***0.1123****	***0.1327****	***0.3804***
PNG	63	***0.0307****	0.0065	***0.1161***	0.0441	***0.1052****	--	0.0299	***0.1130****	***0.4330***	***0.1192***	***0.2265***	0.0377	***0.4533***	***0.1423****	0.0216	***0.3781***
SWAUS	21	***0.0678****	0.0149	***0.1806***	***0.1608****	***0.1130****	0.0431	--	***0.1074****	***0.4384***	***0.1481****	***0.0950****	***0.0913****	***0.6374***	***0.0924****	***0.0998****	***0.3660***
COC	18	0.0328	***0.1029****	0.04561	0.0497	***0.3017***	***0.0873****	0.0699	--	***0.5901***	***0.1050****	***0.2875***	***0.2224****	***0.8413***	***0.3105***	***0.1878***	***0.5007***
SRI	42	***0.5464***	***0.5140***	***0.6374***	***0.6583***	***0.6450***	***0.5795***	***0.6814***	***0.7074***	--	***0.3675***	***0.4328***	***0.3761***	***0.8051***	***0.5197***	***0.4804***	***0.5322***
MAL	33	***0.1632***	***0.2045***	***0.1670***	***0.11421****	***0.3092***	***0.1747***	***0.2650****	***0.1891****	***0.3856***	--	***0.3156***	0.0931	***0.7185***	***0.3122***	***0.1887***	***0.4660***
SEY	31	***0.2466***	***0.1482***	***0.3290***	***0.3123***	***0.1365****	***0.2019***	0.0594	***0.2181****	***0.6264***	***0.3456***	--	***0.2161***	***0.6082***	0.0717	***0.2628***	***0.4174***
ALD	12	***0.1822***	***0.1108****	***0.2705***	***0.2092****	0.0834	***0.1147****	***0.1934****	***0.2808****	***0.4020***	0.0857	***0.2104****	--	***0.6133***	0.0866	0.0159	***0.4151***
MED	40	***0.5148***	***0.3448***	***0.7393***	***0.7957***	***0.4577***	***0.5183***	***0.6749***	***0.8416***	***0.8523***	***0.7062***	***0.4845***	***0.6122***	--	***0.5412***	***0.3502***	***0.6099***
CNI	14	***0.2439***	***0.1036****	***0.3782***	***0.3775***	0.0551	***0.1842****	0.0692	***0.3095****	***0.6926***	***0.3818***	0.0000	***0.1853****	***0.5418***	--	***0.1091****	***0.4329***
WNAO	87	***0.1173***	***0.0183****	***0.2441***	***0.1487****	0.0346	0.0289	***0.0847****	***0.2073***	***0.5592***	***0.2297***	***0.1803***	0.0628	***0.3785***	***0.1234****	--	***0.4148***
GMX	230	***0.5825***	***0.4706***	***0.6766***	***0.6481***	***0.4543***	***0.5540***	***0.5722***	***0.6687***	***0.7258***	***0.6532***	***0.5362***	***0.5241***	***0.4557***	***0.4831***	***0.4866***	--

Table I.4 (Continued) (b)

(b)	*n*	Pacific Ocean					Indian						Atlantic	
		GCA	GPG	PX	KR	PNG	SWAUS	COC	SRI	MAL	SEY	ALD	MED	CNI
GCA	122	--	0.0000	0.0000	***0.0080****	***0.0009****	***0.0085****	0.0103	***0.0041****	0.0023	0.0000	0.0107	***0.0184****	0.0022
GPG	23	0.0000	--	0.0000	0.0123	***0.0016****	0.0078	***0.0164****	***0.0046****	0.0041	0.0000	0.011	***0.0166****	0.0088
PX	37	0.0000	0.0000	--	0.0039	***0.0044****	0.0018	0.0046	***0.0035****	0.0000	0.0000	0.0016	***0.0202****	0.0008
KR	13	***0.0288****	0.0515	0.0135	--	***0.0055****	***0.0201****	0.0000	***0.0143****	0.0017	0.0062	0.0152	***0.0181****	0.0004
PNG	65	***0.0029****	***0.0037****	***0.0147****	***0.0191****	--	***0.0110****	***0.0098****	***0.0075****	***0.0056****	***0.0026****	***0.0157****	***0.0213****	***0.0033****
SWAUS	21	***0.0269****	0.0308	0.0038	***0.0691****	***0.0368****	--	0.0179	***0.0088****	0.0138	0.0078	0.0086	***0.0232****	***0.0190****
COC	18	0.0327	***0.0522****	0.0132	0.0000	***0.0307****	0.0539	--	***0.0071****	0.0027	0.0000	0.0072	***0.0231****	0.0068
SRI	56	***0.0143****	***0.0202****	***0.0112****	***0.0524****	***0.0252****	***0.0246****	***0.0229****	--	***0.0078****	0.0000	0.0015	***0.0260****	***0.0070****
MAL	34	0.0082	0.0256	0.0000	0.0052	***0.0201****	0.0525	0.0086	***0.0292****	--	0.0000	0.0046	***0.0196****	0.0003
SEY	31	0.0000	0.0000	0.0000	0.0223	***0.008****	0.0230	0.0000	0.0002	0.0000	--	0.0000	***0.0144****	0.0033
ALD	12	0.0389	0.0461	0.0064	0.0559	***0.0568****	0.0254	0.0239	0.0059	0.0186	0.0000	--	***0.0268****	0.0144
MED	9	***0.0626****	***0.0604****	***0.0696****	***0.0675****	***0.0712****	***0.0781****	***0.0763****	***0.0827****	***0.0682****	***0.0489****	***0.0860****	--	***0.0161****
CNI	25	0.0079	0.0299	0.0030	0.0011	***0.0116****	***0.0622****	0.0213	***0.0234****	0.0015	0.0112	0.0508	***0.0524****	--

Figure I.6: Genetic landscape based on mtDNA data for *Odyssey* samples where specific spatial coordinates were available. Peaks indicate barriers to gene flow, with troughs indicating areas with greater similarity after accounting for isolation by distance effects. Landscape features corresponding to *a priori* geographic groupings are indicated by labels. (a) and (b) show the same plot rotated to give alternate visual perspectives of the landscape.

Figure I.6 (a) Continued

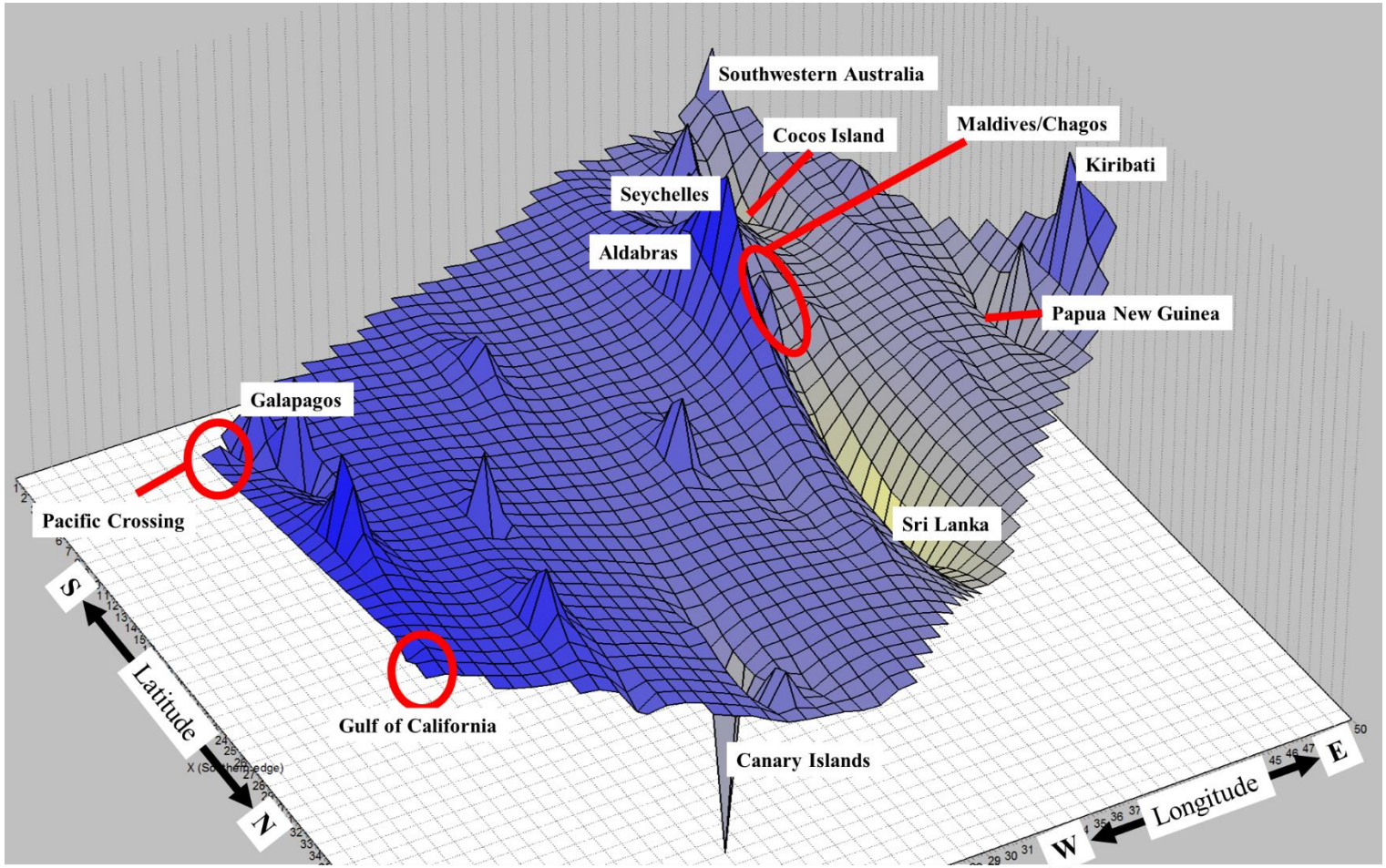


Figure I.6 (b) Continued

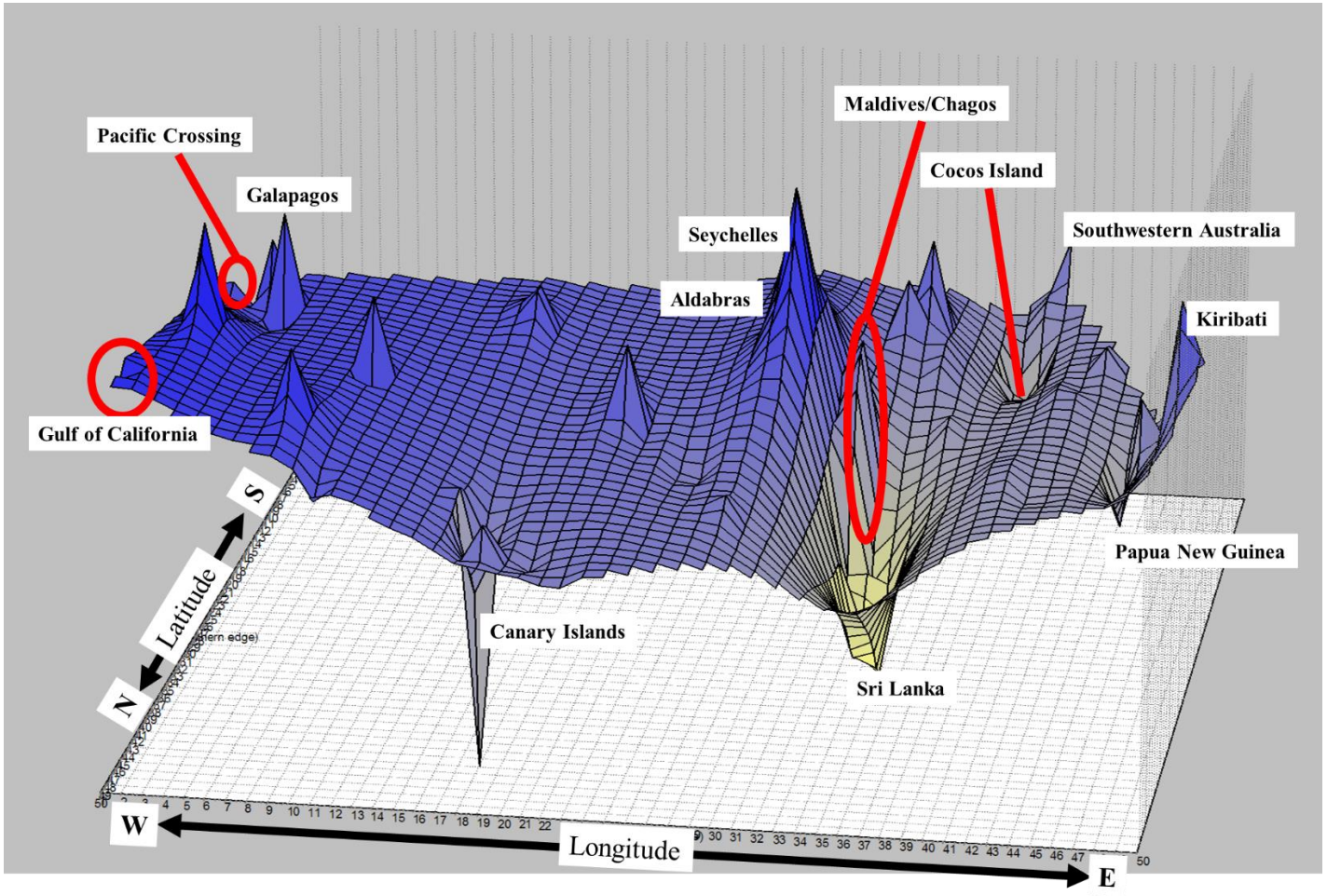


Table I.5: Diversity results by microsatellite locus and for the 'restricted' dataset
(a) Levels of microsatellite diversity and sex-identification. Values for the 'full' dataset are given, with 'restricted' dataset values given in parentheses. As for the full dataset, no significant differences in heterozygosity were detected between oceans, or between regions within oceans for the restricted dataset.

(b) Locus by locus observed and expected heterozygosity results for the 13 microsatellite loci used in this study for both the 'full' and 'restricted' datasets.

(c) Microsatellite differentiation indices for 'full' and 'restricted' datasets.

(d) Pairwise microsatellite regional comparisons for 'restricted' dataset. F_{ST} above diagonal of matrix, G''_{ST} underneath diagonal. Significant values italicized and shown in bold, with * <0.05, ** <0.001. Within-ocean comparisons are shown shaded in gray. Abbreviations: Gulf of California (GCA), Galapagos (GPG), 'Pacific Crossing' (PX), Kiribati (KR), Papua New Guinea (PNG), SW Australia (SWAUS), Cocos Island (COC), Sri Lanka (SRI), Maldives/Chagos Archipelago (MAL), Seychelles (SEY), Aldabras (ALD), Mediterranean (MED), Canary Island (CNI).

(e) Results for the sex-specific F_{ST} comparison by marker. Regional analyses were limited to areas with more than two identified females and males: the Canary Islands, Aldabras, southwestern Australia, Maldives/Chagos Archipelago, Seychelles, Sri Lanka, Gulf of California, Kiribati, Papua New Guinea, and the Pacific Crossing. Due to limited sample sizes, a within-ocean regional analysis was not conducted for the Atlantic. Statistically significant values are bolded and italicized, with * significant at $p < 0.05$, ** significant at $p < 0.001$.

	Microsatellites				Sex	
(a)	No of individuals	Allelic richness	Mean proportion of individuals typed for loci	Mean observed heterozygosity	No. of individuals	Sex ratio (F: M)
Atlantic	41 (39)	8.817 (8.731)	0.9681 (0.9665)	0.6685 (0.6694)	36 (34)	1.12 (1.00)
Canary Islands	*25 (24)*	*5.155 (5.213)*	*0.9754 (0.9744)*	*0.6897 (0.6993)*	*24 (23)*	*1.70 (1.56)*
Mediterranean	*9 (=)*	*4.597 (=)*	*0.9573 (=)*	*0.6307 (=)*	*9 (=)*	*0.13 (=)*
Indian	175 (171)	8.838 (8.722)	0.9178 (0.9163)	0.6863 (0.6841)	132 (128)	3.40 (3.41)
Aldabras	*12 (=)*	*5.147 (=)*	*0.9487 (=)*	*0.6774(=)*	*8 (=)*	*3.00 (=)*
Cocos Island	*18 (=)*	*5.243 (=)*	*0.9359 (=)*	*0.7121 (=)*	*18 (=)*	*F only (=)*
Maldives/Chagos Archipelago	*34 (33)*	*5.315 (5.285)*	*0.9367 (0.9347)*	*0.7002 (0.6926)*	*24 (23)*	*0.60 (0.64)*
Seychelles	*31 (=)*	*5.317 (=)*	*0.8859 (=)*	*0.6973 (=)*	*19 (=)*	*8.50 (=)*
Sri Lanka	*56 (53)*	*5.073 (5.058)*	*0.9272 (0.9245)*	*0.6709 (0.6678)*	*48 (45)*	*7.00 (6.50)*
Southwestern Australia	*21 (=)*	*5.148(=)*	*0.8791 (=)*	*0.6968 (=)*	*13 (=)*	*2.25 (=)*
Pacific	260 (254)	8.870 (8.767)	0.8970 (0.8958)	0.6944 (0.6935)	247 (241)	2.68 (2.60)
Galapagos	*23 (=)*	*5.089 (=)*	*0.8361 (=)*	*0.6770 (=)*	*23 (=)*	*M only (=)*
Gulf of California	*122 (117)*	*5.203 (5.217)*	*0.8909 (0.8889)*	*0.7022 (0.7018)*	*116 (111)*	*4.80 (4.55)*
Kiribati	*13 (=)*	*5.328 (=)*	*0.9112 (=)*	*0.6836 (=)*	*12 (=)*	*5.00 (=)*
"Pacific crossing"	*37 (=)*	*5.101 (=)*	*0.9023 (=)*	*0.7040 (=)*	*34 (=)*	*1 (=)*
Papua New Guinea	*65 (64)*	*5.104 (5.103)*	*0.9243 (0.9231)*	*0.6867 (0.6853)*	*62 (61)*	*6.75 (6.63)*
Total	476 (464)	13.07 (13.03)	0.9108 (0.9093)	0.6896 (0.6884)	415 (403)	2.64 (2.57)

Table I.5 (Continued) (b)

		EV1				EV5				EV94				GT23				GT575				rw4-10				464/465			
"All" Dataset	*n*	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp
Total	541	12	520	0.6	0.641	11	528	0.71	0.707	17	533	0.818	0.804	12	522	0.508	0.497	4	486	0.609	0.61	14	460	0.72	0.768	3	403	0.529	0.54
Atlantic	41	9	41	0.585	0.65	9	40	0.75	0.744	12	41	0.902	0.823	6	38	0.368	0.404	4	41	0.732	0.636	5	40	0.65	0.744	3	39	0.462	0.553
"Atlantic crossing"	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*0*
Bahamas	*3*	*3*	*3*	*0.667*	*0.6*	*3*	*3*	*1*	*0.733*	*3*	*3*	*0.667*	*0.6*	*4*	*3*	*1*	*0.8*	*2*	*3*	*1*	*0.6*	*3*	*3*	*0.667*	*0.6*	*2*	*2*	*0.5*	*0.5*
Central Atlantic	*3*	*3*	*3*	*0.667*	*0.733*	*2*	*3*	*0.333*	*0.333*	*3*	*3*	*1*	*0.733*	*2*	*3*	*0.333*	*0.333*	*2*	*3*	*0.667*	*0.533*	*2*	*3*	*0.333*	*0.333*	*2*	*3*	*0.333*	*0.333*
Canary Islands	*25*	*8*	*25*	*0.48*	*0.618*	*8*	*24*	*0.792*	*0.772*	*10*	*25*	*0.92*	*0.838*	*3*	*23*	*0.261*	*0.305*	*4*	*25*	*0.76*	*0.672*	*5*	*24*	*0.75*	*0.734*	*2*	*24*	*0.5*	*0.507*
Eastern Mediterranean	*1*	*2*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*2*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*0*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*
Central Mediterranean	*3*	*3*	*3*	*1*	*0.733*	*4*	*3*	*0.667*	*0.867*	*4*	*3*	*1*	*0.867*	*2*	*2*	*0.5*	*0.5*	*3*	*3*	*1*	*0.733*	*3*	*3*	*1*	*0.733*	*2*	*3*	*0.333*	*0.6*
Northern Mediterranean	*5*	*4*	*5*	*0.6*	*0.533*	*4*	*5*	*0.8*	*0.711*	*7*	*5*	*0.8*	*0.933*	*3*	*5*	*0.6*	*0.511*	*3*	*5*	*0.6*	*0.689*	*3*	*5*	*0.2*	*0.511*	*2*	*5*	*0.4*	*0.533*
Indian	175	11	168	0.571	0.605	11	171	0.655	0.657	13	175	0.846	0.805	7	171	0.515	0.487	4	153	0.516	0.589	12	151	0.762	0.76	3	124	0.532	0.533
Aldabras	*12*	*6*	*12*	*0.5*	*0.554*	*5*	*11*	*0.545*	*0.472*	*6*	*12*	*1*	*0.772*	*3*	*11*	*0.455*	*0.385*	*3*	*12*	*0.417*	*0.638*	*5*	*12*	*0.583*	*0.75*	*3*	*11*	*0.455*	*0.558*
Southwestern Australia	*21*	*5*	*21*	*0.524*	*0.585*	*7*	*20*	*0.75*	*0.774*	*8*	*21*	*0.905*	*0.819*	*5*	*21*	*0.762*	*0.581*	*3*	*17*	*0.471*	*0.392*	*6*	*14*	*0.857*	*0.783*	*3*	*9*	*0.556*	*0.542*
Cocos Island	*18*	*5*	*18*	*0.5*	*0.47*	*8*	*18*	*0.611*	*0.706*	*9*	*18*	*0.778*	*0.783*	*3*	*17*	*0.647*	*0.528*	*4*	*17*	*0.412*	*0.586*	*7*	*16*	*0.813*	*0.748*	*3*	*12*	*0.583*	*0.583*
"Indian crossing"	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*
Maldives/Chagos Archipelago	*34*	*10*	*32*	*0.688*	*0.673*	*8*	*34*	*0.735*	*0.684*	*10*	*34*	*0.794*	*0.827*	*5*	*34*	*0.5*	*0.525*	*4*	*30*	*0.7*	*0.636*	*10*	*33*	*0.727*	*0.774*	*3*	*25*	*0.52*	*0.611*
Mauritius	*2*	*1*	*2*	*0*	*0*	*2*	*1*	*1*	*1*	*3*	*2*	*0.5*	*0.833*	*1*	*2*	*0*	*0*	*3*	*2*	*0.5*	*0.833*	*3*	*2*	*1*	*0.833*	*1*	*1*	*0*	*0*
Seychelles	*31*	*8*	*29*	*0.448*	*0.538*	*6*	*31*	*0.613*	*0.641*	*10*	*31*	*0.839*	*0.795*	*4*	*30*	*0.467*	*0.49*	*4*	*23*	*0.522*	*0.622*	*7*	*24*	*0.792*	*0.776*	*3*	*22*	*0.636*	*0.542*
Sri Lanka	*56*	*9*	*53*	*0.642*	*0.669*	*7*	*55*	*0.618*	*0.611*	*10*	*56*	*0.857*	*0.804*	*5*	*55*	*0.436*	*0.451*	*4*	*52*	*0.481*	*0.541*	*9*	*49*	*0.755*	*0.755*	*3*	*43*	*0.488*	*0.475*
Pacific	325	12	311	0.617	0.658	9	317	0.735	0.723	15	317	0.792	0.8	9	313	0.521	0.513	4	292	0.64	0.617	13	269	0.706	0.774	3	240	0.538	0.54
Gulf of California	*122*	*12*	*117*	*0.65*	*0.672*	*9*	*121*	*0.727*	*0.728*	*12*	*122*	*0.836*	*0.806*	*4*	*119*	*0.462*	*0.446*	*4*	*111*	*0.649*	*0.603*	*10*	*93*	*0.72*	*0.769*	*3*	*88*	*0.534*	*0.54*
Galapagos	*23*	*9*	*23*	*0.739*	*0.733*	*6*	*23*	*0.652*	*0.596*	*9*	*23*	*0.696*	*0.795*	*5*	*23*	*0.565*	*0.542*	*4*	*16*	*0.5*	*0.694*	*7*	*12*	*1*	*0.833*	*2*	*7*	*0.571*	*0.44*
Kiribati	*13*	*8*	*13*	*0.692*	*0.757*	*7*	*13*	*0.538*	*0.76*	*5*	*12*	*0.75*	*0.743*	*4*	*12*	*0.333*	*0.605*	*4*	*12*	*0.667*	*0.667*	*6*	*11*	*0.636*	*0.727*	*3*	*8*	*0.75*	*0.592*
New Zealand	*62*	*9*	*55*	*0.545*	*0.581*	*8*	*59*	*0.797*	*0.715*	*13*	*59*	*0.797*	*0.832*	*8*	*59*	*0.695*	*0.61*	*4*	*60*	*0.683*	*0.63*	*9*	*60*	*0.7*	*0.799*	*3*	*60*	*0.55*	*0.549*
Oregon	*2*	*2*	*2*	*0.5*	*0.5*	*4*	*2*	*1*	*1*	*4*	*2*	*1*	*1*	*2*	*2*	*0.5*	*0.5*	*2*	*2*	*1*	*0.667*	*3*	*2*	*1*	*0.833*	*2*	*2*	*0.5*	*0.5*
Papua New Guinea	*65*	*9*	*63*	*0.587*	*0.677*	*7*	*63*	*0.73*	*0.744*	*9*	*63*	*0.762*	*0.746*	*5*	*61*	*0.557*	*0.527*	*4*	*59*	*0.61*	*0.63*	*6*	*59*	*0.61*	*0.739*	*3*	*48*	*0.542*	*0.504*
"Pacific crossing"	*37*	*8*	*37*	*0.568*	*0.578*	*8*	*35*	*0.771*	*0.732*	*10*	*35*	*0.771*	*0.828*	*5*	*36*	*0.417*	*0.512*	*4*	*31*	*0.645*	*0.561*	*8*	*31*	*0.742*	*0.781*	*3*	*26*	*0.423*	*0.548*
Samoa	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*0*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*

Table I.5 (Continued) (b)

	GATA417				SW13				EV37				EV14				FCB1				SW19			
"All" Dataset	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp
Total	7	437	0.508	0.532	14	522	0.826	0.836	32	502	0.855	0.905	14	481	0.688	0.717	16	517	0.793	0.836	32	506	0.879	0.921
Atlantic	3	41	0.39	0.401	10	41	0.78	0.835	17	39	0.795	0.885	11	40	0.7	0.734	10	38	0.684	0.817	17	37	0.892	0.929
"Atlantic crossing"	*1*	*1*	*0*	*0*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*1*	*1*	*0*	*0*
Bahamas	*1*	*3*	*0*	*0*	*4*	*3*	*1*	*0.867*	*2*	*2*	*0.5*	*0.5*	*2*	*3*	*0.333*	*0.333*	*4*	*3*	*0.333*	*0.867*	*4*	*2*	*1*	*1*
Central Atlantic	*2*	*3*	*0.333*	*0.333*	*3*	*3*	*0.667*	*0.6*	*5*	*3*	*1*	*0.933*	*3*	*2*	*1*	*0.833*	*4*	*3*	*0.667*	*0.8*	*5*	*3*	*1*	*0.933*
Canary Islands	*3*	*25*	*0.44*	*0.448*	*8*	*25*	*0.68*	*0.829*	*16*	*24*	*0.833*	*0.898*	*9*	*25*	*0.8*	*0.733*	*9*	*24*	*0.833*	*0.803*	*16*	*24*	*0.917*	*0.934*
Eastern Mediterranean	*1*	*1*	*0*	*0*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Central Mediterranean	*2*	*3*	*0.333*	*0.333*	*3*	*3*	*1*	*0.733*	*2*	*3*	*0.333*	*0.333*	*4*	*3*	*0.667*	*0.8*	*4*	*3*	*0.667*	*0.867*	*4*	*2*	*1*	*1*
Northern Mediterranean	*3*	*5*	*0.6*	*0.6*	*5*	*5*	*1*	*0.844*	*6*	*5*	*0.8*	*0.844*	*2*	*5*	*0.2*	*0.556*	*2*	*4*	*0*	*0.429*	*5*	*5*	*0.8*	*0.756*
Indian	5	152	0.572	0.578	12	167	0.814	0.843	24	162	0.833	0.907	12	161	0.702	0.734	14	167	0.743	0.825	27	166	0.861	0.921
Aldabras	*4*	*10*	*0.8*	*0.726*	*9*	*11*	*0.818*	*0.9*	*10*	*10*	*0.9*	*0.921*	*8*	*12*	*0.667*	*0.659*	*8*	*12*	*0.75*	*0.786*	*11*	*12*	*0.917*	*0.92*
Southwestern Australia	*3*	*15*	*0.667*	*0.632*	*9*	*20*	*0.6*	*0.629*	*15*	*21*	*0.762*	*0.931*	*6*	*20*	*0.55*	*0.662*	*8*	*20*	*0.75*	*0.851*	*14*	*21*	*0.905*	*0.935*
Cocos Island	*3*	*15*	*0.6*	*0.508*	*9*	*18*	*0.889*	*0.881*	*13*	*17*	*0.941*	*0.914*	*8*	*18*	*0.944*	*0.833*	*8*	*17*	*0.706*	*0.824*	*12*	*18*	*0.833*	*0.905*
"Indian crossing"	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*1*	*1*
Maldives/Chagos Archipelago	*3*	*32*	*0.5*	*0.555*	*9*	*33*	*0.818*	*0.811*	*18*	*30*	*0.933*	*0.888*	*9*	*31*	*0.613*	*0.756*	*12*	*32*	*0.75*	*0.831*	*14*	*34*	*0.824*	*0.913*
Mauritius	*2*	*2*	*0.5*	*0.5*	*3*	*2*	*0.5*	*0.833*	*2*	*2*	*0*	*0.667*	*2*	*1*	*1*	*1*	*2*	*2*	*0.5*	*0.5*	*2*	*2*	*0*	*0.667*
Seychelles	*3*	*24*	*0.583*	*0.542*	*10*	*31*	*0.871*	*0.874*	*16*	*29*	*0.897*	*0.926*	*10*	*29*	*0.724*	*0.728*	*12*	*28*	*0.75*	*0.855*	*15*	*26*	*0.923*	*0.919*
Sri Lanka	*4*	*53*	*0.528*	*0.594*	*10*	*51*	*0.843*	*0.849*	*19*	*52*	*0.75*	*0.895*	*10*	*49*	*0.714*	*0.713*	*11*	*55*	*0.745*	*0.814*	*19*	*52*	*0.865*	*0.917*
Pacific	7	244	0.488	0.519	14	314	0.838	0.832	29	301	0.874	0.905	14	280	0.679	0.703	16	312	0.833	0.842	29	303	0.888	0.919
Gulf of California	*3*	*77*	*0.429*	*0.515*	*14*	*117*	*0.855*	*0.826*	*24*	*116*	*0.888*	*0.909*	*12*	*102*	*0.647*	*0.69*	*14*	*116*	*0.845*	*0.855*	*26*	*114*	*0.886*	*0.918*
Galapagos	*3*	*12*	*0.333*	*0.366*	*9*	*22*	*0.773*	*0.813*	*16*	*23*	*0.826*	*0.916*	*8*	*22*	*0.545*	*0.648*	*7*	*23*	*0.696*	*0.807*	*13*	*21*	*0.905*	*0.904*
Kiribati	*3*	*12*	*0.25*	*0.62*	*10*	*13*	*0.923*	*0.886*	*10*	*13*	*0.923*	*0.898*	*6*	*10*	*0.9*	*0.8*	*8*	*13*	*0.692*	*0.803*	*11*	*12*	*0.833*	*0.902*
New Zealand	*6*	*57*	*0.649*	*0.554*	*11*	*61*	*0.77*	*0.849*	*20*	*54*	*0.87*	*0.9*	*10*	*52*	*0.731*	*0.714*	*13*	*57*	*0.86*	*0.844*	*20*	*54*	*0.926*	*0.935*
Oregon	*2*	*2*	*1*	*0.667*	*3*	*2*	*1*	*0.833*	*4*	*2*	*1*	*1*	*1*	*1*	*0*	*0*	*2*	*2*	*0.5*	*0.5*	*4*	*2*	*1*	*1*
Papua New Guinea	*4*	*54*	*0.481*	*0.463*	*9*	*63*	*0.81*	*0.82*	*19*	*61*	*0.82*	*0.898*	*11*	*60*	*0.717*	*0.706*	*12*	*64*	*0.844*	*0.842*	*18*	*63*	*0.857*	*0.919*
"Pacific crossing"	*3*	*29*	*0.483*	*0.56*	*12*	*35*	*0.943*	*0.846*	*17*	*32*	*0.938*	*0.891*	*9*	*33*	*0.667*	*0.7*	*12*	*37*	*0.892*	*0.822*	*17*	*37*	*0.892*	*0.908*
Samoa	*1*	*1*	*0*	*0*	*2*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*

Table I.5 (Continued) (b)

	Locus	EV1				EV5				EV94				GT23				GT575				rw4-10				464/465			
"Restricted" Dataset	Measures	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp
Total	526	12	505	0.594	0.637	11	513	0.702	0.706	17	518	0.822	0.805	12	508	0.504	0.497	4	471	0.609	0.611	14	445	0.719	0.769	3	388	0.528	0.54
Atlantic	39	9	39	0.564	0.647	9	38	0.737	0.748	12	39	0.897	0.827	6	36	0.389	0.422	4	39	0.769	0.641	5	38	0.632	0.747	3	37	0.459	0.554
"Atlantic crossing"	*As for "all" dataset*																												
Bahamas	*As for "all" dataset*																												
Central Atlantic	*2*	*2*	*2*	*0.5*	*0.5*	*1*	*2*	*0*	*0*	*2*	*2*	*1*	*0.667*	*2*	*2*	*0.5*	*0.5*	*2*	*2*	*1*	*0.667*	*1*	*2*	*0*	*0*	*1*	*2*	*0*	*0*
Canary Islands	*24*	*8*	*24*	*0.458*	*0.616*	*8*	*23*	*0.783*	*0.781*	*10*	*24*	*0.917*	*0.841*	*3*	*22*	*0.273*	*0.317*	*4*	*24*	*0.792*	*0.675*	*5*	*23*	*0.739*	*0.739*	*2*	*23*	*0.522*	*0.51*
Eastern Mediterranean	*As for "all" dataset*																												
Central Mediterranean	*As for "all" dataset*																												
Northern Mediterranean	*As for "all" dataset*																												
Indian	171	11	164	0.567	0.592	11	167	0.647	0.655	13	171	0.854	0.804	7	167	0.503	0.483	4	149	0.51	0.59	12	147	0.762	0.758	3	120	0.542	0.535
Aldabras	*As for "all" dataset*																												
Southwestern Australia	*As for "all" dataset*																												
Cocos Island	*As for "all" dataset*																												
"Indian crossing"	*As for "all" dataset*																												
Maldives/Chagos Archipelago	*33*	*10*	*31*	*0.677*	*0.655*	*8*	*33*	*0.727*	*0.675*	*10*	*33*	*0.818*	*0.828*	*5*	*33*	*0.485*	*0.523*	*4*	*29*	*0.69*	*0.639*	*10*	*32*	*0.719*	*0.774*	*3*	*24*	*0.5*	*0.614*
Mauritius	*As for "all" dataset*																												
Seychelles	*As for "all" dataset*																												
Sri Lanka	*53*	*9*	*50*	*0.64*	*0.649*	*7*	*52*	*0.596*	*0.61*	*10*	*53*	*0.868*	*0.797*	*5*	*52*	*0.404*	*0.436*	*4*	*49*	*0.469*	*0.541*	*9*	*46*	*0.761*	*0.752*	*3*	*40*	*0.525*	*0.477*
Pacific	316	12	302	0.613	0.658	9	308	0.727	0.721	15	308	0.795	0.801	9	305	0.518	0.513	4	283	0.64	0.617	13	260	0.708	0.776	3	231	0.532	0.538
Gulf of California	*117*	*12*	*112*	*0.634*	*0.669*	*9*	*116*	*0.716*	*0.722*	*12*	*117*	*0.829*	*0.804*	*4*	*115*	*0.461*	*0.447*	*4*	*106*	*0.651*	*0.603*	*10*	*88*	*0.727*	*0.777*	*3*	*83*	*0.518*	*0.537*
Galapagos	*As for "all" dataset*																												
Kiribati	*As for "all" dataset*																												
New Zealand	*59*	*9*	*52*	*0.538*	*0.584*	*8*	*56*	*0.786*	*0.719*	*13*	*56*	*0.839*	*0.84*	*8*	*56*	*0.696*	*0.614*	*4*	*57*	*0.667*	*0.631*	*9*	*57*	*0.702*	*0.798*	*3*	*57*	*0.561*	*0.553*
Oregon	*As for "all" dataset*																												
Papua New Guinea	*64*	*9*	*62*	*0.597*	*0.683*	*7*	*62*	*0.726*	*0.742*	*9*	*62*	*0.758*	*0.745*	*5*	*60*	*0.55*	*0.525*	*4*	*58*	*0.621*	*0.633*	*6*	*58*	*0.603*	*0.734*	*3*	*47*	*0.532*	*0.503*
"Pacific crossing"	*As for "all" dataset*																												
Samoa	*As for "all" dataset*																												

Table I.5 (Continued) (b)

	GATA417				SW13				EV37				EV14				FCB1				SW19			
"Restricted" Dataset	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp	k	N	HObs	HExp
Total	7	423	0.504	0.529	14	507	0.826	0.838	32	487	0.856	0.906	14	469	0.682	0.716	16	502	0.795	0.837	32	491	0.882	0.922
Atlantic	3	39	0.385	0.399	10	39	0.795	0.844	16	37	0.811	0.889	11	38	0.684	0.728	10	36	0.694	0.826	17	35	0.886	0.929
"Atlantic crossing"																								
Bahamas																								
Central Atlantic	*2*	*2*	*0.5*	*0.5*	*3*	*2*	*1*	*0.833*	*4*	*2*	*1*	*1*	*2*	*1*	*1*	*1*	*3*	*2*	*0.5*	*0.833*	*4*	*2*	*1*	*1*
Canary Islands	*3*	*24*	*0.417*	*0.435*	*8*	*24*	*0.667*	*0.836*	*16*	*23*	*0.87*	*0.909*	*9*	*24*	*0.792*	*0.733*	*9*	*23*	*0.87*	*0.812*	*16*	*23*	*0.913*	*0.934*
Eastern Mediterranean																								
Central Mediterranean																								
Northern Mediterranean																								
Indian	5	148	0.561	0.576	12	163	0.81	0.844	24	158	0.835	0.908	12	158	0.696	0.73	14	163	0.742	0.824	27	162	0.864	0.922
Aldabras																								
Southwestern Australia																								
Cocos Island																								
"Indian crossing"																								
Maldives/Chagos Archipelago	*3*	*31*	*0.484*	*0.554*	*9*	*32*	*0.813*	*0.814*	*17*	*29*	*0.931*	*0.881*	*9*	*30*	*0.6*	*0.749*	*11*	*31*	*0.742*	*0.831*	*14*	*33*	*0.818*	*0.91*
Mauritius																								
Seychelles																								
Sri Lanka	*4*	*50*	*0.5*	*0.589*	*10*	*48*	*0.833*	*0.845*	*19*	*49*	*0.755*	*0.902*	*10*	*47*	*0.702*	*0.706*	*11*	*52*	*0.75*	*0.809*	*19*	*49*	*0.878*	*0.921*
Pacific	7	236	0.487	0.515	14	305	0.839	0.834	29	292	0.873	0.904	14	273	0.674	0.704	16	303	0.835	0.843	29	294	0.891	0.92
Gulf of California	*3*	*73*	*0.438*	*0.501*	*14*	*112*	*0.857*	*0.829*	*24*	*111*	*0.892*	*0.908*	*12*	*99*	*0.646*	*0.693*	*14*	*111*	*0.856*	*0.857*	*26*	*109*	*0.899*	*0.919*
Galapagos																								
Kiribati																								
New Zealand	*6*	*54*	*0.63*	*0.552*	*11*	*58*	*0.776*	*0.854*	*20*	*51*	*0.863*	*0.902*	*9*	*49*	*0.714*	*0.715*	*13*	*54*	*0.852*	*0.842*	*20*	*51*	*0.922*	*0.935*
Oregon																								
Papua New Guinea	*4*	*53*	*0.491*	*0.469*	*9*	*62*	*0.806*	*0.821*	*19*	*60*	*0.817*	*0.897*	*11*	*59*	*0.712*	*0.706*	*12*	*63*	*0.841*	*0.844*	*18*	*62*	*0.855*	*0.918*
"Pacific crossing"																								
Samoa																								

Table I.5 (Continued) (c)

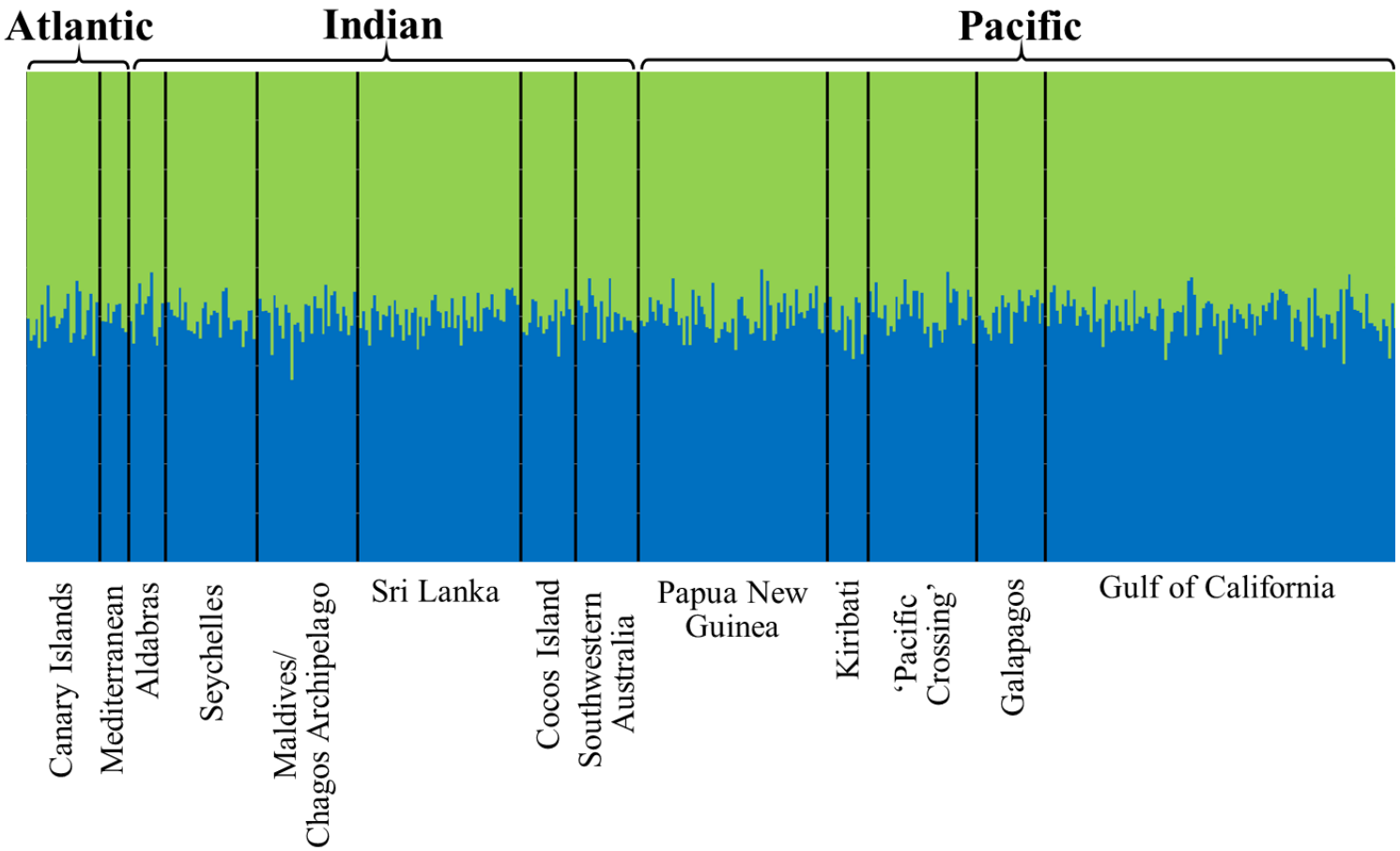
Comparison	Dataset	F_{ST}	p-value	G''_{ST}	p-value
AMOVA among regions	FULL	0.0048	1.07E-07	0.027	0
(worldwide)	RESTRICTED	0.0048	2.42E-06	0.027	0
AMOVA among oceans	FULL	0.0034	3.0E-06	0.016	0.001
	RESTRICTED	0.0034	6.2E-05	0.015	0.002
Pairwise PAC-IO	FULL	0.0026	0.0043	0.009	0.005
	RESTRICTED	0.0027	0.0094	0.009	0.008
Pairwise PAC-ATL	FULL	0.0043	0.0003	0.015	<0.001
	RESTRICTED	0.0040	<0.0001	0.014	0.003
Pairwise IO-ATL	FULL	0.0065	0.0001	0.023	0.000
	RESTRICTED	0.0063	0.0009	0.022	0.001
AMOVA among Pacific regions	FULL	0.0014	0.0117	0.012	0.091
	RESTRICTED	0.0014	0.0121	0.012	0.092
AMOVA among Indian regions	FULL	0.0051	0.0009	0.018	0.034
	RESTRICTED	0.0021	0.0393	0.004	0.326
AMOVA among Atlantic regions	FULL	0.0161	0.0180	0.052	0.042
	RESTRICTED	0.0144	0.0252	0.048	0.057

(d)	GCA	GPG	PX	KR	PNG	SWAUS	COC	SRI	MAL	SEY	ALD	MED	CNI
GCA	--	0.0000	0.0000	***0.0085****	***0.0009****	0.0088	0.0103	***0.0041****	0.0026	0.0000	0.0114	***0.0181****	0.0018
GPG	0.0000	--	0.0000	***0.0123****	***0.0018****	0.0078	***0.0164****	***0.0052****	0.0039	0.0000	0.011	***0.0166****	0.0088
PX	0.0000	0.0000	--	0.0039	***0.0044****	0.0018	0.0046	***0.0043****	0.0000	0.0000	0.0016	***0.0202*** **	0.001
KR	***0.0308****	***0.0515****	0.0135	--	***0.0049****	***0.0201****	0.0000	***0.0145*** **	0.0025	0.0062	0.0152	***0.0181****	0.0001
PNG	***0.0030****	***0.0046****	***0.0149****	***0.0169****	--	***0.0116****	***0.0101****	***0.0069*** **	***0.0054****	***0.0028****	***0.0156****	***0.0218*** **	***0.0027****
SWAUS	0.0282	0.0308	0.0038	***0.0691****	***0.0387****	--	0.0179	***0.0097****	0.0141	0.0078	0.0086	***0.0232****	***0.0177****
COC	0.0325	***0.0522****	0.0132	0.0000	***0.0317****	0.0539	--	***0.0064****	0.003	0.0000	0.0072	***0.0231****	0.0048
SRI	***0.0141****	***0.0216****	***0.0136****	***0.0523*** **	***0.0229*** **	***0.0270****	***0.0200****	--	***0.0086****	0.0000	0.0009	***0.0252*** **	***0.0066****
MAL	0.0093	0.0246	0.0000	0.0084	***0.0192****	0.0532	0.0095	***0.0314****	--	0.0000	0.0031	***0.0189****	0.001
SEY	0.0000	0.0000	0.0000	0.0223	***0.0084****	0.023	0.0000	0.0000	0.0000	--	0.0000	***0.0144****	0.0026
ALD	0.0416	0.0461	0.0064	0.0559	***0.0565****	0.0254	0.0239	0.0036	0.0134	0.0000	--	***0.0268****	0.0148
MED	***0.0617****	***0.0604****	***0.0696*** **	***0.0675****	***0.0731*** **	***0.0781****	***0.0763****	***0.0796*** **	***0.0656****	***0.0489****	***0.0860****	--	***0.0144****
CNI	0.0065	0.0307	0.0035	0.0003	***0.0096****	***0.0588****	0.0151	***0.0220****	0.0038	0.0091	0.0525	***0.0481****	--

Table I.5 (Continued) (e)

		394 bp mtDNA CR				Microsatellites						
		n	F_{ST} (females)	F_{ST} (males)	*p* - value	n	F_{ST} (females)	F_{ST} (males)	*p* -value	var. assignment (females)	var. assignment (males)	*p* - value
Pacific	By region	F: 172 M: 43	0.1113	0.0346	0.0697	F: 174 M: 44	0.0002	0.0091	0.8185	13.53	15.57	0.2832
Indian	By region	F: 67 M: 26	0.4705	0.2988	0.2152	F: 80 M:28	0.0061	0.0049	0.4408	8.801	6.384	0.7720
Worldwide	By region	F: 248 M: 73	0.2643	0.1288	***0.0194****	F: 268 M: 81	0.0066	0.0015	0.1730	11.73	12.45	0.2472
	By ocean	F: 268 M: 99	0.1205	0.0695	0.1028	F: 290 M: 113	0.0074	0.0000	***0.0150****	14.88	14.94	0.4339

Figure I.7: Structure plot (*K = 2*) based on 13 microsatellites, for the *Odyssey* samples where explicit spatial coordinates were available.



Appendix I References:

Bérubé M, Jørgensen H, McEwing R, Palsbøll PJ. 2000. Polymorphic di-nucleotide microsatellite loci isolated from the humpback whale, Megaptera novaeangliae. Molecular Ecology. 9: 2181-2183.

Buchanan FC, Friesen MK, Littlejohn RP, Clayton JW. 1996. Microsatellites from the beluga whale *Delphinapterus leucas*. Molecular Ecology. 5: 571-575.

Drouot V et al. 2004. A note on genetic isolation of Mediterranean sperm whales (*Physeter macrocephalus*) suggested by mitochondrial DNA. Journal of Cetacean Research and Management. 6: 29-32.

Engelhaupt D et al. 2009. Female philopatry in coastal basins and male dispersion across the North Atlantic in a highly mobile marine species, the sperm whale (*Physeter macrocephalus*). Mol Ecol. 18: 4193-4205.

Hamner RM. 2014. All in a DNA's work: conservation genetics and monitoring of the New Zealand endemic Maui's and Hector's dolphins. PhD dissertation. Oregon State University, Corvallis.

Lyrholm T, Gyllensten U. 1998. Global matrilineal population structure in sperm whales as indicated by mitochondrial DNA sequences. P Roy Soc B-Biol Sci. 265: 1679-1684.

Mesnick SL et al. 2011. Sperm whale population structure in the eastern and central North Pacific inferred by the use of single-nucleotide polymorphisms, microsatellites and mitochondrial DNA. Mol Ecol Resour. 11: 278-298.

Nei M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press.

Ortega-Ortiz JG, Engelhaupt D, Winsor M, Mate BR, Hoelzel AR. 2012. Kinship of long-term associates in the highly social sperm whale. Mol Ecol. 21: 732-744.

Palsbøll PJ, Bérubé M, Larsen AH, Jørgensen H. 1997. Primers for the amplification of tri-and tetramer microsatellite loci in baleen whales. Molecular Ecology. 6: 893-895.

Rendell L, Mesnick S, Dalebout M, Burtenshaw J, Whitehead H. 2012. Can genetic differences explain vocal dialect variation in sperm whales, *Physeter macrocephalus?* Behav Genet. 42: 332-343.

Richard KR, Dillon MC, Whitehead H, Wright JM. 1996a. Patterns of kinship in groups of free-living sperm whales (*Physeter macrocephalus*) revealed by multiple molecular genetic analyses. P Natl Acad Sci USA. 93: 8792-8795.

Richard KR, Whitehead H, Wright JM. 1996b. Polymorphic microsatellites from sperm whales and their use in the genetic identification of individuals from naturally sloughed pieces of skin. Molecular Ecology. 5: 313-315.

Schlötterer C, Amos B, Tautz D. 1991. Conservation of polymorphic simple sequence loci in cetacean species. Nature. 354: 63-65.

Valsecchi E, Amos W. 1996. Microsatellite markers for the study of cetacean populations. Molecular Ecology. 5: 151-156.

Waldick RC, Brown MW, White BN. 1999. Characterization and isolation of microsatellite loci from the endangered North Atlantic right whale. Molecular Ecology. 8: 1763.

Whitehead H, Dillon M, Dufault S, Weilgart L, Wright J. 1998. Non-geographically based population structure of South Pacific sperm whales: dialects, fluke-markings and genetics. J Anim Ecol. 67: 253-262.

Appendix II
Supplementary material associated with Chapter 3: Lenient matrilineality in groups of sperm whales from equatorial waters of the Pacific, Indian, and Atlantic Oceans

Figure II.1: R code for generating the null distribution of expected same-haplotype dyads based on background mtDNA haplotype frequencies, for use in calculating the standardized matrilineality index.

```
A <- rep.int(1, 4)
B <- rep.int(2, 36)
C <- rep.int(3, 15)
X <- rep.int(4, 156)
Y <- rep.int(5, 19)

x <- c(A, B, C, X, Y)

# x is an array of your regional haplotypes. These need to be coded as
#numbers

ndyads <- 174
# This is the number of within-group dyads you calculate directly from
#your group data

reps <- 1000
# This is the number of times you want to simulate drawing the
#number of within-group dyad pairs (ndyads) from your total regional
#haplotypes

result_table <- NULL
# This will be the table results are ultimately written to

j <- NULL

for (j in 1:reps) {
e <- 1
# e is used as the counter to get ndyad draws from the regional
#haplotype definitions
matrl_count <- 0
# This will be a count for the number of pairs over each ndyad draw
#that have the same haplotype

while (e <= ndyads) { # while e is less than ndyads: this just means
#that we are doing ndyad draws
g <- sample(x, 2, replace=TRUE, prob=NULL);
# Sampling 2 individuals, i.e. a dyad from our data
if(g[1]==g[2]) {matrl_count <- matrl_count + 1 }#If same haplotype,
#count goes up by one
e <- e+1
}
rep_result <- matrl_count/ndyads #Calculating prop. of same hap dyad
result_table <- rbind(result_table, rep_result) #Recording this
}
#Writing out a table with each 'dyad draw' same haplotype proportion
write.table(result_table, "null_matrilineality_index.txt", sep="\t")
```

Appendix III
A comparison of mass stranding and free-swimming sperm whale social groups

III.1. Abstract

Microsatellites and mtDNA from 106 free-swimming sperm whale social groups (*n* = 678 individuals, described in Chapter 3) were compared to genetic information from 10 mass strandings (*n* = 158 individuals), with field information on a further 67 strandings. I confirm that mass stranding events appear to represent the same entities as free-swimming social groups. My summary of stranding literature also suggests that mass stranding events in the sperm whale are too rare for the previously proposed mass stranding hypothesis to reduce mtDNA diversity.

III.2. Introduction

One similarity between the female social groups and the bachelor groups described in Chapter 3 is that they are both prone to mass stranding events, where apparently healthy individuals strand together. However, it is unknown whether mass stranding events accurately represent 'typical' social groups. The mechanisms leading to mass stranding events are unclear (Bradshaw et al. 2006), but the potential for antagonistic interactions to influence stranding behavior (Oremus et al. 2013) could mean that mass stranding events are not representative of single 'free-swimming' social groups, in terms of sex ratio, relatedness and other parameters of interest (Perrin and Geraci 2009). The comparison of molecular ecology of stranded groups to free-swimming groups is of interest to investigate the untested assumption that mass strandings are a good proxy for

free-swimming sperm whale groups in terms of group composition (Evans et al. 2002). In addition, the composition of mass strandings is of interest because low mtDNA diversity in the sperm whale could be driven by mass stranding events, depending on the frequency of those events (Siemann 1994).

To compare the genetic composition between free-swimming social groups and stranding events, I collate a record of all sperm whale mass stranding events available in the literature or recent news archives. I also genetically characterize several mass stranding events from the New Zealand cetacean tissue archive (Thompson et al. 2013) and where available, supplement these data with previous genetic research on stranding events. I then contrast free-swimming groups characterized in Chapter 3, and the stranding events summarized here, in terms of group size, sex ratio, the existence of first-order relatives, relatedness, and matrilineality.

III.3. Methods

III.3.1. Sample collection and genotyping

Skin and tissue samples (n = 80) were collected from New Zealand sperm whale strandings by New Zealand Department of Conservation staff from 1994 to 2008. DNA was extracted from these samples as detailed in Chapter 2. The samples were then genetically sexed, genotyped at 13 microsatellite loci, and had a 394 bp fragment of their mtDNA CR amplified, as detailed in Chapter 2. Where genetic information was available for mass stranding events in the literature, this was also incorporated in analyses.

III.3.2. Definition of mass stranding events

Mass strandings were identified in the literature/electronic news databases, and in the New Zealand stranding dataset (van Helden 2009), based on events where multiple sperm whales stranded simultaneously and where a total estimate of group size was available. Using field/genetic information, the sex ratio of stranding events was used to characterize the type of stranding event (e.g. bachelor = only males sampled, female social group/'mixed-sex group' = females sampled, unknown = no sex information available). Sex ratios were used in a comparison of free-swimming and mass stranding groups if they were based on sex information from more than half of the total estimated individuals in the stranding.

III.3.3. Average relatedness, kin, and the number of matrilines present in strandings

GROUPRELATE (Valsecchi et al. 2002) was used to calculate average relatedness between members of mass stranding events where more than five individuals were sampled. *GROUPRELATE* was also used to assess whether members of a stranding were significantly more related than expected by chance, by randomly permuting individuals among strandings within each geographic region (geographic region definitions follow Chapter 2, also see Figure 1.1). Potential first-order kin were identified as described in Chapter 2, by identifying pairs of individuals who shared at least one allele at every microsatellite locus through *SOLOMON v. 1.0-1* (Christie et al. 2013). To assess maternal relatedness, the minimum number of matrilines, represented by unique mtDNA CR haplotypes, was recorded for each mass stranding, and the standardized matrilineality

indices described in Chapter 3 calculated at the social group level for mass strandings in this dataset.

III.4. Results

III.4.1. Definition of mass stranding events and comparisons of group size and sex ratio

Among the literature/electronic news sources and the New Zealand stranding database, we identified 77 mass stranding events with estimated group sizes, with the earliest occurring in 1405 and the latest in 2012 (Table III.1). Of these events, 10 were characterized as bachelor group strandings, 18 as female-dominated and 49 as unknown. Sex-ratio information was available for 15 of the 18 'mixed-sex' (female social) groups. Of the 77 total strandings (Table III.1), 10 strandings had the number of mtDNA haplotypes/haplotype frequency information available and more than one individual genetically sampled (this study, Mesnick 2001, Evans et al. 2002, Engelhaupt 2004, Bearzi et al. 2011: summarized in Table III.2). These 10 strandings (n = 158 individuals) were included in downstream analyses of relatedness and matrilineality (Table III.2).

An explicit comparison of the estimated size of free-swimming social groups characterized in Chapter 3 versus stranding events (Figure III.1) revealed no significant differences based on the total dataset (two-tailed Welch's t-test, $p = 0.4436$), or when partitioned into bachelor ($p = 0.6201$) and female groups ($p = 0.2876$). As bachelor groups were defined as strandings containing no identified females, I only examined the percentage of females for the female dominated mass strandings. A comparison of the

percentage of females in free-swimming mixed-sex groups versus mass stranding events revealed no significant differences in the average percentage of females (two-tailed Welch's t-test, $p = 0.2792$, Figure III.2).

III.4.2. Identification of kin and relatedness

Of the five strandings (two female social groups, three bachelor groups) investigated for the presence of first-order kin, only one female-dominated event (PmaNZ035-054) had kin present: a female-female-female triad (potentially consistent with a grandmother-mother-daughter relationship), and an additional mother-daughter pair. The finding of kin within only this female stranding event is consistent with the free-swimming social group data (Chapter 3).

Given only five stranding groups had relatedness values available, no comparison of significance was carried out between free-swimming and mass stranding groups. One stranding event was comprised of individuals more related than expected by chance: a female-dominated New Zealand stranding event (PmaNZ006-009, $R = 0.3113$, $p = 0.0330$). This result is consistent with free-swimming social group data where only a small minority of groups were more related than expected by chance, restricted only to female-dominated ('mixed-sex') groups.

III.4.3. Number of matrilines present in strandings

Of the total 10 strandings with mtDNA information, there were 3 with only a single haplotype. Two of these strandings were female-dominated, with the remaining stranding

a bachelor group that stranded in the Mediterranean in 2009 (Bearzi et al. 2011, Mazzariol et al. 2011). The Mediterranean has previously been characterized as having only a single mtDNA haplotype (Chapter 2, Drouot et al. 2004, Engelhaupt et al. 2009) so this male stranding likely does not represent a true single matriline. The remainder of the strandings showed up to five mtDNA CR haplotypes, consistent with data from the free-swimming groups in Chapter 3. In a comparison of the number of haplotypes per group size between stranded and free-swimming groups, no significant differences were detected for the total dataset (two-tailed Welch's t-test, $p = 0.1829$, Figure III.3), or when partitioned into bachelor ($p = 0.3688$) and female groups ($p = 0.0600$). Standardized matrilineality index values showed a similar range to those of free-swimming groups (Table III.2), but sample sizes were too limited to test for significance

III.5. Discussion

In this study, we survey the molecular ecology of 10 strandings (with field information on a further 67 strandings) and compare these to 106 free-swimming sperm whale social groups (characterized in Chapter 3) in order to show that mass strandings appear to represent the same entities at free-swimming social groups. We found that mass strandings appear consistent with free-swimming groups in terms of size, sex ratio, relatedness, number of kin relationships identified and mtDNA diversity. In addition to consistency between the free-swimming and mass stranding data analyzed in this study, our estimates of the percentage of females within the female social groups (82.1 to 100% depending on ocean) were similar to previous whaling data (e.g. 77% female, Best 1979).

Our results confirm previous assumptions that mass stranding events accurately reflect free-swimming social groups (Evans et al. 2002), and reduces the likelihood that population genetic analyses based on mass stranding events will be biased, as found using single stranding events in other cetacean species (Bilgmann et al. 2011). Despite the potential for mass strandings to act as a proxy for free-swimming groups, I feel genetic studies on free-swimming individuals are still very valuable. Firstly, mass stranding events are relatively rare, and geographically biased towards the North Sea, Tasmania, the Mediterranean, and New Zealand (Table III.1). Secondly, the composition of groups that occur near-shore, and are therefore more likely to strand, might differ from groups found further off-shore e.g. bachelor groups in coastal waters consisting of fewer, but larger, individuals than more off-shore bachelor groups (Letteval et al. 2002). In addition, much like whaling, strandings only give information on a single location of an individual. In contrast, using the *Odyssey* samples from free-swimming whales, I was able to document movements of social groups over time-periods of up to 50 days, and spatial scales up to 500 km (Chapter 3).

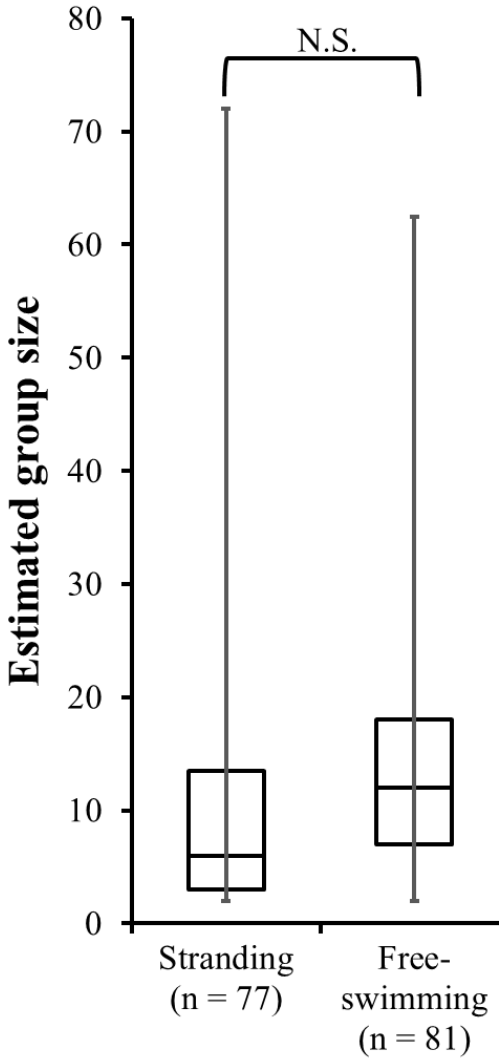


Figure III.1: An assessment of differences in sperm whale social group size between strandings and free-swimming groups. *n* represents the number of groups for each category. Minimum, lower quartile, median, upper quartile and maximum estimated group size is presented.

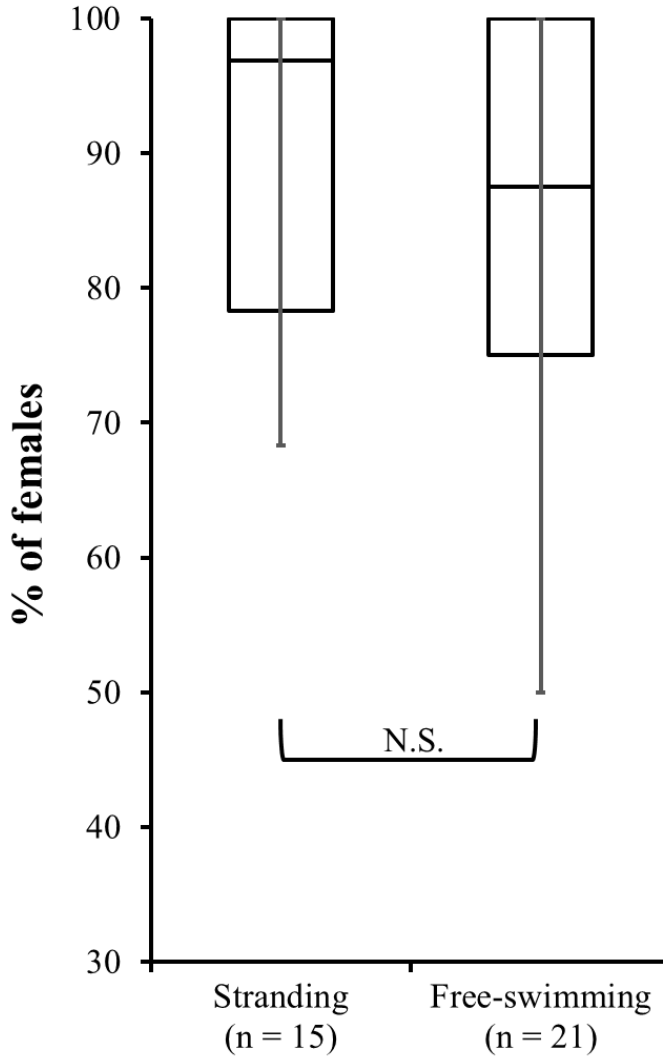


Figure III.2: An assessment of differences in sex ratio between 'mixed-sex' strandings and free-swimming groups. *n* represents the number of groups for each category. Minimum, lower quartile, median, upper quartile and maximum for % of females is presented.

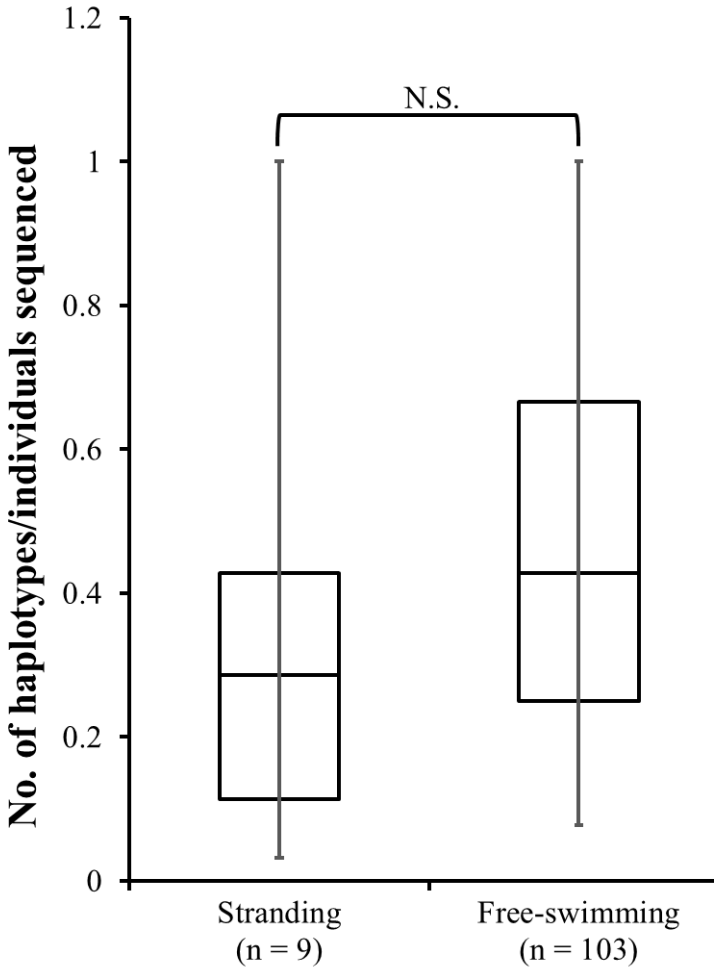


Figure III.3: An assessment of differences in the average number of haplotypes present per social group size (e.g. number of individuals sequenced within a social group) for sperm whale strandings and free-swimming groups. Minimum, lower quartile, median, upper quartile and maximum number of mtDNA haplotypes per group size (no. of individuals sequenced within in a social group) given. Lower values indicate fewer mtDNA haplotypes present within a group.

Table III.1: Summary of sperm whale mass stranding events in the literature/internet where an estimate of the number of stranded individuals was available. These are likely an underestimate, as they are based on written records where the number of stranded individuals was recorded. Type refers to Bachelor (male-only), Female ('mixed-sex' or female-dominated social group) or Unknown (no sex information given). Based on location, it is likely that 'Unknown' North Sea strandings actually reflect bachelor groups. Records shaded in grey were included in genetic analyses of stranded groups. References: [1] (de Smet 1997); [2] (Pierce et al. 2007); [3] (Bearzi et al. 2011); [4] (Sousa and Brito 2011); [5] (Guiler 1978); [6] (Scott and Green 1975); [7] (Evans et al. 2002); [8] (Robson and van Bree 1971); [9] (Stephenson 1975); [10] (Ramos et al. 2001); [11] (McManus et al. 1984); [12] (Harvey et al. 2014); [13] (Rice et al. 1986); [14] (Jauniaux et al. 1998); [15] (Engelhaupt 2004); [16] (Evans 1997); [17] (Mendes et al. 2007); [18] (Santos et al. 2002); [19] (Mesnick 2001); [20] (Mazzariol et al. 2011); [21] (Shears 2009); [22] (Anonymous 2011); [23] (Anonymous 2012); [24] This study.

Year	Location	Ocean	Group type	*n*			Reference
				F	M	Total	
1402	Belgium, North Sea	Atlantic	Unknown			8	[1]
1577	The Netherlands, North Sea	Atlantic	Unknown			13-14	[2]
1577	Belgium/Netherlands, North Sea	Atlantic	Unknown			3	[1, 2]
1584	Italy, Mediterranean	Atlantic	Unknown			7-8	[3]
1604	Germany, North Sea	Atlantic	Unknown			2	[2]
1617	The Netherlands, North Sea	Atlantic	Unknown			2	[2]
1617	The Netherlands, North Sea	Atlantic	Unknown			>1	[2]
1723	Germany, North Sea	Atlantic	Unknown			~21	[2]
1738	Germany, North Sea	Atlantic	Unknown			3	[2]
1751	Germany, North Sea	Atlantic	Unknown			2	[2]
1753	Great Britain, North Sea	Atlantic	Unknown			3	[2]
1757	Denmark, North Sea	Atlantic	Unknown			3	[2]
1761	Great Britain, North Sea	Atlantic	Unknown			2	[2]
1762	The Netherlands, North Sea	Atlantic	Unknown			7-8	[2]

Table III.1 (Continued)

Year	Location	Ocean	Group type	*n*			Reference
				F	M	Total	
1762	Belgium, North Sea	Atlantic	Unknown			2	[2]
1762	Germany, North Sea	Atlantic	Unknown			2	[2]
1762	Great Britain, North Sea	Atlantic	Unknown			12	[2]
1765	Denmark, North Sea	Atlantic	Unknown			2	[2]
1770	Denmark, North Sea	Atlantic	Unknown			2	[2]
1784	Portugal	Atlantic	Unknown			10	[4]
1802,1803 or 1805	Italy, Mediterranean	Atlantic	Unknown	?	1	3	[3]
1853	Croatia, Mediterranean	Atlantic	Unknown			6	[3]
1911	Tasmania, Tasman Sea	Pacific	Bachelor		38	38	[5, 6]
1937	The Netherlands, North Sea	Atlantic	Unknown			2	[2]
1937	France/Belgium, North Sea	Atlantic	Unknown			2	[1, 2]
1938	Italy, Mediterranean	Atlantic	Unknown			7	[3]
1949	Denmark, North Sea	Atlantic	Unknown			5	[2]
1953	Tasmania, Tasman Sea	Pacific	Bachelor		2	2	[7]
1956	Albania, Mediterranean	Atlantic	Female	5	?	8	[3]
1958	New Zealand, Tasman Sea	Pacific	Female	13		13	[8]
1968	Mervey Beach, Australia	Unknown	Unknown			10	[8]
1970	Tasmania, Tasman Sea	Pacific	Female	48	10	58	[5]
1971	Tasmania, Tasman Sea	Pacific	Unknown			32	[5]
1971	New Zealand	Pacific	Female	46	13	59	[5, 8]
1972	New Zealand	Pacific	Female			30	[9]
1972	New Zealand, Tasman Sea	Pacific	Unknown			14	[9]
1972	Brazil	Atlantic	Unknown			33	[10]
1974	New Zealand, Tasman Sea	Pacific	Female	54	18	72	[9]

Table III.1 (Continued)

Year	Location	Ocean	Group type	*n*			Reference
				F	M	Total	
1975	Tasmania, Tasman Sea	Pacific	Female	2		2	[5]
1975	Tasmania, Tasman Sea	Pacific	Unknown			3	[11]
1979	Oregon	Pacific	Female	28	13	41	[12,13]
1980	Tasmania, Tasman Sea	Pacific	Unknown			3	[11]
1981	Tasmania, Tasman Sea	Pacific	Female	7	3	26	[7, 11]
1981	Tasmania, Tasman Sea	Pacific	Bachelor		2	2	[11]
1982	Tasmania, Tasman Sea	Pacific	Female	11	3	14	[11]
1982	Tasmania, Tasman Sea	Pacific	Female	5	?	9	[11]
1984	Denmark, North Sea	Atlantic	Unknown			2	[2]
1989	Tasmania, Tasman Sea	Pacific	Unknown			5	[7]
1989	Tasmania, Tasman Sea	Pacific	Female	16		16	[7]
1991	Denmark, North Sea	Atlantic	Unknown			3	[2]
1994	Belgium, North Sea	Atlantic	Unknown			3	[2]
1994	Belgium, North Sea	Atlantic	Bachelor		4	4	[1, 2, 14]
1994	Orkney, North Sea	Atlantic	Bachelor		11	11	[14, 15, 16, 17]
1994	New Zealand, Tasman Sea	Pacific	Unknown			3	[24]
1995	The Netherlands, North Sea	Atlantic	Bachelor		3	3	[2, 14]
1996	Great Britain, North Sea	Atlantic	Bachelor		6	6	[2, 15, 17]
1996	New Zealand	Pacific	Unknown			3	[24]
1996	Denmark, North Sea	Atlantic	Unknown			16	[2]
1997	The Netherlands, North Sea	Atlantic	Bachelor		4	4	[2, 18]
1997	Great Britain, North Sea	Atlantic	Unknown			2	[2]
1997	Germany, North Sea	Atlantic	Unknown			2	[2]

Table III.1 (Continued)

Year	Location	Ocean	Group type	*n*			Reference
				F	M	Total	
1997	Denmark, North Sea	Atlantic	Unknown			13	[2]
1998	Germany, North Sea	Atlantic	Unknown			6	[2]
2002	Germany, North Sea	Atlantic	Unknown			3	[2]
1998	Tasmania, Tasman Sea	Pacific	Female	62	2	66	[7, 19]
1998	Tasmania, Tasman Sea	Pacific	Female	10	1	11	[7, 19]
1998	Tasmania, Tasman Sea	Pacific	Female	25	10	35	[7, 19]
2000	New Zealand	Pacific	Female	15		18	[24]
2003	New Zealand, Tasman Sea	Pacific	Bachelor		12	12	[24]
2003	Germany, North Sea	Atlantic	Unknown			2	[2]
2004	The Netherlands, North Sea	Atlantic	Unknown			2	[2]
2006	Great Britain, North Sea	Atlantic	Unknown			3	[2]
2009	Italy, Mediterranean	Atlantic	Bachelor		7	7	[3, 20]
2009	Tasmania, Tasman Sea	Pacific	Unknown			50	[21]
2011	Tasmania, Tasman Sea	Pacific	Unknown			22-24	[22]
2012	China	Pacific	Female	1	?	4	[23]
?	Gulf of California	Pacific	Female			7	[19]

Table III.2: Strandings used in genetic analyses, separated by headers indicating region, ocean and year sampled (for some previously published groups, date unknown). Number of individuals that were included in nuclear analyses (where applicable) indicated by *n* (QC). Number of females/males in each group was determined through field and genetic data. Relatedness (*R)* was calculated using *GROUPRELATE*. No *R*-values were significant. A 'Y' in the Kin column indicates first-order relationships were detected in that group. Number of individuals sequenced for mtDNA shown in *n*(mtDNA) column and number of mtDNA CR haplotypes per stranding in *k* column. Standardized matrilineality index (SMI) is presented by group and region. Standardized matrilineality index values bolded, italicized and marked with asterisk if significant at * p <0.05; ** p < 0.001. References: [1] This study; [2] (Bearzi et al. 2011); [3] (Mazzariol et al. 2011); [4] (Engelhaupt 2004); [5] (Mesnick 2001); [6] (Evans et al. 2002).

Group type	Group name	Estimated size	*n* (QC)	*n* (F)	*n* (M)	*R*	Kin	*n*(mtDNA)	*k*	SMI (Hi_{obs} - Hi_{exp})	Reference
Mediterranean (2000)											
♂		7	7	0	7	--	--			--	[2, 3]
North Sea (1994, 1996)								17	3	0.081	
♂	NS2	6	6	0	6	-0.006		6	2	***0.261****	[4]
♂	NS1	11	11	0	11	0.069		3	3	0.033	[4]
Gulf of California (Unknown)											
♀	MES_GCA_1	--	7	--	--	--	--	7	3	--	[5]
New Zealand (1994, 2000, 2003)										***0.126****	
♀	PmaNZ035-054	18	12	15	0	0.042	Y	8	2	***0.278****	[1]
♂	PmaNZ061-072	12	12	0	12	-0.018		10	5	-0.009	[1]
♀	PmaNZ006-008	3	2	3	0	0.311		2	1	0.836	[1]
Tasmania (1998)											
♀	MES_TAS_2	11	10	10	1	--	--	10	1	--	[5, 6]
♀	MES_TAS_1	66	63	62	6	--	--	63	2	--	[5, 6]
♀	MES_TAS_3	35	35	25	10	--	--	35	4	--	[5, 6]

Appendix III References:

Anonymous. 2011. Last whale dies after mass stranding. ABC news. Available from: http://www.abc.net.au/news/2011-11-16/last-whale-dies-after-mass-stranding/3675886.

Anonymous. 2012. Sperm whales die in China. Avax News. Available from http://avaxnews.net/sad/Sperm_Whales_Die_In_China.html.

Bearzi G et al. 2011. Overview of sperm whale *Physeter macrocephalus* mortality events in the Adriatic Sea, 1555–2009. Mammal Review. 41: 276-293.

Best PB. 1979. Chapter 7: Social organization in sperm whales, *Physeter macrocephalus*. In: Winn, HE, Olla, BL, editor(s). *Behavior of Marine Animals. Volume 3: Cetaceans*. New York-London: Plenum Press.

Bilgmann K, Möller LM, Harcourt R, Kemper CM, Beheregaray LB. 2011. The use of carcasses for the analysis of cetacean population genetic structure: a comparative study in two dolphin species. PLoS ONE. 6: e20103.

Bradshaw CJ, Evans K, Hindell MA. 2006. Mass cetacean strandings-a plea for empiricism. Conservation Biology. 20: 584-586.

Christie MR, Tennessen JA, Blouin MS. 2013. Bayesian parentage analysis with systematic accountability of genotyping error, missing data and false matching. Bioinformatics. 29: 725-732.

de Smet WMA. 1997. Five centuries of sperm whale strandings along the Flemish Coast. Biologie. 67: 11-14.

Drouot V et al. 2004. A note on genetic isolation of Mediterranean sperm whales (*Physeter macrocephalus*) suggested by mitochondrial DNA. Journal of Cetacean Research and Management. 6: 29-32.

Engelhaupt D et al. 2009. Female philopatry in coastal basins and male dispersion across the North Atlantic in a highly mobile marine species, the sperm whale (*Physeter macrocephalus*). Mol Ecol. 18: 4193-4205.

Engelhaupt DT. 2004. Phylogeography, kinship and molecular ecology of sperm whales (*Physeter macrocephalus*). University of Durham, Durham City.

Evans K, Morrice M, Hindell M, Thiele D. 2002. Three mass strandings of sperm whales (*Physeter macrocephalus*) in southern Australian waters. Marine Mammal Science. 18: 622-643.

Evans PGH. 1997. Ecology of sperm whales (*Physeter macrocephalus*) in the Eastern North Atlantic, with special reference to sightings and strandings records from the British Isles. Biologie. 67: 37-46.

Guiler ER. 1978. Whale strandings in Tasmania since 1945 with notes on some seal reports. Papers and Proceedings of the Royal Society of Tasmania. 112: 189-213.

Harvey JT, Friend T, McHuron EA. 2014. Cephalopod remains from stomachs of sperm whales (*Physeter macrocephalus*) that mass-stranded along the Oregon coast. Marine Mammal Science. 30: 609-625.

Jauniaux T et al. 1998. Postmortem investigations on winter stranded sperm whales from the coasts of Belgium and The Netherlands. Journal of Wildlife Diseases. 34: 99-109.

Letteval E et al. 2002. Social structure and residency in aggregations of male sperm whales. Canadian Journal of Zoology. 80: 1189-1196.

Mazzariol S et al. 2011. Sometimes sperm whales (*Physeter macrocephalus*) cannot find their way back to the high seas: a multidisciplinary study on a mass stranding. PLoS ONE. 6: e19417.

McManus TJ, Wapstra JE, Guiler ER, Munday BL, Obendorf DL. 1984. Cetacean strandings in Tasmania from February 1978 to May 1983. Papers and Proceedings of the Royal Society of Tasmania. 118: 117-135.

Mendes S, Newton J, Reid RJ, Zuur AF, Pierce GJ. 2007. Stable carbon and nitrogen isotope ratio profiling of sperm whale teeth reveals ontogenetic movements and trophic ecology. Oecologia. 151: 605-615.

Mesnick SL. 2001. Genetic relatedness in sperm whales: evidence and cultural implications. Behavioral and Brain Sciences. 24: 346-347.

Oremus M, Gales R, Kettles H, Baker CS. 2013. Genetic evidence of multiple matrilines and spatial disruption of kinship bonds in mass strandings of long-finned pilot whales, *Globicephala melas*. Journal of Heredity. 104: 301-311.

Perrin WF, Geraci JR. 2009. Stranding. In: Perrin, WF, Wursig, B, Thewissen, JGM, editor(s). *Encyclopedia of Marine Mammals*. San Diego: Academic Press.

Pierce GJ, Santos MB, Smeenk C, Saveliev A, Zuur AF. 2007. Historical trends in the incidence of strandings of sperm whales (*Physeter macrocephalus*) on North Sea coasts: An association with positive temperature anomalies. Fisheries Research. 87: 219-228.

Ramos RMA et al. 2001. A note on strandings and age of sperm whales (*Physeter macrocephalus*) on the Brazilian coast. Journal of Cetacean Research and Management. 3: 321-327.

Rice DW, Wolman AA, Mate BR, Harvey JT. 1986. A mass stranding of sperm whales in Oregon: sex and age composition of the school. Marine Mammal Science. 2: 64-69.

Robson FD, van Bree PJH. 1971. Some remarks on a mass stranding of sperm whales, *Physeter macrocephalus* Linnaeus, 1758, near Gisborne, New Zealand. Zeitschrift für Säugetierkunde. 36: 55-60.

Santos MB et al. 2002. Additional notes on stomach contents of sperm whales *Physeter macrocephalus* stranded in the north-east Atlantic. J. Mar. Biol. Ass. UK. 82: 501-507.

Scott EOG, Green RH. 1975. Recent whale strandings in northern Tasmania. Papers and Proceedings of the Royal Society of Tasmania. 109: 91-96.

Shears R. 2009. Tragedy of 48 whales who died after mass stranding on coast of Tasmania. The Daily Mail. Available at: http://www.dailymail.co.uk/news/article-1126866/Tragedy-48-whales-died-mass-stranding-coast-Tasmania.html.

Siemann LA. 1994. Mitochondrial DNA sequence variation in North Atlantic long-finned pilot whales, *Globicephala melas*. Doctor of Philosophy. Massachusetts Institute of Technology/Woods Hole Oceanographic Institution, Massachusetts, U.S.A.

Sousa A, Brito C. 2011. Historical strandings of cetaceans on the Portuguese coast: anecdotes, people and naturalists. Marine Biodiversity Records. 4: 1-8.

Stephenson AB. 1975. Sperm whales stranded at Muriwai Beach, New Zealand. New Zealand Journal of Marine and Freshwater Research. 9: 299-304.

Thompson KF et al. 2013. A novel conservation approach provides insights into the management of rare cetaceans. Biol Conserv. 331-340.

Valsecchi E, Hale P, Corkeron P, Amos W. 2002. Social structure in migrating humpback whales (*Megaptera novaeangliae*). Molecular Ecology. 11: 507-518.

van Helden A. 2009. New Zealand *Physeter macrocephalus* stranding record. Administered by Te Papa Tongarewa for the Department of Conservation (DOC).

Appendix IV
Supplementary material associated with Chapter 4: Low diversity in the mitogenome of sperm whales revealed by next-generation sequencing

Table IV.1: The 61 cetacean mitogenome sequences, representing 42 species, used in this study. Several species were represented by multiple individuals: the sperm whale (*Physeter macrocephalus*), the bottlenose dolphin (*Tursiops truncatus)*, and the Hector's dolphin (*Cephalorhynchus hectori).* Species/subspecies not previously available on Genbank as reference sequence mitogenomes are shaded in grey and the 44 mitogenomes used in the Bayesian phylogeny are asterisked.

Species	Accession #	Reference
Physeter macrocephalus: PmaOR001	KC312619	This study
Physeter macrocephalus: PmaNZ005*	KC312603	This study
Physeter macrocephalus: PmaNZ008	KC312604	This study
Physeter macrocephalus: PmaNZ010	KC312605	This study
Physeter macrocephalus: PmaNZ013	KC312606	This study
Physeter macrocephalus: PmaNZ015	KC312607	This study
Physeter macrocephalus: PmaNZ016	KC312608	This study
Physeter macrocephalus: PmaNZ032	KC312609	This study
Physeter macrocephalus: PmaNZ034	KC312610	This study
Physeter macrocephalus: PmaNZ036	KC312611	This study
Physeter macrocephalus: PmaNZ039	KC312612	This study
Physeter macrocephalus: PmaNZ045	KC312613	This study
Physeter macrocephalus: PmaNZ056	KC312614	This study
Physeter macrocephalus: PmaNZ058	KC312615	This study
Physeter macrocephalus: PmaNZ076	KC312616	This study
Physeter macrocephalus: PmaNZ082	KC312617	This study
Physeter macrocephalus: PmaNZ085	KC312618	This study
*Cephalorhynchus commersonii**	KC312623	Carraher (2004) – first published with this study
*Cephalorhynchus eutropia**	KC312624	Carraher (2004) – first published with this study
*Cephalorhynchus heavisidii**	KC312625	Carraher (2004) – first published with this study
*Cephalorhynchus hectori hectori**	KC312626	Carraher (2004) – first published with this study
*Cephalorhynchus hectori maui**	KC312627	Carraher (2004) – first published with this study
*Delphinus delphis**	KC312628	Carraher (2004) – first published with this study
*Globicephala melas**	KC312629	Carraher (2004) – first published with this study
*Lagenorhynchus australis**	KC312630	Carraher (2004) – first published with this study
*Lagenorhynchus cruciger**	KC312620	Carraher (2004) – first published with this study
*Lagenorhynchus obscurus**	KC312621	Carraher (2004) – first published with this study
*Tursiops truncatus**	KC312622	Carraher (2004) – first published with this study
*Balaena mysticetus**	NC_005268	Arnason et al. (2004)
*Balaenoptera acutorostrata**	NC_005271	Arnason et al. (2004)
*Balaenoptera bonaerensis**	NC_006926	Sasaki et al. (2005)

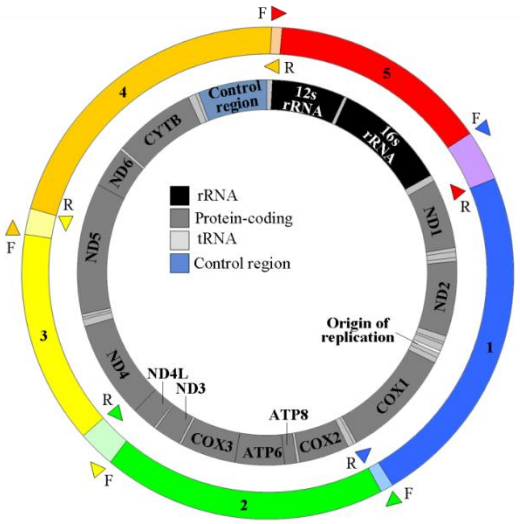
Table IV.1 (Continued)

Species	Accession #	Reference
*Balaenoptera borealis**	NC_006929	Sasaki et al. (2005)
*Balaenoptera brydei**	NC_006928	Sasaki et al. (2005)
*Balaenoptera edeni**	NC_007938	Sasaki et al. (2006)
*Balaenoptera musculus**	NC_001601	Arnason and Gullberg (1993)
*Balaenoptera omurai**	NC_007937	Sasaki et al. (2006)
*Balaenoptera physalus**	NC_001321	Arnason et al. (1991)
*Berardius bairdii**	NC_005274	Arnason et al. (2004)
*Caperea marginata**	NC_005269	Arnason et al. (2004)
*Delphinus capensis**	NC_012061	Xiong et al. (2009)
*Eschrichtius robustus**	NC_005270	Arnason et al. (2004)
*Eubalaena australis**	NC_006930	Sasaki et al. (2005)
*Eubalaena japonica**	NC_006931	Sasaki et al. (2005)
*Grampus griseus**	NC_012062	Xiong et al. (2009)
*Hyperoodon ampullatus**	NC_005273	Arnason et al. (2004)
*Inia geoffrensis**	NC_005276	Arnason et al. (2004)
*Kogia breviceps**	NC_005272	Arnason et al. (2004)
*Lagenorhynchus albirostris**	NC_005278	Arnason et al. (2004)
*Lipotes vexillifer**	NC_007629	Yan et al. (2005)
*Megaptera novaeangliae**	NC_006927	Sasaki et al. (2005)
*Monodon monoceros**	NC_005279	Arnason et al. (2004)
*Orcinus orca**	NC_014682	Foote et al. (2011b)
*Phocoena phocoena**	NC_005280	Arnason et al. (2004)
Physeter macrocephalus	NC_002503	Arnason et al. (2000)
*Platanista minor**	NC_005275	Arnason et al. (2004)
*Pontoporia blainvillei**	NC_005277	Arnason et al. (2004)
*Sousa chinensis**	NC_012057	Xiong et al. (2009)
*Stenella attenuata**	NC_012051	Xiong et al. (2009)
*Stenella coeruleoalba**	NC_012053	Xiong et al. (2009)
*Tursiops aduncus**	NC_012058	Xiong et al. (2009)
*Tursiops truncatus**	NC_012059	Xiong et al. (2009)

Table IV.2: Sample code, location, 400 bp mtDNA CR haplotype (generated using Sanger sequencing: see Table 2 for haplotype definitions), and NGS platform used to sequence the mitogenome of the 17 sperm whales from this study. PmaNZ036 and PmaNZ039 were sampled from the same mass stranding event. All other samples come from separate stranding events. The three samples sequenced on both NGS platforms are highlighted in gray.

Sample name	Location	Haplotype	NGS platform	
			Illumina GAIIx	Roche 454 GS Junior
PmaOR001	Oregon	B		X
PmaNZ005	New Zealand	B	X	X
PmaNZ008	New Zealand	EE	X	
PmaNZ010	New Zealand	C	X	
PmaNZ013	New Zealand	J		X
PmaNZ015	New Zealand	B	X	
PmaNZ016	New Zealand	J	X	
PmaNZ032	New Zealand	B	X	
PmaNZ034	New Zealand	A	X	X
PmaNZ036	New Zealand	B	X	
PmaNZ039	New Zealand	O	X	
PmaNZ045	New Zealand	B	X	X
PmaNZ056	New Zealand	S	X	
PmaNZ058	New Zealand	C	X	
PmaNZ076	New Zealand	Q		X
PmaNZ082	New Zealand	II		X
PmaNZ085	New Zealand	N	X	

Figure IV.1: Location of the primers used to generate the five overlapping LR-PCR fragments for amplification of sperm whale mitogenomes. Primer sequence, length of fragment, and reaction conditions are detailed in the table. Primers were designed for this study with the exception of 1.4UPF (Allan Wilson Center, Massey University, New Zealand), Mys10000ND4LF, Mys130000ND5R (designed by M. Vant) and Pma6916tSerR, PmaHS13660F (modified from Allan Wilson Center primers).



Fragment		Primer name and sequence	Length of fragment (bp)	Annealing temperature (°C)	dNTP concentration (mM)	Polymerase amount (units)
1	Forward	1.4UPF: 5'-AAT CCA GGT CGG TTT CTA TCT-3'	4,391	60	0.4	0.2
	Reverse	Pma6916tSerR: 5'-GTT CGA KTC CTT CCT TTC TT-3'				
2	Forward	Pma6800CO1F: 5'-GAG AAG CMT TYR CAT CCA AAC G-3'	3,608	61	0.2	0.1
	Reverse	PmaHDND4R: 5'-GGG TCA GAG AAG AAC GTA AGG G-3'				
3	Forward	Mys10000ND4LF: 5'-CGA TCC CAC CTA ATR TCC GCA-3'	3,024	60	0.4	0.2
	Reverse	Mys13000ND5R: 5'-GCT CAG GCG TTG GTA TAA GA-3'				
4	Forward	PmaHS13660F: 5'-GCC TYA ACC AAC CYT AYC TRG-3'	3,874	61	0.4	0.2
	Reverse	Pma12sRNAR: 5'-GTG CTT GAT ACC WGC TCC TTT T-3'				
5	Forward	PmaM13tpheF: 5'-AAA GCA ADA CAC TGA AGA TGN NT-3'	3,091	64	0.4	0.1
	Reverse	PmaHD3106R: 5'-TAR ACR GYY AGG CTG GAT ATT GC-3'				

Table IV.3: Primers used to amplify and sequence the 11 dolphin mitogenomes generated in this study. Regions in column 1 relate to labeled regions in NC_005278 (Arnason et al. 2004). The columns to the right indicate those primers that were modified for use on mysticetes. Reference 1: Allan Wilson Centre, Massey University, New Zealand. Reference 2: This study. Reference 3: Cooper *et al*. (2001).

					Mysticete modified			
Region	**Amplifying Primers**	**Sequencing Primers**	**Primer sequence 5`-3`**	**Ref**	**Amplifying Primers**	**Sequencing Primers**	**Primer sequence 5`-3`**	**Ref**
1	1.4UPF - tMet5201R 1360 bp	1.4UPF	AATCCAGGTCGGTTTCTATCT	1				
		HD3106R	GACRGCTAGGCTTGATATTGC	2				
		ND14747F	CCATTCGCCCTATTCTTCCTAGC	3				
		Av4903R	GTTGGTCATATCGGAATCGTG	1				
		BatL3564F	CTTAYCCACGATTCCGGTATGA	1				
		tMet5201R	CCATCATTTTCGGGGTATGG	1				
2	BatL5364F-1LP5100R 1404 bp	BatL3564F	CTTAYCCACGATTCCGGTATGA	1				
		HS4823tRNA-MetF	CCCATACCCCGGAAATGTTG	1				
		BatL4235F	TTTCACTTTTGARTACCAGAAGTT	1				
		BatH4461R	TGGGCRATTGATGAGTATGC	1				
		1LP5100R	AGGCTTTTGAAGGCCTTTGGTCT	1	Mys5200AsnR	Mys5200AsnR	GGAGAAGTAGATTGAAGCCAGTTG	2
3	BatL5310F - Bat6871tSerR 1567 bp	BatL5310F	CCTACTCRGCCATTTTACCTATG	1				
		HS5300F	CCTTCTACCAACAGCAATCGTAC	1	Mys5900CO1F	Mys5900CO1F	AAAACCACCYGCCATRACC	2
		HS6200R	TATTGGGTTATAGCGGGTGGTTT	1				
		BatH7197R	TCTACTTCTTGNGCRTCTAT	1				
		L7318F	ACATTCTTTGAYCCWGCRGGAGG	3				
		H7662R	AGGAAGATGAAGCCYAGAGCTCA	3				
		Bat6871tSerR	GTTCGATTCCTTCCTTTCTT	1				
		Dde7500R (common dolphin only)	CGGATTGTCATTTGTATGGGT	2				
		Dde6900F (common dolphin only)	CCTTGAATGACTAAACGGGTG	2				
4	L7318F - HS8790ATP8R 1893 bp	HS7000F	CATCTCATCAATAAGGCTC	1	Mys6800CO1F	Mys6800CO1F	GAGAAGCATTYACATCCAAACG	2
		BatH7197R	TCTACTTCTTGNGCRTCTAT	1	Mys7800APT8F	Mys7800APT8F	AACATGACTCCTYACYATCC	2
		UP7371F	GGYCATCAATGATAYTGAAGCTA	1				
		2BLP7562R	CCTAGKGATGGSACDGYTCATG	1	Mys7600CO2R	Mys7600CO2R	GGTTGTTTGGTTTAGRCG	2
		HS8790ATP8R	TTGGTTTGGGGGAGGGAGAG	1	Mys8600CO3R	Mys8600CO3R	GTGRTATGAGTGRGTTTGGTGGGTC	2
5	UP7371F - HSealH9849R 1591 bp	UP7371F	GGYCATCAATGATAYTGAAGCTA	1				
		HD7815F	ACAACTAGATACATCAACATGG	2				
		HS8790ATP8R	TTGGTTTGGGGGAGGGAGAG	1				
		HD8351F	CATTAGGCACACCCACTTTC	2				
		HSeal9327R	ACTGCTARKGCTAKGGGTTGA	1				
		HD8386R	GGAATTAGGAAAGTGGGTGTG	2				
		3.4UF	TTYGTRGCMACAGGHTTTCA	1				
		HSealH9849R	GGGGCTAGGCTKGAGTGGTAAAAGGCTCA	1				

Table IV.3 (Continued)

Region	Amplifying Primers	Sequencing Primers	Primer sequence 5`-3`	Ref	Mysticete modified Amplifying Primers	Sequencing Primers	Primer sequence 5`-3`	Ref
6	HD8351F-HD9912R 1552 bp	HD8351F	CATTAGGCACACCCACTTTC	2	Mys8400ATP6F	Mys8400ATP6F	CTAACYGCCAACATYACAGC	2
		HSealH9849R	GGGGCTAGGCTKGAGTGGTAAAAGGCTCA	1	Mys9500R	Mys9500R	CCGAAACTAACTGATTGGAAG	2
		HSealL9746F	ATACCARTGATGACGMGATATTRTCCGAGAA	1				
		HD9912R	GGAYTAGAGACAYTTGGTAYTTG	2	Mys10000ND4LR	Mys10000ND4LR	TGCGGAYATTAGGTGGGATCG	2
7	HSealL9746F – HDND4R 1099 bp	HSealL9746F	ATACCARTGATGACGMGATATTRTCCGAGAA	1				
		HD9912R	GGAYTAGAGACAYTTGGTAYTTG	2				
		HDND4R	GGGTCAGAGAAGAATGTTAAAG	2				
8	HSealL10709F - HD11690R 1863 bp	HSealL10709F	CTTAGCCTACGAATGAACCCA	1	Mys9900ArgF	Mys9900ArgF	TTTCGACYCACTARACTGTG	2
		HD9912R	GGAYTAGAGACAYTTGGTAYTTG	2				
		HDND4R	GGGTCAGAGAAGAATGTTAAAG	2				
		HSealL11521F	GGCAACCAAACCGAAACGACTTAA	1	Mys10400ND4F	Mys10400ND4F	YTRATAGCAAGCCAATCC	2
		HSeal712654R	CCTAAGACCAATGGATTACTTCTATCCT	1	Mys11100ND4R	Mys11100ND4R	CAATTATTAGAGCGGTGG	2
		Bat10878ND4R	TAGTATGCCGTAGCCTCCTAGTT	1				
		HD11690R	GCATGAATTAGCAGTTCTTG	2				
9	HSeal12431ND4F - HS14906NADH6R 2478 bp	HSealL12431ND4F	ACATTATCCCCCTCCTACTCCT	1	Mys11700LeuF	Mys11700LeuF	CCATTGGTCTTAGGARCC	2
		F4UF	ACYGAYGCMAATACAGCAGCC	1				
		H13734R	AGGCCAAATTGRGCTGATTTTCC	1	Mys13100ND5R	Mys13100ND5R	GAAGCGGGGTTGTCCTAGTA	2
		HS13660F	GCMTYAACCAACCACACCTR	1				
		HD13211R	ATTAGGGGGATGTTTGTTGG	2				
		HS14906R	GGGTTGGTTGTGGGTGGTGCT	1	Mys14000ND6R	Mys14000ND6R	TTGATTGTGGGTGGTGGTG	2
10	HS13660F - 14386LR 1643 bp	HS13660F	GCMTYAACCAACCACACCTR	1	Mys13000ND5F	Mys13000ND5F	TCTTATACCAACGCCTGAGC	2
		HD13865F	RTTAGAHACCCAAACCTCAG	2				
		HS14906R	GGGTTGGTTGTGGGTGGTGCT	1				
		14396LR	CGRCAGATRTGTGCGACTGAT	1	Mys14200CytBr	Mys14200CytBr	GGGTGTGTTTTTCGGATGTTG	2
11	HD13865F - HSeal16339tProR 1559 bp	HD13865F	RTTAGAHACCCAAACCTCAG	2				
		14396LR	CGRCAGATRTGTGCGACTGAT	1				
		HS15146F	CATTGACCTACCCACTCCATCT	1				
		HS15745R	GTGTAGTAGGGGTGGAATGGGAT	1				
		HS106031F	CCCGCACACATCAAACCAGAA	1	Mys14300CytBF	Mys14300CytBF	CCTAGCAATACACTACACACC	2
		HSeal16339tProR	TGGGTGTTGGTGGTGAGGT	1	Mys15400ProR	Mys15400ProR	TTGGGTGCTGATGGTGGAGTG	2

Table IV.4: Primers used for Sanger sequencing of NGS low QC sites. Primers were designed for this study, with the exception of Pma7913ATP8R, Pma3902tMetF and Pma6916tSerF (modified from Allan Wilson Center primers); Mys10000ND4LR, Mys13000ND5F, Mys14300CytBR (designed by M. Vant) and Pma6387COX1F (modified from Cooper *et al.* (2001)).

Fragment	Primer name	Sequence
1	Pma3902tMetF	5'-CCCATACCCCGAAAATGTTG-3'
1	Pma4275ND2R	5'-GAC TTC GGG TAY TCA GAA GTG GAA-3'
1	Pma6387COX1F	5'-TGAGCCTTRGGCTTCATTTTTCT-3'
2	Mys10000ND4LR	5'-TGCGGAYATTAGGTGGGATCG-3'
2	Pma6916tSerF	5'-AAGAAAGGAAGGAMTCGAAC-3'
2	Pma7913ATP8R	5'-TTGGTTTGGGATTAGGGGTG-3'
4	Eu14629R	5'-CAGAATGATATTTGTCCTCAGG-3'
4	Eu15639R	5'-GAATGCACGATTATACATAGC-3'
4	Mys13000ND5F	5'-TCTTATACCAACGCCTGAGC-3'
4	Mys14300CytBR	5'-GGTGTGTAGTGTATTGCTAGG-3'
4	PmaM13tpheR	5'-ANNCATCTTCAGTGTHTTGCTTT-3'
5	1.4UPR	5'-AGATAGAAACCGACCTGGATT-3'
5	Eu1241R	5'-ACGGTACTATCTCTATAGCG-3'
5	Eu1635F	5'-AGCAGCCATCAATTAAGAAAGC-3'
5	Eu1701R	5'-AGGAGTTGATTTGTTTAATGTTGG-3'

Figure IV.2: The *.xml *BEAST* files used in Bayesian analyses in this paper

Estimating lineage specific substitution rates using fossil calibrations
Supplementary_material_6A.xml:
http://gbe.oxfordjournals.org/content/suppl/2012/12/21/evs126.DC1/Supplementary_material_6A.xml

Phylogenetic reconstruction using coding region only
Supplementary_material_6B.xml:
http://gbe.oxfordjournals.org/content/suppl/2012/12/21/evs126.DC1/Supplementary_material_6B_17Nov2012.xml

Table IV.5: A summary of the sites that showed differences between sequencing platforms for each sample (used in calculations of cross-platform discrepancy rates). For each platform (Illumina, 454, Sanger) the base call at these sites is given, with "-" indicating an indel event relative to other platforms at that position, and *"NA"* indicating that sequence was not available from that platform. Cells shaded in blue indicate differences associated with homopolymer sites longer than 5 bp, those shaded in gray occurred on primer positions.

Platforms compared	Sample	Site	Call			Final QC call	Notes
			Illumina	454	Sanger		
Illumina, 454, and Sanger	PmaNZ005	16279	C	-	C	C	Corrected based on Illumina
	PmaNZ045	16018	-	G	-	-	Corrected based on Illumina
	PmaNZ045	16047	T	-	T	T	Corrected based on Illumina
	PmaNZ005	44	G	T	T	T	Corrected based on higher coverage 454
	PmaNZ005	178	A	T	T	N	Primer sequence
	PmaNZ005	10006	G	A	A	N	Primer sequence
	PmaNZ005	12982	A	G	G	G	In overlap between two PCR fragments, resolved when duplicate reads removed
454 and Sanger	PmaNZ076	16018	*NA*	G	-	-	Corrected based on Sanger
	PmaNZ076	16279	*NA*	-	C	C	Corrected based on Sanger
	PmaNZ082	1279	*NA*	-	A	A	Corrected based on Sanger
	PmaNZ082	16279	*NA*	-	C	C	Corrected based on Sanger
	PmaNZ082	16047	*NA*	-	T	T	Corrected based on Sanger
Illumina and Sanger	PmaNZ036	10006	G	*NA*	A	N	Primer sequence
	PmaNZ056	6793	A	*NA*	C	N	Primer sequence
	PmaNZ056	6796	T	*NA*	C	N	Primer sequence
	PmaNZ058	23	A	*NA*	T	T	Directly adjacent to primer site, may be from mispriming. Sanger used to verify identity at site.

Table IV.5 (Continued)

Platforms compared	Sample	Site	Call			Final QC call	Notes
			Illumina	454	Sanger		
	PmaNZ005	1140	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	1141	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	1142	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	1143	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	1144	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	1145	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	1146	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	1147	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	1148	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	1149	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	1279	A	-	*NA*	A	Corrected based on Illumina
	PmaNZ005	2252	A	-	*NA*	A	Corrected based on Illumina
	PmaNZ005	4082	A	-	*NA*	A	Corrected based on Illumina
	PmaNZ034	1142	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ034	1143	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ034	1144	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ034	1145	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ034	1146	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ034	1147	C	-	*NA*	C	Corrected based on Illumina
Illumina and 454	PmaNZ034	1148	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ034	16279	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ045	1141	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ045	1142	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ045	1143	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ045	1144	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ045	1145	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ045	1146	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ045	1147	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ045	1148	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ045	1923	A	-	*NA*	A	Corrected based on Illumina
	PmaNZ045	16279	C	-	*NA*	C	Corrected based on Illumina
	PmaNZ005	6793	A	C	*NA*	N	Primer sequence
	PmaNZ005	6796	T	C	*NA*	N	Primer sequence

Table IV.5 (Continued)

Platforms compared	Sample	Site	Call			Final QC call	Notes
			Illumina	454	Sanger		
	PmaNZ034	178	A	T	*NA*	N	Primer sequence
	PmaNZ034	3110	C	T	*NA*	N	Primer sequence
	PmaNZ034	6793	A	C	*NA*	N	Primer sequence
	PmaNZ034	6796	T	C	*NA*	N	Primer sequence
	PmaNZ034	10006	G	A	*NA*	N	Primer sequence
	PmaNZ045	178	A	T	*NA*	N	Primer sequence
	PmaNZ045	3110	C	T	*NA*	N	Primer sequence
Illumina and 454	PmaNZ045	6793	A	C	*NA*	N	Primer sequence
	PmaNZ045	6796	T	C	*NA*	N	Primer sequence
	PmaNZ045	10006	G	A	*NA*	N	Primer sequence
	PmaNZ034	1129	C	A	*NA*	C	Corrected based on Illumina
	PmaNZ034	12982	A	G	*NA*	G	In overlap between two PCR fragments, resolved when duplicate reads removed
	PmaNZ034	16348	A	G	*NA*	A	Corrected based on Illumina
	PmaNZ034	16351	A	T	*NA*	A	Corrected based on Illumina
	PmaNZ034	16377	C	T	*NA*	C	Corrected based on Illumina
	PmaNZ045	12982	A	G	*NA*	G	In overlap between two PCR fragments, resolved when duplicate reads removed

Figure IV.3: Average NGS coverage on variable sites in sperm whale mitogenome assemblies, with error bars reflecting 95 % confidence intervals. Samples with blue background were sequenced only on the 454 platform, those with a green background only on the Illumina platform, and those with the orange background were sequenced with both platforms (for each variable site, platform with highest coverage used in the construction of this figure).

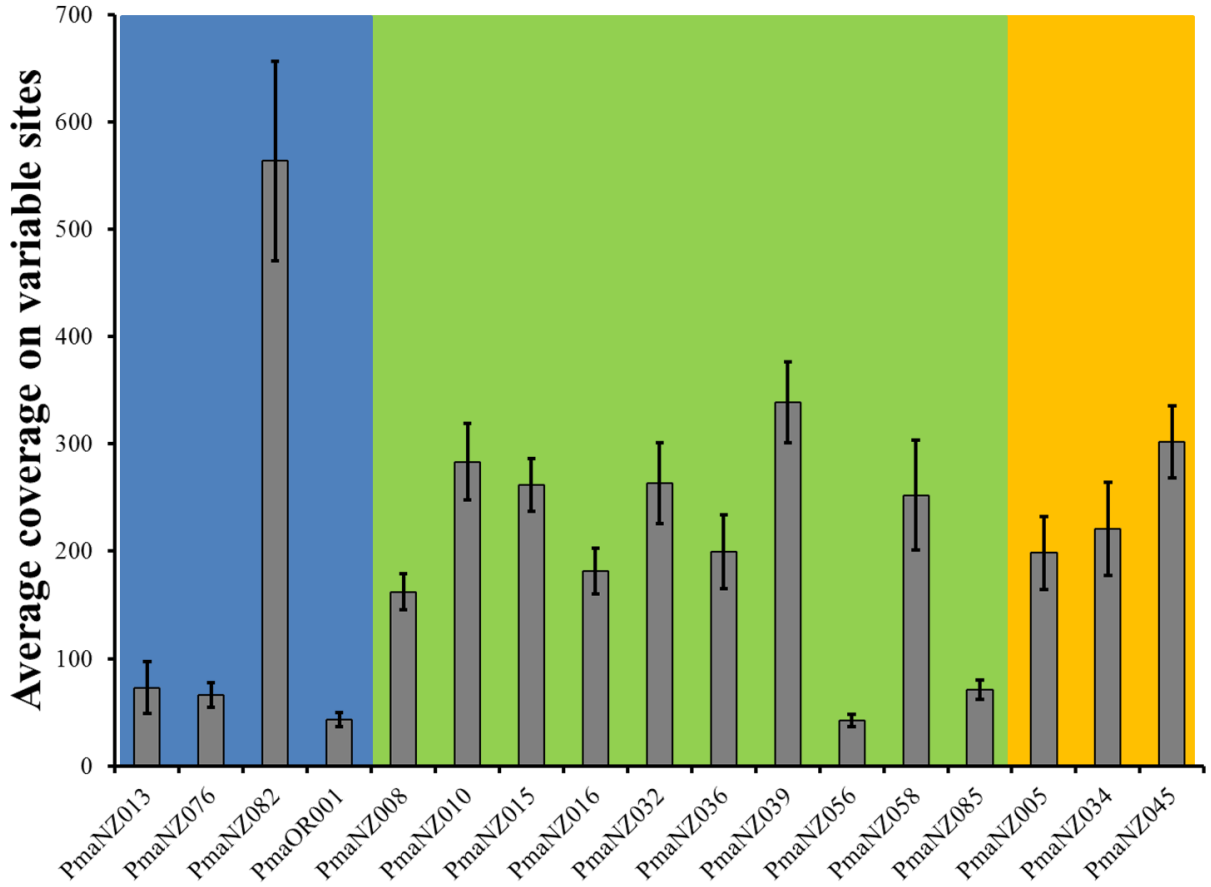


Table IV.6: The maximum pairwise divergence among the 17 sperm whale mitogenomes generated in this study. Maximum pairwise divergence is reported by gene. Length of each gene (bp) in the sperm whale is also reported**.** Regions are in the gene order represented by NC_002503 (Arnason et al. 2000).

Gene	Gene length in the sperm whale (bp)	Max pairwise % difference
tRNA-Phe	73	0
12srRNA	969	0
tRNA-Val	67	0
16srRNA	1589	0.13
tRNA-Leu	75	0
ND1	955	0.32
tRNA-Ile	69	0
tRNA-Gln	75	0
tRNA-Met	69	1.53
ND2	1042	0.39
tRNA-Trp	69	0
tRNA-Ala	69	1.52
tRNA-Asn	74	0
Origin of replication	35	0
tRNA-Cys	67	0
tRNA-Tyr	66	0
COX1	1551	0.39
tRNA-Ser	69	1.52
tRNA-Asp	68	0
COX2	684	0.59
tRNA-Lys	64	1.65
ATP8	192	0
ATP6	679	0.30
COX3	786	0.26
tRNA-Gly	69	0
ND3	346	0.59
tRNA-Arg	69	0
ND4L	297	0.29
ND4	1378	0.70
tRNA-His	69	0
tRNA-Ser_2	60	0
tRNA-Leu_2	70	0
ND5	1821	0.33
ND6	528	0
tRNA-Glu	69	0.44
CytB	1140	0.63
tRNA-Thr	70	0
tRNA-Pro	67	0
D-loop	954	0.96

Table IV.7: Variable sites among the 17 samples sequenced with positioning relative to the sperm whale mitogenome reference sequence NC_002503 (Arnason et al. 2000). Gene, codon position (if protein coding), amino acid substitutions, and mutational status of substitution (Transition/Transversion) are given. Singleton variable sites are shaded in gray. Some missing data (due to primer sites) were located at the following positions (bp): 44, 178; 3,107; 3,110; 6,793; 6,796; 10,006; 10,393; 12,762; 12,765.

Gene	16srRNA		*ND1*				tRNA-Met	*ND2*								tRNA-Ala
Codon position			3	3	3	3		3	1	1	3	3	2	3	1	
Amino acid substitution									A > T	V > M			T > M		A > T	
Transition/Transversion/Indel	Tv	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts
Position (relative to NC_002503)	1149	1858	2798	3206	3479	3650	3896	4104	4132	4234	4437	4566	4592	4749	4765	5078
PmaNZ005	**C**	**G**	**A**	**G**	**T**	**T**	**A**	**G**	**G**	**G**	**A**	**T**	**T**	**A**	**G**	**T**
PmaNZ008	**A**	.	.	.	.	.	.	.	.	.	**G**	.	**C**	.	.	.
PmaNZ010	**A**	**A**	.	.	.	.	.	.	.	.	**G**	.	.	.	**A**	.
PmaNZ013	**A**	.	.	.	.	.	.	.	.	.	**G**	.	.	**G**	.	.
PmaNZ015	**A**	.	.	.	.	.	.	**A**	.	.	**G**	.	.	.	.	**C**
PmaNZ016	**A**	.	.	.	.	.	.	.	**A**	.	**G**	.	.	**G**	.	.
PmaNZ032	**A**	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.
PmaNZ034	**A**	.	.	.	.	.	.	.	.	.	**G**	.	**C**	.	.	.
PmaNZ036	**A**	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.
PmaNZ039	**A**	.	.	.	**C**	**C**	.	.	.	.	**G**	.	.	.	.	.
PmaNZ045	**A**	.	.	.	.	.	.	.	.	.	**G**	.	.	.	.	.
PmaNZ056	**A**	.	.	.	**C**	**C**	.	.	.	.	**G**	.	.	.	.	.
PmaNZ058	**A**	**A**	.	.	.	.	.	.	.	.	**G**	**C**	.	.	**A**	.
PmaNZ076	**A**	.	.	.	**C**	**C**	.	.	.	.	**G**	.	.	.	.	.
PmaNZ082	**A**	.	.	.	.	.	**G**	.	.	.	**G**	.	.	.	.	.
PmaNZ085	.	.	.	**A**	.	.	.	.	.	**A**	**G**	.	.	.	.	.
PmaOR001	.	.	**G**	.	.	.	.	.	.	.	**G**	.	.	.	.	.
# of haplotypes per partition	3		4				2	9								2

Table IV.7 (Continued)

Gene	*COX1*							tRNA-Ser	*COX2*						tRNA-Lys
Codon position	3	3	3	3	3	3	3		3	3	1	3	3	1	
Amino acid substitution											I > V			P > S	
Transition/Transversion/Indel	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts
Position (relative to NC_002503)	5425	5659	5995	6034	6607	6655	6838	6937	7112	7214	7248	7319	7529	7701	7773
PmaNZ005	**T**	**A**	**G**	**G**	**T**	**T**	**T**	**T**	**G**	**T**	**A**	**A**	**G**	**C**	**T**
PmaNZ008	**.**	**.**	**.**	**.**	**.**	**.**	**C**	**C**	**.**	**.**	**.**	**.**	**.**	**.**	**.**
PmaNZ010	**.**	**G**	**.**	**.**	**.**	**.**	**.**	**C**	**.**	**C**	**.**	**G**	**.**	**.**	**.**
PmaNZ013	**.**	**.**	**.**	**A**	**.**	**.**	**C**	**C**	**.**	**.**	**.**	**.**	**.**	**.**	**.**
PmaNZ015	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**
PmaNZ016	**C**	**.**	**A**	**.**	**C**	**C**	**C**	**C**	**.**	**.**	**.**	**.**	**.**	**.**	**.**
PmaNZ032	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**
PmaNZ034	**.**	**.**	**.**	**.**	**.**	**.**	**C**	**C**	**.**	**.**	**.**	**.**	**.**	**.**	**.**
PmaNZ036	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**
PmaNZ039	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**A**	**.**	**.**	**.**	**A**	**.**	**.**
PmaNZ045	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**T**	**.**
PmaNZ056	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**A**	**.**	**.**	**.**	**A**	**.**	**.**
PmaNZ058	**.**	**G**	**A**	**.**	**.**	**.**	**.**	**C**	**.**	**.**	**.**	**G**	**.**	**.**	**.**
PmaNZ076	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**A**	**.**	**.**	**.**	**A**	**.**	**.**
PmaNZ082	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**C**
PmaNZ085	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**G**	**.**	**.**	**.**	**.**
PmaOR001	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**
# of haplotypes per partition	6							2	6						2

Table IV.7 (Continued)

Gene	*ATP6*				*COX3*		*ND3*		*ND4L*			*ND4*				
Codon position	1	1	2	3	3	3	3	1	3	1	3	3	2	3	3	3
Amino acid substitution	T > A	A > T	I > T							V > I			A > V			
Transition/Transversion/Indel	Ts	Ts	Ts	Ts	Ts	Tv	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Tv	Ts	Ts
Position (relative to NC_002503)	8213	8351	8406	8482	8754	9279	9518	9840	10090	10160	10183	10965	10988	11010	11226	11418
PmaNZ005	**G**	**G**	**T**	**G**	**A**	**T**	**C**	**T**	**A**	**G**	**C**	**T**	**C**	**A**	**T**	**A**
PmaNZ008	**A**	.	.	.	.	.	**T**	.	**G**	.	.	.	.	.	.	.
PmaNZ010	.	.	**C**	.	**G**	.	**T**	**C**	**G**	.	.	.	.	.	**C**	.
PmaNZ013	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.	.	.
PmaNZ015	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ016	.	**A**	.	**A**	.	.	**T**	.	.	.	.	.	.	.	.	.
PmaNZ032	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ034	**A**	.	.	.	.	.	**T**	.	**G**	.	.	.	.	.	.	.
PmaNZ036	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ039	.	.	.	.	.	**A**	.	.	.	.	.	.	**T**	**T**	.	.
PmaNZ045	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**G**
PmaNZ056	.	.	.	.	.	**A**	.	.	.	.	.	.	**T**	.	.	.
PmaNZ058	.	.	**C**	.	**G**	.	**T**	**C**	.	.	**T**	**C**	.	.	**C**	.
PmaNZ076	.	.	.	.	.	**A**	.	.	.	**A**	.	.	**T**	.	.	.
PmaNZ082	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ085	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaOR001	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
# of haplotypes per partition	4				3		3		4			6				

Table IV.7 (Continued)

Gene	*ND5*										*ND6*		
Codon position	3	3	2	1	3	3	3	3	3	3	1	3	3
Amino acid substitution			S > N	F > L							I > V		
Transition/Transversion/Indel	Tv	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts
Position (relative to NC_002503)	11808	11823	11861	12031	12108	12732	12873	13083	13227	13515	13615	13655	13712
PmaNZ005	**C**	**C**	**G**	**T**	**A**	**G**	**A**	**G**	**C**	**T**	**T**	**C**	**G**
PmaNZ008	.	.	.	.	**G**	.	.	.	.	.	.	.	.
PmaNZ010	.	**T**	.	**C**	.	.	**G**	.	.	.	**C**	.	**A**
PmaNZ013	**A**	.	.	.	.	.	.	.	.	.	.	**T**	.
PmaNZ015	.	.	**A**	.	.	.	.	.	.	**C**	.	.	.
PmaNZ016	**A**	.	.	.	.	.	.	.	.	.	.	**T**	.
PmaNZ032	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ034	.	.	.	.	**G**	.	.	.	.	.	.	.	.
PmaNZ036	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ039	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ045	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ056	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ058	.	.	.	**C**	.	**A**	.	**A**	**T**	.	.	.	**A**
PmaNZ076	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ082	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ085	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaOR001	.	.	.	.	.	.	.	.	.	.	.	.	.
# of haplotypes per partition	6										4		

Table IV.7 (Continued)

Gene	*CytB*						
Codon position	2	3	3	3	3	3	3
Amino acid substitution	T > N						
Transition/Transversion/Indel	Tv	Ts	Ts	Ts	Ts	Ts	Ts
Position (relative to NC_002503)	14203	14414	14438	14549	14597	14957	15152
PmaNZ005	**C**	**C**	**A**	**G**	**G**	**T**	**A**
PmaNZ008	.	.	.	.	.	**C**	.
PmaNZ010	.	.	.	.	.	.	.
PmaNZ013	.	.	.	.	.	.	.
PmaNZ015	.	.	.	.	.	.	.
PmaNZ016	.	**T**	.	**A**	**A**	.	**G**
PmaNZ032	.	.	.	.	.	.	.
PmaNZ034	.	.	.	.	.	**C**	.
PmaNZ036	.	.	.	.	.	.	.
PmaNZ039	.	.	**G**	.	.	.	.
PmaNZ045	.	.	.	.	.	.	.
PmaNZ056	.	.	**G**	.	.	.	.
PmaNZ058	.	.	.	.	.	.	.
PmaNZ076	.	.	**G**	.	.	.	.
PmaNZ082	**A**	.	.	.	.	.	.
PmaNZ085	.	.	.	.	.	.	.
PmaOR001	.	.	.	.	.	.	.
# of haplotypes per partition	5						

Table IV.7 (Continued)

Gene	Control region														
Codon position															
Amino acid substitution															
Transition/Transversion/Indel	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Tv	Ts	Ts	Ts	Ts
Position (relative to NC_002503)	15536	15579	15581	15658	15734	15746	15747	15762	15793	15798	16048	16082	16098	16269	16407
PmaNZ005	**T**	**C**	**A**	**T**	**A**	**A**	**C**	**A**	**G**	**C**	**G**	**A**	**G**	**T**	**A**
PmaNZ008	**C**	.	**G**	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ010	.	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.
PmaNZ013	**C**	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.
PmaNZ015	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ016	**C**	.	.	.	.	.	.	**G**	.	.	.	.	.	.	**G**
PmaNZ032	.	.	.	.	.	.	.	.	.	.	**T**	.	.	**C**	.
PmaNZ034	**C**	.	.	.	.	.	.	.	.	.	.	**G**	.	.	.
PmaNZ036	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
PmaNZ039	.	.	.	**C**	**G**	.	.	.	**A**	**T**	.	.	.	.	.
PmaNZ045	.	.	.	.	.	.	.	.	.	.	.	.	**A**	.	.
PmaNZ056	.	**T**	.	**C**	**G**	**G**	.	.	**A**	**T**	.	.	.	.	.
PmaNZ058	.	.	.	.	.	.	.	**G**	.	.	.	.	.	.	.
PmaNZ076	.	.	.	**C**	**G**	**G**	.	.	**A**	**T**	.	.	.	.	.
PmaNZ082	.	.	.	.	.	.	**T**	.	.	.	.	.	.	.	.
PmaNZ085	.	.	.	.	.	.	.	.	**A**	.	.	.	.	.	.
PmaOR001	,	.	.	.	.	.	.	.	.	.	.	.	.	**C**	.
# of haplotypes per partition	14														

Table IV.8: Mean and standard error for gene-specific (and for the protein-coding regions, codon-specific) nucleotide diversity (in %) for the 17 sperm whale mitogenomes sequenced in this study. Regions are in the gene order represented by NC_002503 (Arnason et al. 2000).

Gene/codon	Length (bp)	π (%)	S.E.
tRNA-Phe	73	0.00	0.00
12s rRNA	969	0.00	0.00
tRNA-Val	67	0.00	0.00
16s rRNA	1590	0.03	0.03
tRNA-Leu	75	0.00	0.00
ND1 Total	957	0.09	0.05
ND1 1st codon	319	0.00	0.00
ND1 2nd codon	319	0.00	0.00
ND1 3rd codon	319	0.28	0.15
tRNA-Ile	69	0.00	0.00
tRNA-Gln	76	0.00	0.00
tRNA-Met	69	0.18	0.21
ND2 Total	1044	0.12	0.04
ND2 1st codon	348	0.13	0.08
ND2 2nd codon	348	0.06	0.06
ND2 3rd codon	348	0.17	0.09
tRNA-Trp	69	0.00	0.00
tRNA-Ala	69	0.18	0.20
tRNA-Asn	74	0.00	0.00
Origin of light strand replication	35	0.00	0.00
tRNA-Cys	67	0.00	0.00
tRNA-Tyr	66	0.00	0.00
COX1 Total	1551	0.08	0.03
COX1 1st codon	517	0.00	0.00
COX1 2nd codon	517	0.00	0.00
COX1 3rd codon	517	0.26	0.11
tRNA-Ser	69	0.74	0.80
tRNA-Asp	68	0.00	0.00
COX2 Total	684	0.18	0.08
COX2 1st codon	228	0.10	0.07
COX2 2nd codon	228	0.00	0.00
COX2 3rd codon	228	0.45	0.26
tRNA-Lys	65	0.19	0.22
ATP8 Total	192	0.00	0.00
ATP8 1st codon	64	0.00	0.00

Table IV.8 (Continued)

Gene/codon	Length (bp)	π (%)	S.E.
ATP8 2nd codon	64	0.00	0.00
ATP8 3rd codon	64	0.00	0.00
ATP6 Total	681	0.10	0.05
ATP6 1st codon	227	0.15	0.11
ATP6 2nd codon	227	0.10	0.10
ATP6 3rd codon	227	0.05	0.06
COX3 Total	786	0.07	0.05
COX3 1st codon	262	0.00	0.00
COX3 2nd codon	262	0.00	0.00
COX3 3rd codon	262	0.20	0.16
tRNA-Gly	69	0.00	0.00
ND3 Total	348	0.21	0.15
ND3 1st codon	116	0.20	0.20
ND3 2nd codon	116	0.00	0.00
ND3 3rd codon	116	0.44	0.47
tRNA-Arg	69	0.00	0.00
ND4L Total	297	0.19	0.12
ND4L 1st codon	99	0.12	0.13
ND4L 2nd codon	99	0.00	0.00
ND4L 3rd codon	99	0.47	1.99
ND4 Total	1380	0.06	0.03
ND4 1st codon	460	0.00	0.00
ND4 2nd codon	460	0.07	0.06
ND4 3rd codon	460	0.13	0.06
tRNA-His	69	0.00	0.00
tRNA-Ser	60	0.00	0.00
tRNA-Leu	70	0.00	0.00
ND5 Total	1821	0.08	0.03
ND5 1st codon	607	0.04	0.03
ND5 2nd codon	607	0.02	0.02
ND5 3rd codon	607	0.19	0.07
ND6 Total	528	0.11	0.06
ND6 1st codon	176	0.07	0.07
ND6 2nd codon	176	0.00	0.00
ND6 3rd codon	176	0.26	0.18
tRNA-Glu	69	0.00	0.00
CytB Total	1140	0.10	0.04
CytB 1st codon	380	0.00	0.00

Table IV.8 (Continued)

Gene/codon	Length (bp)	π (%)	S.E.
CytB 2nd codon	380	0.03	0.03
CytB 3rd codon	380	0.27	0.12
tRNA-Thr	70	0.00	0.00
tRNA-Pro	67	0.00	0.00
Overall mitogenome	16,428	0.10	0.02
Control region	954	0.35	0.10
Overall coding region	10,815	0.10	0.02
Coding total 1st codon	3605	0.04	0.02
Coding total 2nd codon	3605	0.03	0.01
Coding total 3rd codon	3605	0.22	0.04
rRNAs total	2559	0.02	0.01
tRNAs total	1514	0.06	0.03

Table IV.9: Heteroplasmies detected for sperm whale mitogenomes sequenced in this study. Mitogenomes with detected heteroplasmies shaded in gray.

Sample	Position	Platform	Coverage	Major allele	Frequency	Minor allele	Frequency	Gene	Codon position	Substitution	Verified by sanger
PmaNZ005	NA	Illumina	*No heteroplasmies observed*								
PmaNZ005	NA	454	*No heteroplasmies observed*								
PmaNZ008	NA	Illumina	*No heteroplasmies observed*								
PmaNZ010	NA	Illumina	*No heteroplasmies observed*								
PmaNZ013	NA	454	*No heteroplasmies observed*								
PmaNZ015	NA	Illumina	*No heteroplasmies observed*								
PmaNZ016	NA	Illumina	*No heteroplasmies observed*								
PmaNZ032	NA	Illumina	*No heteroplasmies observed*								
PmaNZ034	12975	Illumina	170	A	0.6824	C	0.3176	ND5	3rd	No	NA
PmaNZ034	NA	454	*Excluded due to the inconsistencies shown between Illumina and 454 in variable site QC process*								
PmaNZ036	158	Illumina	1244	G	0.6214	A	0.3778	12s rRNA	NA	NA	NA
PmaNZ036	943	Illumina	175	T	0.6914	C	0.3086	12s rRNA	NA	NA	NA
PmaNZ036	1858	Illumina	145	G	0.6621	A	0.3379	16S rRNA	NA	NA	NA
PmaNZ036	6095	Illumina	120	G	0.6750	A	0.3167	COX1	3rd	No	NA
PmaNZ036	8754	Illumina	232	A	0.6983	G	0.3017	COX3	3rd	No	NA
PmaNZ036	14831	Illumina	139	T	0.6187	C	0.3813	CYTB	3rd	No	NA
PmaNZ036	15658	Illumina	59	T	0.6949	C	0.3051	D-loop	NA	NA	NA
PmaNZ039	NA	Illumina	*No heteroplasmies observed*								
PmaNZ045	NA	Illumina	*No heteroplasmies observed*								
PmaNZ045	NA	454	*No heteroplasmies observed*								

Table IV.9 (Continued)

Sample	Position	Platform	Coverage	Major allele	Frequency	Minor allele	Frequency	Gene	Codon position	Substitution	Verified by sanger
PmaNZ056	NA	Illumina	*No heteroplasmies observed*								
PmaNZ058	NA	Illumina	*No heteroplasmies observed*								
PmaNZ076	NA	454	*No heteroplasmies observed*								
PmaNZ082	NA	454	*No heteroplasmies observed*								
PmaNZ085	1610	Illumina	43	G	0.6826	C	0.3175	16S rRNA	NA	NA	NA
PmaNZ085	4235	Illumina	30	A	0.6333	G	0.3667	ND2	1st	V > M	Confirmed by Sanger
PmaOR001	NA	454	*No heteroplasmies observed*								

Table IV.10: Mitogenomic diversity estimates for selected mammalian species

Species	**Sample size**	**Nucleotide diversity (%) +/- SE**	**Reference**
Sperm whales	17	0.096 +/- 0.015	This study
Killer whale (*Orcinus orca*) Pacific transient Atlantic Type 1 Pacific resident Pacific offshore	24 21 15 16	0.047 0.035 0.021 0.024	(Foote et al. 2011a)
Pig (*Sus scrofa*) Indigenous Chinese breeds	17	0.091 +/- 0.010	(Yang et al. 2003)
Fisher (*Martes pennant*)	40	0.088	(Knaus et al. 2011)
Przewalksi's horse (*Equus przewalskii)*	4	0.54 +/- 0.064	(Goto et al. 2011b)
Common chimpanzee (*Pan troglodytes)*	10	1.48	(Stone et al. 2010)
Human (*Homo sapiens)*	53	0.37	(Stone et al. 2010)
Human (*H. sapiens)* European Chinese Melanesian Maori	189 67 52 50 20	0.174 0.145 0.186 0.172 0.018	(Benton et al. 2012)
Human (*H. sapiens*) African Non-African Tanzanian	320 94 226 49	0.269 0.392 0.181 0.380	(Gonder et al. 2007)
Tasmanian tiger (*Thylacinus cynocephalus*)	2	0.032	(Miller et al. 2009)

Appendix IV References:

Árnason Ú, Gullberg A. 1993. Comparison between the complete mtDNA sequences of the blue and the fin whale, two species that can hybridize in nature. J Mol Evol. 37: 312-322.

Arnason U, Gullberg A, Gretarsdottir S, Ursing B, Janke A. 2000. The mitochondrial genome of the sperm whale and a new molecular reference for estimating Eutherian divergence dates. J Mol Evol. 50: 569-578.

Arnason U, Gullberg A, Janke A. 2004. Mitogenomic analyses provide new insights into cetacean origin and evolution. Gene. 333: 27-34.

Arnason U, Gullberg A, Widegren B. 1991. The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balaenoptera physalus*. J Mol Evol. 33: 556-568.

Benton M et al. 2012. Complete mitochondrial genome sequencing reveals novel haplotypes in a Polynesian population. PLoS ONE. 7: e35026.

Carraher CJF. 2004. Comparative mitogenomics of the Southern Hemisphere dolphin genus *Cephalorhynchus.* MSc Thesis. The University of Auckland, Auckland.

Cooper A et al. 2001. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. Nature. 409: 704-707.

Foote AD et al. 2011a. Out of the Pacific and back again: insights into the matrilineal history of Pacific killer whale ecotypes. PLoS ONE. 6: e24980.

Foote AD et al. 2011b. Genetic differentiation among North Atlantic killer whale populations. Mol Ecol. 20: 629-641.

Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA. 2007. Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol. 24: 757-768.

Goto H et al. 2011. A massively parallel sequencing approach uncovers ancient origins and high genetic variability of endangered Przewalski's horses. Genome Biol Evol. 3: 1096-1106.

Knaus BJ, Cronn R, Liston A, Pilgrim K, Schwartz MK. 2011. Mitochondrial genome sequences illuminate maternal lineages of conservation concern in a rare carnivore. BMC Ecol. 11: 10.

Miller W et al. 2009. The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*). Genome Res. 19: 213-220.

Sasaki T et al. 2005. Mitochondrial phylogenetics and evolution of mysticete whales. Syst Biol. 54: 77-90.

Sasaki T et al. 2006. *Balaenoptera omurai* is a newly discovered baleen whale that represents an ancient evolutionary lineage. Mol Phylogenet Evol. 41: 40-52.

Stone AC et al. 2010. More reliable estimates of divergence times in *Pan* using complete mtDNA sequences and accounting for population structure. P Roy Soc B-Biol Sci. 365: 3277-3288.

Xiong Y, Brandley M, Xu S, Zhou K, Yang G. 2009. Seven new dolphin mitochondrial genomes and a time-calibrated phylogeny of whales. BMC Evol Biol. 9: 20.

Yan J, Zhou K, Yang G. 2005. Molecular phylogenetics of 'river dolphins' and the baiji mitochondrial genome. Mol Phylogenet Evol. 37: 743-750.

Yang J et al. 2003. Genetic diversity present within the near-complete mtDNA genome of 17 breeds of indigenous Chinese pigs. J Hered. 94: 381-385.

Appendix V
Supplementary material associated with Chapter 4: A bottleneck for *Moby Dick*? Low mtDNA and nuclear diversity in the sperm whale

The files listed below are archived with the author, as well as the Cetacean Conservation and Genomics Laboratory at Oregon State University's Hatfield Marine Science Center. Requests for data access can be submitted to Alana Alexander (alanamaryalexander@gmail.com) and C. Scott Baker (scott.baker@oregonstate.edu), and should be accompanied by a proposal describing how the data would be used. A two-year embargo applies to the data, however, requests made before two years will be considered on a case-by-case basis, depending upon the proposed use and potential for collaboration.

Nuclear genotypes for individuals sequenced in this study

File name: Nuclear_genotypes_Pma_Kbr_25Aug2014.txt

Individual code and allele calls for the nuclear DNA sequence generated in this study. Sequence definition for allele calls given in 'Nuclear allele definitions' below.

Nuclear allele definitions

Folder name: Nuclear_allele_definitions_Pma_Kbr_25Aug2014

Fasta files with the DNA sequence for the pygmy and sperm whale alleles defined for each nuclear locus sequenced in this study.

mtDNA sequence definitions

Folder name: mtDNA_definitions_Pma_Kbr_25Aug2014

Fasta files with the mtDNA control region and protein-coding region sequence for the pygmy sperm and sperm whales sequenced in this study.

BEAST files

Folder name: BEAST_files_Pma_Kbr_25Aug2014

Multispecies *.xml files for generating species-specific substitution rates and files for implementing the sperm whale- and pygmy sperm whale-specific extended Bayesian skyline plot analysis in *BEAST*.

Table V.1: Primers used for amplification of PCR fragments in this study, along with length of amplified fragment (including primers), annealing temperature, dNTP concentration and amount of polymerase (fragments > 1400 bp: Phusion, fragments < 1400 bp: Platinum *Taq*). Some primers were species specific (Pma: sperm whale/Kbr: pygmy sperm whale). References: 1. (Lyons et al. 1997) 2. (Alexander et al. 2013); 3. (Meredith et al. 2011); [4] B. Slikas, personal communication; [5] This study. * indicates modified from primer first introduced by this reference.

Fragment	Species	Primer name and sequence	Position relative to PmaNZ005	Ref	Length of fragment (bp)	Annealing temperature (°C)	dNTP (mM)	Polymerase amount (units)
CHRNA1	Pma/Kbr	CHRNA1F: 3’ - GAC CAT GAA GTC AGA CCA GGA G - 5’	-22 to -1	[1]	387	62	0.1875	0.1
	Pma/Kbr	CHRNA1R: 3’ - GGA GTA TGT GGT CCA TCA CCA T - 3’	344-365					
ESD	Pma/Kbr	ESDF: 3’ - TTT TGG ACA CTC CAT GGG AG - 3’	-97 to -74	[1]	825	55	0.1875	0.1
	Pma/Kbr	ESDR: 3’ - TTC CAT TTA CTT TGA TCT GTT CCC - 3’	709 to 728					
FGG	Pma/Kbr	FGGF: 3’ - TCA AGA CAT TGC CAA TAA GGG - 3’	-60 to -40	[1]	1,162	62	0.1875	0.1
	Pma/Kbr	FGGR: 3’ - CCA GTA GGA GAC AGA TGT CCA AA - 3’	1,080 to 1,102					
G6PD	Pma/Kbr	G6PDF: 3’ - CGT GAG GAC CAG ATC TAC CG - 3’	-54 to -37	[1]	557	62	0.1875	0.1
	Pma/Kbr	G6PDR: 3’ - GCA GGA GGT TCT GCA - 3’	484 to 503					
RHO	Pma/Kbr	RHOF: 3’ - AGG GGA GGT CAC TTT ATA AGG G - 3’	-80 to -59	[1]	268	62	0.1875	0.1
	Pma/Kbr	RHOR: 3’ - CCA GCA TGG AGA ACT GCC - 3’	171 to 188					
Mito 1A	Pma/Kbr	1.4UPF: 5’-AAT CCA GGT CGG TTT CTA TCT-3’	2,545-2,565	[2]	1,752	60	0.2	0.1
	Pma/Kbr	Pma4275ND2R: 5'-GAC TTC GGG TAY TCA GAA GTG GAA-3'	4,274-4,297					
Mito 1B	Pma/Kbr	Pma3902tmetF: 5’-CCC ATA CCC CGA AAA TGT TG-3’	3,903-3,922	[2]	3,033	60	0.2	0.2
	Pma/Kbr	Pma6916tserR: 5’-GTT CGA KTC CTT CCT TTC TT-3’	6,917-6,936					
Mito 2	Pma/Kbr	Pma6800CO1F: 5’-GAG AAG CMT TYR CAT CCA AAC G-3’	6,787-6,808	[2]				
	Pma	PmaHDND4R: 5’-GGG TCA GAG AAG AAC GTA AGG G-3’	10,373-10,394	[2]	3,607: Pma	61	0.2	0.2
	Kbr	Kbr10365ND4R: 5' – GAG TCG GAG AAG AAT GTT AGG G - 3'	10,373-10,394	[5]	3,617: Kbr	61	0.4	0.2
Mito 3	Pma/Kbr	Mys10000ND4LF: 5’-CGA TCC CAC CTA ATR TCC GCA-3’	9,992-10,012	[2]	3,024	60	0.2	0.2
	Pma/Kbr	Mys13000ND5R: 5’-GCT CAG GCG TTG GTA TAA GA-3’	12,997-13,016					

Table V.1 (Continued)

Fragment	Species	Primer name and sequence	Position relative to PmaNZ005	Ref	Length of fragment (bp)	Annealing temperature (°C)	dNTP (mM)	Polymerase amount (units)
Mito	Pma/Kbr	Eu12831F: 5'-TGG CTC CAT CCA TAA CC-3'	12,822-12,841	[4]	3,797	61	0.4	0.2
4*	Pma/Kbr	Pma12sRNAR: 5'-GTG CTT GAT ACC WGC TCC TTT T-3'	170-191	[2]				
Mito	Pma/Kbr	EjapM13tpheF: 5' – AAA GCA AGA CAC TGA AAA TGT CT – 3'	25-47	[4]	2,540	60	0.4	0.2
5A	Pma/Kbr	1.4UPR: 5'-AGA TAG AAA CCG ACC TGG ATT-3'	2,545-2,565	[2]				
Mito	Pma/Kbr	Eu1635F: 5'-AGC AGC CAT CAA TTA AGA AAG C-3'	1,638-1,659	[2]	1,478	63	0.4	0.2
5B	Pma/Kbr	PmaHD3106R: 5'-TAR ACR GYY AGG CTG GAT ATT GC-3'	3,094-3,116					
BRCA1	Pma/Kbr	BRCA1_CetartF1: 5' - CTG AGA MGC ATC CAG AAA ART ATS AGG G – 3'	-28 - -1	[3]	1,898	66	0.4	0.2
1	Pma/Kbr	PmaBRCA1_1834R: 5' – AGA GYG GGC ACA GAC TGT TGC – 3'	1,850-1,870	[5]				
BRCA1	Pma/Kbr	PmaBRCA1_1065F: 5' – TGA ACT AGT CAG TAG AAA CC – 3'	1,081-1,103	[5]	1,802	53	0.4	0.2
2	Pma/Kbr	PmaBRCA1_2844R: 5' – GGA AGC ARG GAA GCT CTT C – 3'	2,864-2882	[3]*				
ApoB	Pma/Kbr	PmaApoB_22F: 5' - AAC AAT GAG AAA AGC ATY GAG GC – 3'	-23 - -1	[5]	1,897	51	0.2	0.2
1	Pma/Kbr	PmaApoB_1891R: 5' - GTT TCA GTC CRG GAA TAA TGA ATT TAT C – 3'	1,847-1,874	[3]*				
ApoB	Pma/Kbr	PmaApoB_755F: 5' - CAC TGA ATA CCA ATG TTG GTC TTT ATA ACC - 3'	711-740	[3]*	1,703	67	0.4	0.2
2	Pma/Kbr	PmaApoB_2433R: 5'- TGT GGA ATC CAG GAA TGT TTC AAC G - 3'	2,389-2,413	[5]				
TTN	Pma/Kbr	PmaTTN_1F: 5' - GGT GAA ATC CGA GAT GCA GCT ATM ATY G - 3'	-28 - -1	[3]*	1,942	69	0.4	0.2
1	Pma/Kbr	PmaTTN_1834R: 5'- CC ATT GAA ATT CTT GGG AGT TCG AC - 3'	1,890 - 1,914					
TTN	Pma/Kbr	PmaTTN_1353F: 5' - TGA GAG AAA GGA AAG AAA TAG TAT TTT GTG G - 3'	1,409 - 1,439	[3]*	1,715	67	0.4	0.2
2	Pma/Kbr	PmaTTN_3040R: 5' - TTA ACT CAG AGT CAA GGT CAA GTT CAG G - 3'	3,096 - 3,123					
TTN	Pma/Kbr	PmaTTN_2780F: 5' - CAW TTA CAC TTG CCT GGG GTA AAC C - 3'	2,836-2,860	[3]*	1,636	66	0.4	0.2
3	Pma/Kbr	PmaTTN_4390R: 5' - AAA ATT CTA AAC TGG TAC TCA GTG CC - 3'	4,446-4,471	[5]				

* Forward primer modified in comparison with Alexander et al. (2013) for this fragment

Figure V.1: Code for *PEppyphase*, a program written to phase nuclear DNA next-generation data by using reads physically linked between variable sites. *PEppyphase* has three components: *PEphase_shell* (the Linux script that executes the component commands), *generate_bed.R* (an R script that generates files to be used with *BEDtools*), and *PEphase.R* (the R script that carries out the summary of phase across the defined variable sites). *PEppyphase* expects a fastq file of paired-end reads (although the code in *PEphase_shell* could be modified to carry out a single-end assembly) called 'Input.fastq', and a reference sequence in fasta format called 'Reference.fas'. 'Reference.fas' is used as the reference for the position of variable sites listed in the last file required for *PEppyphase:* 'variable_sites'. 'variable_sites' is a tab-delimited file of the position of variable sites, and the number of nucleotides to record at each of the variable sites (in case of a repeated insertion where multiple bases would be needed to confirm the variants). 'variable_sites' should have the variable site position in the first column (in the first row), and separated by a tab, the number of bases to record at the variable site e.g.

287 1
1807 1
1981 7
2328 1

In addition, *PEppyphase* depends on the installation of *BEDTools* (Quinlan and Hall 2010), *samtools* (Li et al. 2009a), *bwa* (Li and Durbin 2009) and *R* (R Core Team 2013). The following directories are listed in *PEphase_shell* and will need to be modified if an updated installation is used, or the directories containing the programs are in a different location:

~/bin/bedtools2-2.19.1
/local/cluster/bin/samtools
~/bin/bwa-0.7.4

PEppyphase expects to have all of its scripts located in the same directory as your 'Input.fastq', 'Reference.fas', and 'variable_sites' files. *PEppyphase* outputs a tab-delimited file: 'phased_reads_output.txt' listing the name of all reads that cover more than one variable site, and their nucleotide at each of these sites. If a read does not cover a specific site, 'NA' is returned. If forward and reverse pairs disagree about the base call at a site, 'X' is returned. This file can then be summarized by sorting or using a pivot table in data visualization software. Ongoing development will aim to make *PEppyphase* more flexible and automated. Code for *PEphase_shell*, *generate_bed.R* and *PEphase.R* is given on the following pages. Code for the function 'substrRight' in *PEphase.R* obtained from http://stackoverflow.com/a/7963963.

Figure V.1 (Continued)

PEPhase_shell

```
#!/bin/bash

dos2unix generate_bed.R
dos2unix PEphase.R
dos2unix variable_sites
dos2unix Reference.fas

Rscript generate_bed.R

~/bin/bwa-/bwa index -a is Reference.fas

/local/cluster/bin/samtools faidx Reference.fas

~/bin/bwa-/bwa mem -p Reference.fas Input.fastq > Aln-pe.sam

/local/cluster/bin/samtools view -bS -o Aln-pe.bam Aln-pe.sam

/local/cluster/bin/samtools sort Aln-pe.bam Aln-pe_sorted

/local/cluster/bin/samtools rmdup -s Aln-pe_sorted.bam Aln-
pe_sorted_rmdup.bam

/local/cluster/bin/samtools mpileup -f Reference.fas Aln-
pe_sorted_rmdup.bam > Pileup

rm -rf Aln-pe.bam

rm -rf Aln-pe.sam

~/bin/bedtools2-/bin/bamToBed -i Aln-pe_sorted_rmdup.bam >
BedReference.bed

~/bin/bedtools2-/bin/subtractBed -a BedReference.bed -b B.bed >
Extracted_regions.bed

~/bin/bedtools2-/bin/intersectBed -a Aln-pe_sorted_rmdup.bam -b
Extracted_regions.bed > Regions_of_interest.bam

/local/cluster/bin/samtools view -o Regions_of_interest.sam
Regions_of_interest.bam

rm -rf Aln-pe_sorted_rmdup.bam

rm -rf Aln-pe_sorted.bam

Rscript PEphase.R

rm -rf BedReference.bed

rm -rf Extracted_regions.bed
```

Figure V.1 (Continued)

```
rm -rf Regions_of_interest.bam
```

Figure V.1 (Continued)

generate_bed.R

```
rm(list=ls())

#Pulling in our list of variable sites
variable_sites <- read.table("variable_sites", header=FALSE, sep="\t",
stringsAsFactors=FALSE)

#Reading in reference to get the total length of the alignment
reference <- read.csv("Reference.fas", header=TRUE,
stringsAsFactors=FALSE)
len <- nchar(reference[1,1])

bedfile <- matrix(NA, nrow=dim(variable_sites)[1]+1, ncol=3)
for_2_column <- c(0,variable_sites[,1])
bedfile[,2] <- for_2_column

for(j in 2:dim(bedfile)[1]) {
    bedfile[j-1,3] <- as.numeric(bedfile[j,2])-1
    }

bedfile[dim(bedfile)[1],3] <- len

bedfile[,1] <- "Reference.fas"

write.table(bedfile, "B.bed", sep="\t", row.names=F, col.names=F,
quote=FALSE)

q()
```

Figure V.1 (Continued)

PEPhase.R

```
rm(list=ls())

#Pulling in our list of variable sites
variable_sites <- read.table("variable_sites", header=FALSE, sep="\t",
stringsAsFactors=FALSE)

#Reading in a *.sam file that has already been piped through BEDTools
#etc.
in_table <- (read.table("Regions_of_interest.sam", header=FALSE,
sep="\t", col.names=paste0("V",seq_len(15)), fill=TRUE,
stringsAsFactors=FALSE))

#Reworking the data into a matrix so it can be manipulated
matrix_1 <-
as.matrix(in_table[1:(dim(in_table)[1]),1:dim(in_table)[2]])

dim(matrix_1) <- c(nrow=(dim(in_table)[1]),ncol=dim(in_table)[2])

matrix_1 <- matrix_1[,1:14]

rm(in_table)

# Making space for the additional variables we will be generating
num_variable_sites <- dim(variable_sites)[1]
max_rows <- dim(matrix_1)[1]
num_extra_cols <- 1+num_variable_sites+num_variable_sites

addcolumns <- matrix(NA, nrow=max_rows, ncol=num_extra_cols)

matrix_1 <- cbind(matrix_1, addcolumns)

rm(addcolumns)
rm(num_extra_cols)


#Defining variables
i <- NULL
j <- NULL
F_end <- NULL

# For loop calculating position of end of read and what variable sites
#the read covers
for (i in 1:max_rows)
    {
# First off, we are calculating the end position of our F reads (the
#beginning is in [i,4])
F_end <- as.numeric(matrix_1[i,4])+nchar(matrix_1[i,10])-1
matrix_1[i,15] <- F_end
#Now we have our F and R positions, we can work out what variable sites
#our F and R reads cover
for (j in 1:num_variable_sites) {
```

Figure V.1 (Continued)

```
          if(as.numeric(matrix_1[i,4])<=as.numeric(variable_sites[j,1]))
{

if(as.numeric(variable_sites[j,1])<=as.numeric(matrix_1[i,15])) {
                    F_count <- 1                                    } else
{F_count <- 0}
                                                              } else
{F_count <- 0}
#An F_count of 1 means our F pair covers that site, 0 that it does not

          if(F_count==0) {matrix_1[i,15+j] <- 0 }
          if(F_count==1) {matrix_1[i,15+j] <- 1 }
                                  }
    }

rm(F_count)
rm(F_end)
rm(max_rows)
rm(j)

#Reading the *.pileup in order to get flanking regions for the variable
#sites. Finding those variable sites
pileup <- read.table("Pileup", colClasses=c(rep("character",
3),rep("NULL",3)), fill=TRUE, header=FALSE, sep="\t", quote="",
stringsAsFactors=FALSE)

pileup_1 <- as.matrix(pileup[1:(dim(pileup)[1]),1:dim(pileup)[2]])

dim(pileup_1) <- c(nrow=(dim(pileup)[1]),ncol=dim(pileup)[2])

rm(pileup)
j <- NULL
var_site <- NULL
flank <- NULL
flanking_left_finish <- NULL
flanking_right_start <- NULL
flanking_left_start_20 <- NULL
flanking_right_end_20 <- NULL
flanking_left_start_10 <- NULL
flanking_right_end_10 <- NULL
flanking_left_start_5 <- NULL
flanking_right_end_5 <- NULL
left_flank_20 <- NULL
left_flank_10 <- NULL
left_flank_5 <- NULL
right_flank_20 <- NULL
right_flank_10 <- NULL
right_flank_5 <- NULL
alignmentlength <- dim(pileup_1)[1]

for (j in 1:num_variable_sites) {
     var_site <- as.numeric(variable_sites[j,1])
     flank <- as.numeric(variable_sites[j,2])
```

Figure V.1 (Continued)

```
     flanking_left_finish <- var_site - 1
     flanking_right_start <- var_site + flank
     flanking_left_start_20 <- flanking_left_finish - 20
     flanking_right_end_20 <- flanking_right_start + 20
     flanking_left_start_10 <- flanking_left_finish - 10
     flanking_right_end_10 <- flanking_right_start + 10
     flanking_left_start_5 <- flanking_left_finish - 5
     flanking_right_end_5 <- flanking_right_start + 5
     if(flanking_left_start_20<=1) {left_flank_20 <- NA} else {
     left_flank_20 <-
paste(pileup_1[flanking_left_start_20:flanking_left_finish,3],
collapse="")}
     if(flanking_left_start_10<=1) {left_flank_10 <- NA} else {
     left_flank_10 <-
paste(pileup_1[flanking_left_start_10:flanking_left_finish,3],
collapse="")}
     if(flanking_left_start_5<=1) {left_flank_5 <- NA} else {
     left_flank_5 <-
paste(pileup_1[flanking_left_start_5:flanking_left_finish,3],
collapse="")}
     if(flanking_right_end_20>=alignmentlength) {right_flank_20 <- NA}
else {
     right_flank_20 <-
paste(pileup_1[flanking_right_start:flanking_right_end_20,3],
collapse="")}
     if(flanking_right_end_10>=alignmentlength) {right_flank_10 <- NA}
else {
     right_flank_10 <-
paste(pileup_1[flanking_right_start:flanking_right_end_10,3],
collapse="")}
     if(flanking_right_end_5>=alignmentlength) {right_flank_5 <- NA}
else {
     right_flank_5 <-
paste(pileup_1[flanking_right_start:flanking_right_end_5,3],
collapse="")}
     variable_sites[j,3] <- left_flank_20
     variable_sites[j,4] <- left_flank_10
     variable_sites[j,5] <- left_flank_5
     variable_sites[j,6] <- right_flank_20
     variable_sites[j,7] <- right_flank_10
     variable_sites[j,8] <- right_flank_5
                                }
rm(pileup_1)
rm(var_site)
rm(flank)
rm(flanking_left_finish)
rm(flanking_right_start)
rm(flanking_left_start_20)
rm(flanking_right_end_20)
rm(flanking_left_start_10)
rm(flanking_right_end_10)
rm(flanking_left_start_5)
rm(flanking_right_end_5)
```

Figure V.1 (Continued)

```
rm(left_flank_20)
rm(left_flank_10)
rm(left_flank_5)
rm(right_flank_20)
rm(right_flank_10)
rm(right_flank_5)
rm(alignmentlength)
rm(i)
rm(j)

#Splitting the reads on the flanking region to get the variable sites
matrix_1 <- matrix_1[order(matrix_1[,1]),]

max_rows <- dim(matrix_1)[1]

max_variable_sites_pos <- 15+num_variable_sites

x <- NULL
y <- NULL
j <- NULL
i <- NULL
temp <- NULL


substrRight <- function(x, y){
  substr(x, nchar(x)-y+1, nchar(x))
}

for(i in 1:max_rows)
      {
for (j in 1:num_variable_sites) {
   if(as.numeric(matrix_1[i,4])<=as.numeric(variable_sites[j,1]) &&
as.numeric(variable_sites[j,1])<=as.numeric(matrix_1[i,15])) {
                if((is.na(variable_sites[j,3]))==FALSE){
                temp <- unlist(strsplit((matrix_1[i,10]),
(variable_sites[j,3])))
                matrix_1[i,max_variable_sites_pos+j] <-
substring(temp[2],1,as.numeric(variable_sites[j,2]))
                }
                if((is.na(matrix_1[i,max_variable_sites_pos+j])) &
(is.na(variable_sites[j,6]))==FALSE){
                temp <- unlist(strsplit((matrix_1[i,10]),
(variable_sites[j,6])))
                matrix_1[i,max_variable_sites_pos+j] <-
substrRight(temp[1],(as.numeric(variable_sites[j,2])))
                }
                if((is.na(matrix_1[i,max_variable_sites_pos+j])) &
(is.na(variable_sites[j,4]))==FALSE){
                temp <- unlist(strsplit((matrix_1[i,10]),
(variable_sites[j,4])))
                matrix_1[i,max_variable_sites_pos+j] <-
substring(temp[2],1,as.numeric(variable_sites[j,2]))
                }
```

Figure V.1 (Continued)

```
                if((is.na(matrix_1[i,max_variable_sites_pos+j])) &
(is.na(variable_sites[j,7]))==FALSE){
                temp <- unlist(strsplit((matrix_1[i,10]),
(variable_sites[j,7])))
                matrix_1[i,max_variable_sites_pos+j] <-
substrRight(temp[1],(as.numeric(variable_sites[j,2])))
                }
                if((is.na(matrix_1[i,max_variable_sites_pos+j])) &
(is.na(variable_sites[j,5]))==FALSE){
                temp <- unlist(strsplit((matrix_1[i,10]),
(variable_sites[j,5])))
                matrix_1[i,max_variable_sites_pos+j] <-
substring(temp[2],1,as.numeric(variable_sites[j,2]))
                }
                if((is.na(matrix_1[i,max_variable_sites_pos+j])) &
(is.na(variable_sites[j,5]))==FALSE){
                temp <- unlist(strsplit((matrix_1[i,10]),
(variable_sites[j,8])))
                matrix_1[i,max_variable_sites_pos+j] <-
substrRight(temp[1],(as.numeric(variable_sites[j,2])))
                }
              }
            }
          }

rm(x)
rm(y)
rm(i)
rm(j)
rm(temp)
rm(substrRight)

# Generating a vector of the variable site names to act as a header for
#our output table, plus first row
matrix_1 <- matrix_1[order(matrix_1[,1]),]
max_rows <- dim(matrix_1)[1]
maxrowsminus1 <- max_rows-2
var_site_threshold <- num_variable_sites-2

newrow <- NULL
to_output <- NULL
for (j in variable_sites[,1]) {to_output <- cbind(to_output, j)}
to_output <- cbind("read_name",to_output)

for(j in 1:num_variable_sites) {
newrow[1] <- matrix_1[1,1]
newrow[j+1] <- matrix_1[1,15+num_variable_sites+j]
}
if(sum(is.na(newrow))<=var_site_threshold) {
to_output <- rbind(to_output,newrow)
            }

rm(j)
```

Figure V.1 (Continued)

```
# Combining phasing over F and R reads (where available), and
#generating output file
j <- 1
newrow <- NULL

for(i in 2:maxrowsminus1)
            {
             if ((matrix_1[i-1,1]) in (matrix_1[i,1])==TRUE){
             for(j in 1:num_variable_sites) {
             if((is.na(matrix_1[i-1,15+num_variable_sites+j]))==TRUE
&&((is.na(matrix_1[i,15+num_variable_sites+j]))==TRUE)) {
             newrow[1] <- matrix_1[i,1]
             newrow[j+1] <- matrix_1[i,15+num_variable_sites+j]
             } else {
             if((is.na(matrix_1[i,15+num_variable_sites+j]))==TRUE) {
             newrow[1] <- matrix_1[i,1]
             newrow[j+1] <- matrix_1[i-1,15+num_variable_sites+j]
             } else {
             if((is.na(matrix_1[i-1,15+num_variable_sites+j]))==TRUE){
             newrow[1] <- matrix_1[i,1]
             newrow[j+1] <- matrix_1[i,15+num_variable_sites+j]
             } else
             if((is.na(matrix_1[i-1,15+num_variable_sites+j]))==FALSE
&&((is.na(matrix_1[i,15+num_variable_sites+j]))==FALSE)) {
             if((matrix_1[i-
1,15+num_variable_sites+j])==(matrix_1[i,15+num_variable_sites+j])) {
             newrow[1] <- matrix_1[i,1]
             newrow[j+1] <- matrix_1[i,15+num_variable_sites+j]
             } else {
             newrow[1] <- matrix_1[i,1]
             newrow[j+1] <- c("X")
             }
             }
             }
             }
             }
             if(sum(is.na(newrow))<=var_site_threshold) {
             to_output <- rbind(to_output,newrow)
             }
             }
             else {
             if ((matrix_1[i,1]) in (matrix_1[i+1,1])==FALSE){
             for (j in 1:num_variable_sites) {
             newrow[1] <- matrix_1[i,1]
             newrow[j+1] <- matrix_1[i,15+num_variable_sites+j]
             }
             if(sum(is.na(newrow))<=var_site_threshold) {
             to_output <- rbind(to_output,newrow)
             }
             }
           }
         }
```

Figure V.1 (Continued)

```
write.table(to_output, "phased_reads_output.txt", sep="\t")

rm(list=ls())

q()
```

Table V.2: Generation times (in years) used to calculate locus-specific N_e for cetacean species presented in Figure 5.6. All estimates for generation time obtained from Taylor et al. (2007). The estimate for the common dolphin represents an average for the species in the *Delphinus* genus.

Species	Generation time (years)
Sperm whale (*Physeter macrocephalus*)	31.9
Pygmy sperm whale (*Kogia breviceps*)	12.1
Antarctic minke whale (*Balaenoptera bonaerensis*)	22.1
Gray whale (*Eschrichtius robustus*)	22.9
Humpback whale (*Megaptera novaeangliae*)	21.5

Appendix V References:

Alexander A et al. 2013. Low diversity in the mitogenome of sperm whales revealed by next-generation sequencing. Genome Biology and Evolution. 5: 113-129.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25: 1754-1760.

Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25: 2078-2079.

Lyons LA et al. 1997. Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nature Genetics. 15: 47-56.

Meredith RW et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. Science. 334: 521-524.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26: 841-842.

R Core Team. 2013. R: A language and environment for statistical computing. URL http://www.R-project.org/.

Taylor BL, Chivers SJ, Larese J, Perrin WF. 2007. Generation length and percent mature estimates for IUCN assessments of cetaceans. Administrative Report LJ-07-01, Southwest Fisheries Science Center, 8604 La Jolla Shores Blvd., La Jolla, CA 92038, USA. 24 pp.