Demography or selection on linked cultural traits or genes? Investigating the driver of low mtDNA diversity in the sperm whale using complementary mitochondrial and nuclear genome analyses

Morin, Phillip; Foote, Andrew; Baker, C. Scott; Hancock-Hanser, Brittany; Kaschner, Kristin; Mate, Bruce; Mesnick, Sarah; Pease, Victoria; Rosel, Patricia; Alexander, Alana

Molecular Ecology

DOI:
10.1111/mec.14698

Published: 01/06/2018

Peer reviewed version

Dyfyniad o’r fersiwn a gyhoeddwyd / Citation for published version (APA):

Hawliau Cyffredinol / General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Demography or selection on linked cultural traits or genes? Investigating the driver of low mtDNA diversity in the sperm whale using complementary mitochondrial and nuclear genome analyses

Phillip A. Morin\textsuperscript{1,*}, Andrew D. Foote\textsuperscript{2}, C. Scott Baker\textsuperscript{3,4}, Brittany L. Hancock-Hanser\textsuperscript{1}, Kristin Kaschner\textsuperscript{5}, Bruce R. Mate\textsuperscript{3,4}, Sarah L. Mesnick\textsuperscript{1}, Victoria L. Pease\textsuperscript{1}, Patricia E. Rosel\textsuperscript{6}, Alana Alexander\textsuperscript{7,8}

\textsuperscript{1}Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, La Jolla, California, USA
\textsuperscript{2}Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, Gwynedd, LL57 2UW, UK
\textsuperscript{3}Marine Mammal Institute, Hatfield Marine Science Center, Oregon State University, 2030 SE Marine Science Drive, Newport, OR 97365, USA,
\textsuperscript{4}Department of Fisheries and Wildlife, College of Agricultural Sciences, 104 Nash Hall, Corvallis, OR 97330, USA
\textsuperscript{5}Department of Biometry and Environmental System Analysis, Albert-Ludwigs-University of Freiburg, Tennenbacher Strasse 4, 79106 Freiburg, Germany
\textsuperscript{6}Southeast Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Lafayette, Louisiana, USA
\textsuperscript{7}Biodiversity Institute, University of Kansas, 1345 Jayhawk Blvd, Lawrence, KS 66045, USA
\textsuperscript{8}Current address: Department of Anatomy, University of Otago, Dunedin 9016, New Zealand

*Corresponding author: phillip.morin@noaa.gov

Short title: Global phylogeography of the sperm whale

Keywords: cetacean, PSMC, population structure, Pleistocene
Abstract

Mitochondrial DNA has been heavily utilized in phylogeography studies for several decades. However, underlying patterns of demography and phylogeography may be misrepresented due to coalescence stochasticity, selection, variation in mutation rates, and cultural hitchhiking (linkage of genetic variation to culturally transmitted traits affecting fitness). Cultural hitchhiking has been suggested as an explanation for low genetic diversity in species with strong social structures, counteracting even high mobility, abundance and limited barriers to dispersal. One such species is the sperm whale, which shows very limited phylogeographic structure and low mtDNA diversity despite a worldwide distribution and large population. Here, we use analyses of 175 globally distributed mitogenomes and three nuclear genomes to evaluate hypotheses of a population bottleneck/expansion versus a selective sweep due to cultural-hitchhiking or selection on mtDNA as the mechanism contributing to low worldwide mitochondrial diversity in sperm whales. In contrast to mtDNA control region (CR) data, mitogenome haplotypes are largely ocean-specific, with only one of 80 shared between the Atlantic and Pacific. Demographic analyses of nuclear genomes suggest low mtDNA diversity is consistent with a global reduction in population size that ended approximately 125,000 years ago, correlated with the Eemian interglacial. Phylogeographic analysis suggests that extant sperm whales descend from maternal lineages endemic to the Pacific during the period of reduced abundance, and have subsequently colonized the Atlantic several times. Results highlight the apparent impact of past climate change, and suggest selection and hitchhiking are not the sole processes responsible for low mtDNA diversity in this highly social species.
**Introduction**

There is a growing body of literature indicating that the commonly held tenet of population genetics, that genetic diversity should be correlated with population size, is often violated, particularly for mitochondrial DNA (mtDNA) (Bazin et al. 2006). When mtDNA diversity is found to be low in currently abundant populations, several hypotheses have been invoked to explain the discord, including population bottlenecks (e.g., Hoelzel et al. 2002; O’Brien 1994), selection on mtDNA (e.g., Finch et al. 2014; Foote et al. 2011), variation in mutation rates (Lyrholm et al. 1996), and cultural hitchhiking (linkage of genetic variation to culturally transmitted traits affecting fitness; Kopps et al. 2014; Premo & Hublin 2009; Whitehead 1998, 2005).

Among globally distributed large whales, most baleen whales exhibit high mtDNA diversity relative to toothed whales and are divided into multiple subspecies and genetically distinct populations (e.g., fin, humpback, gray and blue whales; Archer et al. 2013; Baker et al. 2013; Jackson et al. 2014; Lang et al. 2014; Leduc et al. 2016). Among toothed whales, however, unusually low mitochondrial DNA (mtDNA) diversity in some of the social odontocetes (e.g., sperm, pilot, killer, and false-killer whales; Alexander et al. 2016; Alexander et al. 2013; Hoelzel et al. 2002; Martien et al. 2014; Van Cise et al. 2016) has limited power to infer population structure, phylogeography and historical demography using traditional genetic tools. The sperm whale (*Physeter macrocephalus*) is particularly enigmatic in this respect, as it is one of the most cosmopolitan and abundant of the large odontocetes, and known to move over large ranges of up to thousands of kilometers over annual or longer time periods (Mizroch & Rice 2013; Steiner *et al.* 2012; Straley *et al.* 2014), yet it exhibits low mtDNA diversity and evidence of female philopatry (Alexander *et al.* 2016; Lyrholm & Gyllensten 1998; Lyrholm *et al.* 1999; Mesnick *et al.* 2011).

In addition to the sperm whale’s broad distribution throughout the world’s oceans, it is considered relatively abundant even post-industrial whaling, with the global population size estimated in 2002 at 360,000, approximately 1/3 of its pre-whaling population size (Whitehead 2002). Previous genetic studies based on mtDNA control region (CR)
sequences showed haplotype frequency differences among oceans, major within-ocean geographic regions and female-led social groups, but limited phylogeographic signal, largely due to the low CR diversity and occurrence of three common haplotypes in all large ocean basins (Alexander et al. 2016; Engelhaupt et al. 2009; Lyrholm & Gyllensten 1998; Mesnick et al. 2011; Whitehead 1998).

In other species with low levels of CR diversity, complete mitochondrial genome (mitogenome) sequences have been used to detect phylogeographic structure and estimate divergence times and historical demography with higher precision (e.g., Archer et al. 2013; Buddhakosai et al. 2016; Morin et al. 2010; Morin et al. 2015; Shamblin et al. 2012). Several hypotheses have been proposed to explain the extremely low mtDNA diversity in sperm whales, including a population bottleneck (Lyrholm & Gyllensten 1998; Lyrholm et al. 1996), low mutation rate (Lyrholm et al. 1996; Whitehead 1998), stochastic variation in maternal lineage survival (Amos 1999; Tiedemann & Milinkovitch 1999), cultural hitchhiking (Whitehead 1998, 2005; Whitehead et al. 2017), and selective constraints on CR sequences that limit accumulation of variation and lead to saturation of sites free to vary (Alexander et al. 2013). Alexander et al. (2013) used complete mitogenome sequences (N=17) to evaluate support for some of these hypotheses, and concluded that the data were consistent with a bottleneck or selective sweep (involving selection on mitochondrial protein-coding regions or cultural hitchhiking), and found no evidence for selective constraint on the control region or slow substitution rates in sperm whale mtDNA relative to other cetaceans. In this study, we analyze complete mitogenomes from 175 globally distributed sperm whale samples to investigate phylogeography, selection, and historical demography. We compare demographic inferences from mitogenomes and nuclear genomes to evaluate the previously proposed hypotheses of a population bottleneck/expansion versus a selective sweep due to cultural-hitchhiking or selection on mtDNA as the mechanism contributing to low worldwide mitochondrial diversity in sperm whales.

Materials and Methods

DNA extraction and library preparation
Sperm whale tissue samples (n=158) collected by live-biopsy or from dead stranded animals were stored in salt-saturated 20% DMSO at -20°C in the US National Marine Fisheries Service (NMFS) Marine Mammal and Sea Turtle Research (MMASTR) Collection at the Southwest Fisheries Science Center (SWFSC), or in 70% ethanol at -20°C at the Oregon State University Cetacean Conservation and Genomics Laboratory (CCGL). Sample information is in supplemental Table S1. DNA from SWFSC samples was extracted from tissue samples using either a silica-membrane method (Qiaxtractor® DX reagents, Qiagen, Valencia, CA, USA) or a simple salt-precipitation procedure (Miller et al. 1988). DNA libraries for these samples were constructed and enriched for mitochondrial DNA according to Hancock-Hanser et al. (2013), and sequenced in 3 pools of 49-66 samples with Illumina GEII, HiSeq (100 bp) and NextSeq (75 bp) single-end reads. Eleven samples were repeated in two pools to increase read depth of coverage.

CCGL samples were extracted using a standard phenol/chloroform method (Sambrook et al. 1989), modified for smaller samples (Baker et al. 1994). Libraries for these samples were constructed from long-range PCR products following Alexander et al. (2013), individually-barcoded and prepared for sequencing using a Nextera XT DNA Sample Preparation Kit (Illumina, La Jolla, CA, USA). Sample libraries were pooled and sequenced in three Illumina MiSeq paired-end runs (two 250 bp, one 75 bp).

**Sequence read quality control and assembly**

Assembly of sequence reads to the reference mitogenome (KC312603) was performed using custom scripts in R (R Core Team 2014) and publicly available programs as previously described (Hancock-Hanser et al. 2013; Dryad data repository doi: 10.5061/dryad.cv35b). The reference mitogenome was modified to improve assembly coverage at the ‘ends’ of the linearized mitogenome by adding 40 bp from each end of the sequence to the opposite end (so that reads could map across the artificial break point of the linearized sequence). Nucleotide sites in the consensus sequence for each sample were called “N” if there were <3 reads, <9 reads where there was nucleotide variation at the site among reads, or >9 reads where a single nucleotide did not represent >70% of the reads. All sequences were aligned and visually inspected in the program GENEIOUS (V. 6.0.5, Biomatters, Auckland, New Zealand), with indels and unique variants identified in
GENEIOUS then verified by visual inspection of the read alignments of individual sample assemblies in the BAM files. New and 17 previously published sequences (Alexander et al. 2013) were aligned and checked for frame-shift indels and coding sequence start and stop codons within protein-coding regions, based on published annotation of the reference sperm whale mitogenome.

Genetic diversity analyses and diagnosability

The R package strataG (Archer et al. 2017a) was used to calculate haplotype and nucleotide diversity by geographically-defined population (stratum). To test if haplotypic diversity and nucleotide diversity within strata were significantly different from one another, we estimated the variance of each measure via a stratified bootstrap (supplemental methods). In each iteration of the bootstrap, individuals were randomly chosen with replacement from each stratum, with sample sizes determined by the empirical data set. At each of the 1000 bootstrap replicates, differences in nucleotide and haplotype diversity were calculated between each stratum (stratum 1 diversity - stratum 2 diversity, for each measurement). Strata were considered to have significantly different haplotype and/or nucleotide diversity if the observed difference was in the lower 5% of the distribution (strata 2 significantly greater diversity than stratum 1), or in the upper 5% of the distribution (strata 1 significantly greater than stratum 2).

Diagnosability is a measure of the ability to correctly determine whether a specimen of unknown origin can be correctly assigned to a group based on a trait or traits (Archer et al. 2017b). To test the diagnosability of mitogenome sequences between Atlantic and Pacific sperm whales, we conducted a Random Forest analysis (Breiman 2001) to create a model to classify samples to ocean basins following Archer et al. (2017b). The Random Forest was built with 50,000 trees. Each tree was built using a random draw of 25 samples (half of the smallest oceanic sample size of $n = 50$ for the Atlantic, including the Gulf of Mexico and Mediterranean Sea) from each ocean basin to avoid biasing the model due to differences in sample size. The upper and lower confidence intervals of the correct classification score as well as the expected classification score (prior) were calculated as described in Archer et al. (2017b).
**Phylogenetic analyses**

Time-calibrated phylogenetic analysis was performed using two methods: creating a prior distribution for the time to most recent ancestor (TMRCA) using a two-phase process (described below), and estimating the TMRCA using previously published substitution rates (see Demographic reconstruction section). The two-phase method utilized BEAST (v 1.8; Drummond et al. 2012). Aligned coding sequences were extracted based on published start/stop codons of the reference sequence. Optimal substitution models for individual loci were determined based on Bayesian Information Content (BIC) using PartitionFinder (v 1.1.1; Lanfear et al. 2012). All loci were constrained to having the same underlying topology and clock rate (as loci on the mtDNA are fully linked without recombination). Phase 1 estimated the TMRCA for all sperm whale mitogenomes based on the coding partitions from four sperm whale haplotypes (mtGen11, 22, 30, 47) selected to represent the four major clades identified in a maximum likelihood tree of all full mitogenome unique haplotypes (PhyML v. 2.2.0; Guindon et al. 2009, implemented in Geneious), along with aligned coding region sequences from the pygmy sperm whale (*Kogia breviceps*, accession No. NC005272) and four beaked whale species (*Berardius bairdii*, NC005274; *Ziphius cavirostris*, KC776696; *Mesoplodon densirostris*, KF032860; *Hyperodon ampulatus*, NC005273). The Bayesian phylogeny was calibrated using lognormal distributions on two calibration nodes: Odontoceti (34.67 Myr: 95% CI 29.93 – 40.17 Myr) (McGowen et al. 2009) and Ziphiidae (22 Myr: 14.86 – 32.56 Myr) (Dornburg et al. 2012; McGowen et al. 2009). The uclcd prior was set to a uniform distribution (min. 1E-6, max. 1), the tree prior was speciation, Yule process, and the chain was 100M MCMC steps, logged every 10k steps. Convergence of two replicates and mixing were checked using TRACER (v1.5; Rambaut et al. 2014) and RWTY (Warren et al. 2017a). The maximum clade credibility tree was generated with TreeAnnotator (v1.8.1) in the BEAST software cluster (Drummond et al. 2012) after removal of the first 10% of trees.

The phylogenetic analysis of all sperm whale mitogenomes (Phase II) was conducted as above, separately for all complete mitogenome unique haplotypes (N=80) and unique
haplotypes of concatenated coding loci (N=60), with TMRCA calibration based on the lognormal distribution (log(stdev = 0.2)) of the median TMRCA estimate of sperm whales from Phase I (136.7 thousand years ago (K YA), 95% CI 85.2 – 201.1 K YA) following previously described methods (Morin et al. 2010; Morin et al. 2015). A strict clock was used with a constant size coalescent tree model. 100M MCMC steps and a single mutation model (TN93), selected based on the AIC in jModeltest v2.1 (Darriba et al. 2012), were used for the full-length mitogenome sequences. For the concatenated protein coding loci, we used 10M MCMC steps and 3rd position sites only, with two mutation models applied to NADH6 and all other genes combined (HKY and TrN, respectively), based on analysis of individual loci in PartitionFinder. For both analyses, convergence was checked based on four replicate runs using TRACER v1.5 and RWTY.

A haplotype median joining network (MJN: Bandelt et al. 1999) was created using the program POPART (Leigh & Bryant 2015) with default settings.

Demographic reconstruction

We used the codon-partitioned concatenated protein-coding regions of the sperm whale mitogenomes (excluding NADH6 following Alexander et al. 2013; Ho & Lanfear 2010) in a skygrid analysis (implemented in BEAST v1.8.3; Drummond et al. 2012) and a skyline analysis (implemented in BEAST v2.3.0; Bouckaert et al. 2014). All 175 samples were included in both analyses to approximate a “balanced” sampling strategy (multiple samples from multiple populations) found to result in the least bias in inferring demographic change (Heller et al. 2013), and a separate GTR+G+I substitution model was used for each codon partition. Prior shapes for parameters within the substitution models and the relative rates for each codon position were derived from Alexander et al. (2013) (summarized in supplemental Table S2). For the skygrid analysis, two chains of 100M states, sampling every 100,000 states, were run. For the skyline analysis, two chains of 50 million states with sampling every 10,000 states were carried out. Convergence was assessed for the two analyses by comparing posterior distributions between chains in TRACER v1.6, and for topologies through RWTY.
reconstructions were generated through TRACER. Model comparison was performed in
TRACER using AICM with 1,000 bootstraps (Baele et al. 2012).

For demographic analysis based on the nuclear genome, we employed pairwise
sequentially Markovian coalescent (PSMC; Li & Durbin 2011) analysis of three sperm
whale high-coverage genomes to infer changes in effective population size \( (N_e) \). We
aligned published short read data from three sperm whales (GMX-SRS38925; BioSample
SAMN01906698 (reference genome, Gulf of Mexico); read files SRR680161,
SRR680169, SRR674482; SC991024-177-1; BioSample SAMN06187412 (Pacific); read
files SRR5136496, SRR5136506, SRR5136508, SRR5146847; BioSample
SAMN06187413 (Indian); read files SRR5136491, SRR5136493, SRR5136497,
SRR5146843), to the repeat-masked sperm whale reference genome v2.0.2 (accession
No. GCA_000472045.1; Warren et al. 2017b). Details of read quality filtering and
assembly are provided with supplemental Figure S1. The PSMC plot was scaled to an
autosomal mutation rate \( (\mu_A) \) of \( 2.9 \times 10^{-8} \) substitutions per nucleotide per generation
(Dornburg et al. 2012) and a generation time of 31.9 years (Taylor et al. 2007), and we
conducted 100 bootstrap resamplings on all PSMC analyses. Sensitivity testing for
different mutation rates and generation times were conducted by re-scaling the PSMC
plots (see supplemental Figure S1). While PSMC can provide useful insights into
demographic change inferred from changes in coalescent rates (Li & Durbin 2011), it is
also sensitive to changes in population structure (Mazet et al. 2016). Therefore, we
extended the PSMC analyses presented by Warren et al. (2017b) by also carrying out
PSMC analysis of a pseudo-diploid genome made by randomly sampling an allele at each
site from each of the individual genome assemblies using seqtk
(https://github.com/lh3/seqtk; see supplemental Figure S1 for details). The pseudo-
diploid analysis provides information on changes in the rate of coalescence between two
individual genomes through time, and therefore the timing of changes in population
structure relative to the changes in \( N_e \) inferred by the single genome PSMC analyses (see
Cahill et al. 2016; Chikhi et al. 2018).

\textit{Ancestral range reconstruction}
For biogeographic ancestral range reconstruction analyses, the full mitogenome phylogeny was pruned to twelve clades that had high posterior support. A custom R-script was then used to prune the tree topology to one branch per clade (N. Matzke; http://phylo.wikidot.com/example-biogeobears-scripts#pruning_a_tree). The terminal branches were assigned to ocean basin(s) based on the origins of samples found within the non-pruned clades, and the tree and geographic information were used for biogeographic analysis using the R package BioGeoBEARS (Matzke 2012, 2013). The package was used to compare three basic models (DEC, DIVALIKE, BAYARELIKE), and two forms of each model (with and without founder-event cladogenesis; Matzke 2014). These models allowed for different combinations of range expansion, range splitting (vicariance), and extinction, with the model conferring the highest likelihood on the tip data selected based on AIC. Likelihood ratio tests were used to compare the nested models, with and without founder-event cladogenesis. Clades were subsampled to test for the effect of uneven sampling (see methods in the legend of supplemental Figure S2).

Suitable-habitat maps were generated with the AquaMaps approach to species distribution modelling (Kaschner et al. 2016; Kaschner et al. 2011; Ready et al. 2010) based on physical and oceanographic parameters for the last glacial maximum (LGM, ~20 KYA), current, and future conditions (year 2100, based on IPCC A2 emissions scenario). Since adult male and female sperm whales have different habitat use patterns, with males migrating to higher latitudes to feed, we altered the global sperm whale suitable habitat model (males and females) to generate female-specific models by setting the minimum mean sea surface temperature (minSST, preferred minSST) to approximate the observed summer distributions of adult female and mixed family groups in the North Pacific identified from commercial whaling records (Ivashchenko et al. 2014). The adjusted mean SST does not perfectly reflect the summer SST patterns, but allows approximate habitat models for the different time periods. All models exclude the primary production envelope because this information is not available for the LGM, but comparison of the present-day distribution maps with and without primary production indicated little effect of this parameter (supplemental Figure S3 A-D, www.aquamaps.org, supported by previous research that suggested little correlation of
primary production with sperm whale distributional patterns: Jaquet & Whitehead 1996). Total sizes of distributions were subsequently calculated for both sexes in different ocean basins for all three time periods. We calculated areas using both a probability threshold of 0.0 (approximating the mean annual native range of a species including some potentially lesser utilized habitat) and ≥ 0.6 (shown to correspond to the core habitat of a species; Kaschner et al. 2011).

Selection analyses on mitogenome data

We aligned the codon-partitioned concatenated protein coding regions of newly generated sperm whale mitogenomes with the multispecies mitogenome alignment of Alexander et al. (2013), and updated this with more recently published cetacean mitogenomes/refseq versions of cetacean mitogenomes (see supplemental Table S3 for all accession numbers). Using the priors detailed in supplemental Table S4, and a Yule model of speciation, we ran two chains of 100 M generations sampling every 1,000 states in BEAST v2.3.0. After checking for convergence using TRACER and AWTY (Nylander et al. 2008; Wilgenbusch et al. 2004), we used the tree we obtained for tree-based analyses of selection using TreeSAAP v3.2 (Woolley et al. 2003) and PAML v4.9 (Yang 2007). In addition to these methods, we estimated selection for each codon using HyPhy (Pond et al. 2005) as implemented in MEGA v6.06 (Tamura et al. 2013). We further examined selection using the MEME (Murrell et al. 2012) and FUBAR methods (Murrell et al. 2013), as implemented on the Datamonkey webserver (Delport et al. 2010; Pond & Frost 2005). MEME is designed to detect sites that have experienced episodic diversification, while FUBAR is designed to detect sites that have experienced pervasive diversification. We used both methods to detect any sites that have experienced positive selection in the sperm whale. We conducted all of these analyses using just one representative sperm whale haplotype (mtGen01), and GTR models of nucleotide substitution (or REV models where GTR was not available). For investigating specific sites inferred to be under positive selection we used α = 0.05 as the threshold for statistical significance in HyPhy, MEME and FUBAR. Within PAML M8, we used Bayes Empirical Bayes (BEB) (Yang et al. 2005), with a threshold of P > 95%. TreeSAAP results were restricted to putative sites under selection where at least one
property had a magnitude category of six or more in every pairwise comparison between
the representative sperm whale and other cetaceans.

Following Caballero et al. (2015), 3D homology models of proteins for genes where
positive selection was detected in sperm whales (ND1, ND2, ATP8, COX3, ND4L, COX3,
ND4, ND5, CYTB) were generated by the SWISS-MODEL server (Schwede et al. 2003),
using the templates specified in supplemental Figure S4 (best-fitting model based out of
the top four templates for each gene region), and visualized using UCSF Chimaera v1.11
(Pettersen et al. 2004). This 3D model was annotated with domains based on alignment to
the best-fitting template. We annotated a 2D model constructed with Protter (Omasits et
al. 2014) with the domains and transmembrane topology based on alignment with the
template, as well as additional secondary structure from the 3D model of the sperm whale
constructed using DPSS (Kabsch & Sander 1983).

Results

Mitogenome dataset assembly

We generated mitogenome sequences for 158 new samples for this study. After
combining these with 17 previously published sequences from Alexander et al. (2013),
our dataset consisted of 175 globally distributed mitogenomes (Figure 1). We replicated
three sequences from Alexander et al. (2013) using our capture enrichment methods and
verified consistency between our approach and the long-range PCR approach of
Alexander et al. (2013). Mean depth of coverage was 126× (range 18-170), and all but 13
of the newly generated sequences contained ≤10 unresolved or missing nucleotides
(maximum = 65). After alignment and verification of unique indels and variant sites,
there were 80 unique haplotype sequences (Table S5).

Genetic diversity analyses and relationship of haplotypes

Mitogenome nucleotide diversity (π = 0.093%) was very similar to that reported
previously for samples predominantly obtained near New Zealand (0.096%, Alexander et
al. 2013) despite a 10-fold increase in sample size and global sample distribution. This
estimate may still be slightly inflated due to selection of some specimens to maximize
coverage of control region haplotype diversity both in this study and in Alexander et al. (2013). Sperm whales have some of the lowest documented mitogenomic diversity among cetaceans (see Table 1 in Alexander et al. 2013), but haplotype diversity was high overall (0.975) and differed significantly by ocean basin in all comparisons except “main basin” Atlantic vs. Gulf of Mexico (GoMx: both haplotype and nucleotide diversity) and Pacific vs. GoMx (nucleotide diversity) (Table 1; supplemental Table S6). Differences in diversity remained significant for all comparisons except GoMx vs. Mediterranean (haplotype diversity) even after removal of replicate haplotypes collected from the same social group to control for oversampling of close relatives (supplemental Table S6).

**Phylogenetic analyses**

The time calibrated phylogeny for TMRCA of the four divergent sperm whale haplotypes with outgroups (Phase I) is in supplemental Figure S5. The mean substitution rate estimate for all codons was 7.596E-3 substitutions/site/Myr (95% CI = 6.357E-3 – 8.860E-3), lower than that estimated previously for the sperm whale (1.04E-2) and odontocetes (1.0E-2) based on different parameters (Alexander et al. 2013). The median TMRCA estimate of 136.7 KYA (95% CI = 85.1 – 201.1 KYA) was used as the prior for analysis of the full set of sperm whale haplotypes (Phase II). The input xml, log and output trees files are available in the Dryad digital archive (doi:10.5061/dryad.57271).

The sperm whale phylogeny based on the full set of complete mitogenome haplotypes is shown in Figure 2. All equivalent sample size (ESS) values were >200 and RWTY indicated convergence of topology among separate runs. The median TMRCA estimates from the full-length mitogenomes and the 3rd positions of concatenated coding loci were nearly identical: 126.4 KYA (95% CI = 81.04 – 180.8 KYA) and 126.5 KYA (80.066 – 178.4), respectively. These estimates were older than the TMRCA estimated based on 17 mitogenomes from the Pacific alone (103 KYA (95% HPD 72.8 - 137.4); Alexander et al. 2013), likely due to the lower median substitution rate inferred in the current analysis (7.034E-3 substitutions per site per million years: 95% CI = 4.068E-3 – 1.078E-2). The input xml, log and output trees files are available in the Dryad digital archive (doi:10.5061/dryad.57271).
Full length mitogenome haplotypes tended to be separated by only a few nucleotide substitutions (Figure 3). Despite this low diversity, a high degree of phylogeographic structure was evident: 65 of the 80 haplotypes were found only in the Pacific, 14 were found only in the Atlantic, and only one haplotype (mt03) was found in both of these ocean basins (1x N. Pacific, 9x N. Atlantic). Sampling in the Indian Ocean was limited to one sample from the Maldives that had a unique haplotype (mt33) that differed by 1 bp from a haplotype found in Tasmania (mt04). One additional Indian Ocean mitogenome (from the Seychelles) assembled from the genome data of Warren et al. (2017b) was haplotyped as mt54, also found in the Tasmania. The three most common control region (CR) haplotypes globally (A, B, C based on 394bp; supplemental Table S7; Alexander et al. 2016), which are shared between the Pacific and Atlantic Oceans (and the Mediterranean Sea), constituted 67% of the samples used in this study. These three CR haplotypes were further divided into 17, 16, and 14 mitogenome haplotypes, respectively, and the four samples from the Mediterranean Sea, all CR haplotype C, were split into two unique haplotypes (supplemental Table S7). Apart from mt03 (CR haplotype A), mentioned above, all of the 48 mitogenome haplotypes corresponding to these three CR haplotypes were ocean-basin specific.

We conducted random forest analysis to determine the probability of assigning known and newly discovered mitogenome haplotypes to ocean basin. When the GoMx and Mediterranean (Med.) samples were collapsed into the Atlantic population for purposes of assignment, results indicated 100% probability of correct assignment to the Pacific, but only 78% probability for the Atlantic Ocean (Table 2). Without the GoMx and Med. samples, the assignment probability was slightly lower for the Atlantic (71%). The large number of ocean-specific haplotypes (all but one shared Pacific/Atlantic haplotype) could suggest high diagnosability, but the lower success is due to both the single shared haplotype (mt03) and high similarity of several Atlantic haplotypes to Pacific haplotypes (Figures 2, 3).

**Demographic reconstruction**
The two skygrid chains showed convergence, and all skygrid parameters had ESS values above 200. The estimated TMRCA was 87 KYA (95% HPD: 60.3-119.4). The skyline analysis also showed convergence between both chains, and all parameters had ESS values >200 after combining the chains. A TMRCA very similar to the skygrid analysis was recovered (Median: 86 KYA; 95% HPD: 59-118.8 KYA). The optimal model based on AICM was the skyline. Both analyses showed patterns consistent with a population expansion (Figure 4), but the timing appeared to differ, with the skyline suggesting the expansion began 30-35 KYA and the skygrid analysis suggesting a more gradual increase from >80 KYA (Figure 4). The 95% confidence intervals for the two methods indicate little resolution in the timing of the population expansion. Since these analyses used all samples from both ocean basin populations, assumptions of the models were violated. However, when limiting the sample set to only the Atlantic or Pacific Oceans (supplemental Figure S6), the pattern of population expansion is still recovered for the Pacific, the location of the inferred mitogenome MRCA (see ancestral reconstruction section). In contrast, the skyline plot is flat for the Atlantic, as might be expected where mitogenome clusters do not coalesce within the ocean basin. Input xml and output log and tree files are available in the Dryad digital archive (doi:10.5061/dryad.57271).

Although the nuclear genome PSMC analysis also recovered evidence of a recent population expansion, it indicated that this expansion was just one in a series of population fluctuations through time. Plots from analysis of all three ocean-basin samples declined starting about 2-3 million years ago (MYA) to a low around 1 MYA, with some fluctuation through the early and mid-Pleistocene climate cycles (Figure 4). In the late Pleistocene, the low effective population size ($N_e$) approximately 120 -150 KYA, appears to have increased to a peak $N_e$ roughly 50 KYA, then declined again to a low at about 20-30 KYA just before the last glacial maximum (Figure 4). This pattern is nearly identical between samples from the Atlantic, Pacific and Indian Oceans, differing only during the last glacial cycle, where the estimated population size estimate in the Atlantic was lower (though variation is higher, so differences are less certain; Supplemental Figure S1). These results are concordant with previous PSMC analyses of sperm whales (Warren et al. 2017b). However, the y-axis of a PSMC plot is most accurately interpreted as the
inverse of the coalescent rate, and changes in population structure can wrongly be interpreted as changes in effective population size using this method (Foote et al. 2016; Mazet et al. 2015; Mazet et al. 2016). To further tease apart changes in structure from changes in effective population size, we generated pseudo-hybrids of pairs of individuals from different populations (Cahill et al. 2016). As populations undergo a gradual fission, the rate coalescence between the two haploid genomes that comprise the pseudo diploid will decline forward in time, and should cease altogether when the two populations are completely isolated (Cahill et al. 2016). PSMC will thus infer a gradual increase in Ne as the coalescence rate decreases, and an infinite increase in effective population size at the point in time when the lineages become completely isolated.

PSMC analyses of the pseudo-diploids indicate high uncertainty at >1 MYA, but closely resemble the shared Ne of the three ocean-specific samples 0.15-1 MYA, indicative of a stable rate of coalescence between lineages (although possibly still structured, see Chikhi et al. 2018) during this period (Figure 4). During the Eemian inter-glacial, at around 120 KYA, the pseudo-diploid estimates of Ne increase rapidly and abruptly rises to infinity for all three population-pair pseudo-diploids. We interpret this as indicative of divergence between sperm whale populations associated with colonization of ocean-basins at this time. Immediately following this period, estimates of Ne fall in both the Atlantic and Pacific (and to a lesser extent, the Indian Ocean) individual PSMC plots. We interpret these combined inferences as being indicative that the population split started in the Eemian, likely with some ongoing gene flow (resulting in additional coalescence events leading to the PSMC inferring a larger Ne in all samples), and that gene flow (and hence coalescence) then ceased altogether after the Eemian, leading to the inferred decline in Ne by PSMC in the individual samples. Thus, the increase and then decline in inferred Ne for each ocean between ~125,000 and 20,000 years ago is likely to be an artifact of population differentiation, masking the true effective population sizes.

Ancestral range reconstruction and suitable habitat models
Biogeographic ancestral range reconstruction analyses with BioGeoBEARS found dispersal-extinction-cladeogenesis (DEC) as the most likely model, with the nested form
of the model allowing for founder-event cladogenesis (DEC+J) significantly more likely than the traditional DEC model (p = 0.0057). The resulting phylogeographic inference for the pruned tree indicates that the root of the tree was in the Pacific, with multiple colonizations of the Atlantic around approximately 20,000 and 60,000 years ago (Figure 5). The majority of our samples originated from the Pacific, which could potentially bias our ancestral area reconstruction (Moyle et al. 2016). We examined the effect of this discrepancy in sample size by down-sampling the Pacific to equal that of the Atlantic (see supplemental Figure S2). Across n = 1,000 replicates down-sampling the Pacific to 35 individuals, the root of the tree was inferred as Pacific in 99.9% of the replicates, indicating that uneven sampling is not driving this pattern.

Suitable habitat models for females in particular show a striking change between the present-day and the LGM, especially in the Atlantic Ocean (Figure 6), where core suitable habitat for females was reduced by 50% at the LGM (supplemental Table S8). The latitudinal shift also indicates that the Atlantic and Pacific habitats were likely completely separated by land masses, with only marginal potential for female dispersal between ocean basins even in the current warmer period. Projections of future habitat at the beginning of the next century suggest that dispersal between ocean basins may be more likely, but that northern and southern hemisphere populations may become separated (supplemental Figure S3).

Selection analyses on mitogenome data

PAML detected a pervasive pattern of purifying selection across the cetacean mitogenomes (supplemental Table S9), with the ratio of non-synonymous to synonymous changes estimated as ω = 0.093. However, all methods except FUBAR detected at least one site putatively under positive selection in the lineage leading to sperm whales (after restricting sites inferred to be under positive selection to those where all sperm whales showed a fixed amino acid substitution in comparison with the rest of the cetacean species, supplemental Table S10). In total, 19 amino acids spanning 8 mitochondrial-encoded proteins were found to show signatures of positive selection in the sperm whale (supplemental Table S10, Figure S4). Of these 19 amino acid changes, 12 of them
occurred in or adjacent to transmembrane regions, contrasting with the overall pattern of mammalian evolution of most adaptive variation being restricted to loop regions (da Fonseca et al. 2008).

Discussion

Despite the high abundance and global extent of the sperm whale’s distribution, mitogenomic diversity within this species is markedly low (this study; Alexander et al. 2013; Lyrholm & Gyllensten 1998; Whitehead 2005). However, without detailed examination of demographic patterns indicated by the nuclear genome, previous research was unable to distinguish between demographic causes and selective sweeps as the most likely cause of this low diversity. Here, we showed that both the mitogenome and nuclear genome analyses provide evidence of a population expansion and ocean-basin divergence since the last interglacial period. The presence of this pattern across both genomes is consistent with a historically small effective population size (suggested by the PSMC plots for all three ocean basins) rather than selective sweeps on the mitogenome or cultural hitchhiking of haplotypes as the primary cause of low mtDNA diversity in sperm whales, despite our detection of sites under positive selection in the sperm whale mitogenome. In contrast, the expectation proposed under the mitochondrial selective sweep hypotheses (due to either direct selection on the mitogenome or a matrilineal cultural trait, and hitchhiking of linked neutral mtDNA diversity) is discordant patterns between the nuclear and mitochondrial genomes, with a decline in mitochondrial diversity over time (Whitehead 2005) but limited or no concomitant decline in nuclear DNA diversity. This expectation is at least partially due to the high rates of sex-specific dispersal of males in sperm whales (Alexander et al. 2016), and recombination in the nuclear genome, unlinking the nuclear and mitochondrial genomes in these processes. However, it is important to note that given the strong influence of social structure on sperm whales (Whitehead et al. 2017), and the detection of positive selection on the protein-coding regions, we cannot rule out other forces, including cultural hitchhiking, having further reduced mitogenome diversity in the sperm whale. Nevertheless, the consistency of the inferred reduction in population size based on the nuclear genome occurring at approximately the same time as the TMRCA of the mitogenomes suggests
demographic processes as the primary cause of low present-day sperm whale mtDNA diversity.

Our mitogenomic analyses suggest that the current global distribution of sperm whales results from a relatively recent expansion (20-40 KYA). The inferred mitogenomic MRCA suggests expansion from a single, refugial population most likely located in the Pacific Ocean (though we cannot rule out an Indian Ocean refugium due to lack of sampling). “Ice-house” conditions are believed to have developed approximately 3 MYA, characterized by the development of perennial Arctic Ocean sea ice and continental ice sheets in North America and Eurasia (Greene et al. 2008). This global change has been implicated in a global extinction event among marine megafauna (Pimiento et al. 2017) and coincides with an inferred decline in the global sperm whale population (revealed by PSMC analyses of individual sperm whale genomes), and an evident restriction in the distribution of sperm whales (revealed by pseudo-diploid PSMC analysis of the nuclear genome, Figure 4), potentially reducing the distribution to a single refugial population from which present day sperm whale lineages descended. The global abundance appears to have increased from a low of <10,000 breeding individuals ($N_e$) since the last long cold period (Saale glaciation, ~80 kyr) that ended approximately 125 KYA. Since then, there have been several shorter cold periods, lasting <10 kyr each (Figure 4), which appear to correspond approximately with coalescence nodes in the mitogenome tree (Figure 2), after which haplotypes diversified and mitochondrial lineages dispersed between oceans. Although timing estimates are approximate, the pattern suggests effects of climate on population size, distribution, and persistence of specific maternal lineages through time in the sperm whale. While male sperm whales are known to travel to and feed in cold temperate/polar waters, females are generally restricted to warmer tropical and temperate waters, and mtDNA data suggest that they may only be able to disperse between the Pacific and Atlantic oceans during warm periods that allow them to extend their latitudinal range. However, the pseudo-diploid analysis and divergence of the nuclear genome PSMC plots from samples located in different oceans over the last ~100,000 years (Figure 4) both suggest that males also rarely disperse between ocean basins, resulting in genetic divergence between the Atlantic and Pacific (see also...
Alexander et al. 2016; Lyrholm & Gyllensten 1998; Lyrholm et al. 1999). Nuclear genome-wide SNP analysis of more samples from each ocean basin – particularly the under-sampled Indian Ocean – will be needed to infer degree of divergence and levels of current and/or historical gene flow.

The use of complete mitogenomes instead of control region sequences allows for substantially greater power to detect phylogeographic structure in low-diversity species such as the sperm whale, and can also improve power in population genetic studies. The sampling design for this study was not appropriate for traditional frequency-based population genetic analyses, but comparative analysis of mitogenome and control region haplotypes for large geographic regions (supplemental Figure S7) indicated mitogenomes were more sensitive at detecting significant differentiation among strata based on $F_{ST}$, $\Phi_{ST}$ and $\chi^2$ statistics. Although patterns of differentiation were largely concordant between these data sets, the lack of power from control region data indicates caution must be used when genetic structure among potential strata is not detected solely based on CR data.

Sperm whales are among the deepest diving marine mammals, and it is likely that their mitochondrially-encoded proteins have been under intense selection for oxygen efficiency (Janik 2001) and robustness to pressure changes (Somero 1992; Warren et al. 2017b). Changes in transmembrane regions could more strongly affect protein structure and function than in the loop regions (Saier 1994), potentially directly addressing these selective pressures. Future comparisons to patterns of positive selection on other deep-diving species would be needed to see if this was a pervasive pattern across other taxa.

Sperm whales are not considered rare, but have been depleted to approximately 1/3 of their pre-whaling population sizes and are protected by various national and international laws and treaties (e.g., the US Endangered Species Act, IUCN Red List, International Whaling Commission). Given evidence of population structure within and between ocean basins (Alexander et al. 2016; Engelhaupt et al. 2009; Lyrholm & Gyllensten 1998; Mesnick et al. 2011), the likelihood that populations were unevenly depleted by whaling...
(Ivashchenko et al. 2014), and lack of information on population recovery globally, it is highly probable that some sperm whale populations are still endangered or at risk (Carroll et al. 2014; Gero & Whitehead 2016; Notarbartolo-Di-Sciara 2014). Our results provide support for current isolation of populations located in different ocean basins, with episodic dispersal of females restricted to only warm climate periods. Although the current warming trends predict expansion of sperm whale habitat that allows both growth and opportunities for inter-ocean dispersal, we cannot predict how rapid climate change may affect the ecosystems on which sperm whales depend (see, e.g., Dawe et al. 2007; Ibanez et al. 2011; Jaquet et al. 2003; Pecl & Jackson 2008). In particular, climate change can affect the Atlantic and Pacific oceans quite differently (Boyle & Keigwin 1985; Cheng et al. 2009; Greene et al. 2008; Howard 1997), which may explain the apparent refugial population of sperm whales in the Pacific Ocean during the last long glacial period. Further research on trends in abundance and health of sperm whale populations and protection from other anthropogenic effects such as pollution (e.g., de Stephanis et al. 2013; Savery et al. 2014a; Savery et al. 2014b; Unger et al. 2016), ocean noise (Mate et al. 1994), ship strikes (Jensen & Silber 2003), entanglement (Barlow & Cameron 2003), and prey competition (Hucke-Gaete et al. 2004) should be top priorities to prevent loss of already depleted populations.

Acknowledgements

stranding network. All samples have been imported to the US under CITES permit, and samples from the US were collected under Marine Mammal Protection Act permits. We are grateful for technical and analytical assistance from The Scripps Research Institute core sequencing lab, Sebastian Duchene, Simon Ho, D. Steel, K. Hoekzema, C. Sislak and Milaya Nykänen. Cristina Garilao generously provided customized distribution maps from AquaMaps. We are very grateful for helpful discussion and comments on the manuscript provided by Hal Whitehead, Kim Parsons, Karen Chambers and four anonymous reviewers. Funding was provided by the National Marine Fisheries Service Southwest Region; the Marine Mammal and Turtle Division, Southwest Fisheries Science Center; U.S. Navy Pacific Fleet and NMFS Pacific Islands Science Center. Financial support for ADF was provided by the Welsh Government and Higher Education Funding Council for Wales through the Sêr Cymru National Research Network for Low Carbon, Energy and Environment, and from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 663830.

References


while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* 29, 2157-2167.


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


Data Accessibility

Mitogenome haplotype sequences generated for this study have been submitted to GenBank (accession numbers KU891329 - KU891394). Sequence alignment input and output xml file from Beast phylogenetic and skyline/skygrid analyses are available from the Dryad Digital Repository doi:10.5061/dryad.57271.

Author Contributions

PAM and SLM conceived of the study. CSB, BRM, SLM, HW, KK, PER and AA contributed samples, data, and expert interpretation. BLH, VLP and AA generated genetic data. PAM, AA and ADF conducted analyses and wrote the manuscript.
Figures legends

Figure 1. Map of all samples used in this study. Black circles = new mitogenome data, gray triangles = previously published mitogenome sequences (Alexander et al. 2013).

Figure 2. Time-calibrated (millions of years) phylogeny of all unique sperm whale mitochondrial genomes. Posterior probability support >0.5 is shown at nodes. Haplotype ID’s are provided as “mt##” (see supplemental Table S1 for samples associated with haplotypes) followed by abbreviated locations (see supplemental Table S1). The clades to the right correspond to the pruned tree clades used for phylogeographic analysis (Figure 5). Branches are color coded by ocean basin where haplotypes are found: Blue = Pacific, Red = Atlantic, Green = Indian, Orange = Mediterranean, Pink = Pacific/Atlantic (shared). *Mitogenome haplotypes assembled from SRA data (Warren et al. 2017b).

Figure 3. Median Joining Network of all unique haplotypes, colored by ocean. Tick marks on branches indicate the number of nucleotide differences. Not all 80 unique haplotypes are included in the network, as the algorithm collapses haplotypes with missing or ambiguous positions.

Figure 4. Demographic reconstructions based on the sperm whale mitogenomes and nuclear genome PSMC. The mitogenome skyline (black) and skygrid (gray) plot lower and upper 95% highest posterior density intervals are shown by thinner lines and median values by thicker lines. Demographic estimates were converted to effective population size (females) by dividing by a $\tau$ of 31.9 years (Taylor et al. 2007). The nuclear PSMC plots are for three sperm whales from the Atlantic (red), Pacific (blue), and Indian (yellow) Oceans. Pseudo-diploid plots are shown for each ocean pair: Atlantic-Pacific (green), Atlantic-Indian (pink), and Pacific-Indian (purple). The X axis starts at 10,000 yr before present, and the Y axis is truncated at $3.5 \times 10^4$, as pseudo-diploid plots increase exponentially and bootstrap variation at $<2 \times 10^4$ years was very large. Glacial maxima and Eemian warm period are shown with beige and yellow shading, respectively. Bootstrap plots for each PSMC analysis are presented in supplemental Figure S1.

Figure 5. Most probable ancestral ranges for sperm whale populations estimated using the best fit model DEC+J from BioGeoBEARS and full mitogenomes. Clade IDs (tips) are as in Figure 2, and represent the tree topology pruned to one tip per clade. The boxes at nodes indicate the most likely geographic range immediately after cladogenesis. Oceanic regions are indicated by P = Pacific, A = Atlantic, I = Indian, M = Mediterranean.

Figure 6. AquaMaps environmental envelope models for distribution of female sperm whale habitat mapped to environmental conditions for a) last glacial maximum and b) current habitat. Dark red color indicates core suitable habitat; yellow indicates marginal habitat.
Table 1
Number of haplotypes, haplotypic diversity and nucleotide diversity by ocean basin.

<table>
<thead>
<tr>
<th>Region</th>
<th>No. Samples</th>
<th>No. Haplotypes</th>
<th>Haplotype Diversity</th>
<th>Nucleotide Diversity (π)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic*</td>
<td>35</td>
<td>10</td>
<td>0.839</td>
<td>0.000991</td>
</tr>
<tr>
<td>GoMx</td>
<td>11</td>
<td>5</td>
<td>0.818</td>
<td>0.001019</td>
</tr>
<tr>
<td>Mediterranean</td>
<td>4</td>
<td>2</td>
<td>0.500</td>
<td>0.000000</td>
</tr>
<tr>
<td>Pacific</td>
<td>124</td>
<td>66</td>
<td>0.972</td>
<td>0.000795</td>
</tr>
<tr>
<td>Global†</td>
<td>175</td>
<td>80</td>
<td>0.975</td>
<td>0.000934</td>
</tr>
</tbody>
</table>

*excluding the Gulf of Mexico and Mediterranean Sea.
†includes single sample from the Indian Ocean.
Table 2
Random forest assignment of samples to ocean basin, percent correctly assigned, and lower (LCI) and upper (UCI) 95% confidence limits. All sample origins were known (rows), but assigned to ocean basins based on the training subset of mitogenomes.

<table>
<thead>
<tr>
<th></th>
<th>Atlantic</th>
<th>Pacific</th>
<th>pct correct</th>
<th>LCI</th>
<th>UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic</td>
<td>39</td>
<td>11</td>
<td>78</td>
<td>64.0</td>
<td>88.5</td>
</tr>
<tr>
<td>Pacific</td>
<td>0</td>
<td>124</td>
<td>100</td>
<td>97.1</td>
<td>100</td>
</tr>
</tbody>
</table>
“ice-house” conditions

LGM
Eemian warm period
Figure 6. AquaMaps environmental envelope models for distribution of female sperm whale habitat mapped to environmental conditions for a) last glacial maximum and b) current habitat. Dark red color indicates core suitable habitat; yellow indicates marginal habitat.