Oceanographic barriers, divergence, and admixture: Phylogeography and taxonomy of two putative subspecies of short-finned pilot whale

Amy M. Van Cise1,2,3 | Robin W. Baird4 | Charles Scott Baker5 | Salvatore Cerchio6 | Diane Claridge7 | Russell Fielding8 | Brittany Hancock-Hanser2 | Jacobo Marrero9,10 | Karen K. Martien2 | Antonio A. Mignucci-Giannoni11 | Erin M. Oleson12 | Marc Oremus5 | M. Michael Poole13 | Patricia E. Rosel14 | Barbara L. Taylor2 | Phillip A. Morin2

1Scripps Institution of Oceanography, La Jolla, California
2Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, La Jolla, California
3Woods Hole Oceanographic Institution, Woods Hole, Massachusetts
4Cascadia Research Collective, Olympia, Washington
5Marine Mammal Institute, Hatfield Marine Science Center, Oregon State University, Newport, Oregon
6New England Aquarium, Boston, Massachusetts
7Bahamas Marine Mammal Research Organisation, Abaco, Bahamas
8Department of Earth & Environmental Systems, University of the South, Sewanee, Tennessee
9Asociación Tonina, Investigación y Divulgación del medio natural Marino, San Cristóbal de La Laguna, Spain
10BIOECOMAC, Department of Animal Biology, La Laguna University, La Laguna, Tenerife, Canary Islands, Spain
11Centro de Conservación de Manatíes de Puerto Rico, Universidad Interamericana, Bayamón, Puerto Rico
12Pacific Islands Fisheries Science Center, National Marine Fisheries Service, NOAA, Honolulu, Hawaii
13Marine Mammal Research Program, Moorea, French Polynesia
14Southeast Fisheries Science Center, National Marine Fisheries Service, NOAA, Lafayette, Louisiana

Abstract
Genomic phylogeography plays an important role in describing evolutionary processes and their geographic, ecological, or cultural drivers. These drivers are often poorly understood in marine environments, which have fewer obvious barriers to mixing than terrestrial environments. Taxonomic uncertainty of some taxa (e.g., cetaceans), due to the difficulty in obtaining morphological data, can hamper our understanding of these processes. One such taxon, the short-finned pilot whale, is recognized as a single global species but includes at least two distinct morphological forms described from stranding and drive hunting in Japan, the “Naisa” and “Shiho” forms. Using samples (n = 735) collected throughout their global range, we examine phylogeographic patterns of divergence by comparing mitogenomes and nuclear SNP loci. Our results suggest three types within the species: an Atlantic Ocean type, a western/central Pacific and Indian Ocean (Naisa) type, and an eastern Pacific Ocean and northern Japan (Shiho) type. mtDNA control region differentiation indicates...
Genomic phylogeography is the modern continuation of classic taxonomic disciplines, and as such has an important role in the description of evolutionary processes such as isolation, selection, and speciation (Bowen et al., 2016). Marine phylogeography attempts to evaluate these processes in an environment with few obvious barriers and many widely distributed species, some with large home ranges and long migratory routes. However, meta-analyses of phylogeography across several marine taxa reveal some common, large-scale barriers between what may be considered biogeographic provinces. These include the Isthmus of Panama, separating the Pacific and Atlantic Oceans; the East Pacific Barrier, which is the large, oligotrophic, deep open ocean that limits the dispersal of many tropical species between the eastern Pacific Ocean and the central/western Pacific Ocean; the Indo-Pacific Barrier (the Indo-Malay Archipelago), separating the western Pacific and Indian Oceans; the Benguela Barrier, separating the Indian and Atlantic Oceans; and the equatorial tropics, separating temperate species in the northern and southern hemispheres (Bowen et al., 2016; Davies, 1963; Gaither, Bowen, Rocha, & Briggs, 2016; Lessios, 2008; Perrin, 2007).

These barriers often cause genetic divergence that gives rise to populations, subspecies, or species. However, taxonomic under-classification can limit our understanding of evolutionary processes such as isolation and divergence, inhibit our understanding of the ecological drivers of species evolution, and undermine conservation efforts (Bowen et al., 2016; Taylor, Archer, et al., 2017; Taylor, Perrin, et al., 2017). Taxonomic species delineation based on morphological characteristics sometimes overlooks the existence of cryptic species, subspecies, or evolutionarily important population structure (Rosel et al., 2017; Taylor, Perrin, et al., 2017). Recent advances in genomic techniques allow researchers to describe species’ taxonomy and population structure with higher resolution than was previously possible (Cammen et al., 2016), oftentimes revealing cryptic speciation in the absence of physical barriers to dispersal, driven by local adaptation and/or social behaviour (Leslie & Morin, 2016; Morin et al., 2015; Pazmiño et al., 2018; Podos, 2010; Rendell, Mesnick, Dalebout, Burtenshaw, & Whitehead, 2012; Rocha, Craig, & Bowen, 2007; Smith & Friessen, 2007; Yoshino, Armstrong, Izawa, Yokoyama, & Kawata, 2008).

This is especially true for some cetaceans, which, despite being highly mobile, often exhibit high site fidelity and adaptation to local environments (Andrews et al., 2010; Bowen et al., 2016; Foote et al., 2016; Hamner et al., 2012; Mahaffy, Baird, McSweeney, Webster, & Schorr, 2015). Others have ranges that cover entire ocean basins, yet exhibit socially-driven population structure (Balcazar et al., 2015; Carroll et al., 2015; Rendell et al., 2012; Witteveen et al., 2011). In this study, we use genetic data to understand the evolutionary phylogeography and propose taxonomic revision of a data-deficient cetacean species, the short-finned pilot whale, in order to improve our ability to understand evolutionary processes within this taxonomic unit.

Short-finned pilot whales (Globicephala macrorhynchus) are recognized as a single species with a pan-tropical and pan-temperate distribution, strong social structure (Alves et al., 2013; Mahaffy et al., 2015; Whitehead, 1998), site fidelity (Mahaffy et al., 2015), and low mtDNA diversity with widely distributed mtDNA control region haplotypes (Oremus et al., 2009; Van Cise et al., 2016). Two morphologically and genetically distinct types, originally described off Japan (Kasuya, Miyashita, & Kasamatsu, 1988; Oremus et al., 2009; Yamase, 1760), have largely nonoverlapping distributions throughout the Pacific Ocean based on samples examined to date (Van Cise et al., 2016), as well as distinct vocal repertoires in tested regions (Van Cise, Roch, Baird, Mooney, & Barlow, 2017). These two types, called “Naisa” and “Shiho” types, were originally described in 1760 based on morphological characteristics (Yamase, 1760). Their parapatric distributions around Japan remain segregated due to differing habitat preferences associated with thermally differentiated currents (Kasuya et al., 1988).

The Naisa and Shiho types differ in body size, melon and skull shape, colour pattern (specifically the brightness of the saddle patch), and number of teeth (Kasuya et al., 1988; Miyazaki & Amano, 1994; Polisini, 1980; Yonekura, Matsui, & Kasuya, 1980). Naisa-type individuals are the smaller of the two types (females 316–405 cm, males 422–525 cm) (Chivers, Perryman, Lynn, West, & Brownell, 2018), with square-shaped melons and a dark, barely visible saddle patch. Shiho-type individuals are larger by 1–2 m, with rounded melons and a bright saddle patch.

Based on morphological data and mtDNA control region sequences from outside Japan, the Shiho type has been found in the eastern Pacific Ocean from the northern to southern extent of the

**KEYWORDS**
cetacean, Globicephala macrorhynchus, phylogeography, population structure, taxonomy
short-finned pilot whale range, while the Naisa type has been found in the central/western Pacific Ocean and in the Indian Ocean (Chen et al., 2014; Chivers et al., 2018; Oremus et al., 2009; Polisini, 1980; Van Cise et al., 2016). A single skull collected from Alaska indicates that the historical range of the Shiho type may have extended between the eastern North Pacific Ocean and northern Japan. Nuclear sequences from samples collected in Hawaii, the Mariana Islands, and the eastern Tropical Pacific Ocean (ETP) suggest that the two types may be genetically distinct in their nuclear DNA, with no male-mediated gene flow between them (Van Cise, Morin, Baird, Oleson, & Martien, 2016). This evidence suggests that Naisa- and Shiho-type short-finned pilot whales may be subspecies or species, but further genetic sampling and analyses from throughout the global range of the species is needed to determine the correct taxonomic delimitation of these two types.

To date, a global taxonomic study of short-finned pilot whales has been inhibited by a lack of samples from the Indian and Atlantic Oceans. The limited data that are published suggest that the distribution of the Naisa type may extend into the Indian Ocean (Van Cise et al., 2016), while samples from the Atlantic Ocean have not yet been classified. These data are insufficient to resolve how short-finned pilot whales from these two ocean basins are related to the two types described in the Pacific Ocean. In this study we examine the global phylogeography of short-finned pilot whales, and present a formal proposal to recognize two subspecies of short-finned pilot whale.

2 | MATERIALS AND METHODS

2.1 | Sample collection and sequencing

Samples were obtained from NOAA’s SWFSC Marine Mammal and Sea Turtle Research (MMASTR) Collection (n = 268) and from other contributors and collections throughout the world (n = 53). The majority of tissue samples were collected by dart-biopsy of free-ranging whales, using an 8 mm diameter biopsy dart deployed from a crossbow. All samples were collected under permit and according to protocol to minimize disturbance to the animals. When possible, sampling was limited to one to two individuals per encounter in order to minimize the effect of related individuals on population structure. Additional samples were obtained from stranded animals. Sampling locations are shown in Figure 1, and sample details in Table S1.

DNA was extracted from tissue samples using the methods described by Martien, Chivers, et al. (2014). Genomic libraries were prepared and pooled for separate capture enrichment of mitogenome and nuclear SNP loci according to the methods described in Hancock-Hanser et al. (2013) with minor modifications. The libraries used for nuclear locus enrichment were prepared using 400 ng of DNA per sample, pooled in equimolar amounts prior to capture enrichment on capture arrays containing nuclear loci only. The nuclear SNP capture array was modified from Van Cise et al. (2017) to include 54 targeted loci, rather than the original 78 targeted loci, based on results from capture arrays used in Van Cise et al. (2017). Single-end 100 bp reads were sequenced on an Illumina HiSeq500.

2.2 | Mitogenome assembly

Mitogenome sequences were assembled using custom R scripts (R Core Team, 2016) written at SWFSC (Dryad data repository https://doi.org/10.5061/dryad.cv35b), which call on the Burrows-Wheeler alignment program (BWA; Li & Durbin, 2009). The reference sequence used to assemble mitogenome sequences (GenBank
Accession No. JF339976) was modified at the ends to include 40 bp from the opposite end, in order to improve coverage in these regions across the artificial break point in the linearized sequence.

In most cases, nucleotides were called at a locus if there were at least 10 reads and that nucleotide was called in >80% of the reads, or if a locus had at least five reads with 100% agreement in nucleotide calling. Due to the possibility of "index-hopping" during library amplification and by Illumina sequencers (Jun et al., 2012; Kircher, Sawyer, & Meyer, 2012), the R scripts were modified to include additional quality control steps. If the 80% threshold was not met, an additional filtering step was used to probabilistically call the nucleotide at that locus. First, we determined the "common" nucleotide at each locus across all samples in the data set, which was defined as the nucleotide that was represented in >50% of the samples at the locus. We also determined the "rare" nucleotide at each locus, or the nucleotides(s) represented in <50% of the samples at that locus. Next, for each locus in each sample we calculated the number of reads that matched the common nucleotide, and the number of reads that matched one or more rare nucleotides. The common nucleotide was called if the proportion of common reads at that locus in the sample (common read proportion, crp) was greater than the common read proportion at that locus across the entire sample pool (pooled common read proportion, pcrp). In other words, if crp > pcrp, the final call for that site went to the common nucleotide. Finally, rare nucleotides were called using a conservative, two-step approach that required (a) a high ratio of rare reads at that position in that sample versus the pooled data set, and (b) a high binomial probability of the rare nucleotide at that site. If the proportion of rare nucleotide reads at that locus in the sample (rare read proportion, rrp) was greater than the proportion of the rare read at that site across all samples (pooled rare read proportion, prrp) by a ratio of at least rrp = (prrp + 0.25)/1.25, then the locus in question was passed to the binomial probability test based on the frequency of each nucleotide at that site across the entire sample set. The rare read at each locus had to pass the binomial probability test with a binomial probability greater than 95%. If the rare nucleotide passed each of these tests at a given locus for a given sample, the final call for that site went to the rare nucleotide. The modified R script is included as Appendix S1. Finally, consensus sequences for each individual were aligned and visually inspected in Geneious (V. 7.1.5; Biomatters), and unique variants were verified by visual comparison with the BAM files.

2.3 | Nuclear assembly and SNP genotyping

Nuclear sequences containing 112 SNPs in 54 previously selected and quality controlled loci (Van Cise et al., 2017) were assembled as in Morin et al. (2015). Sequences were assembled using custom scripts (Dryad data repository https://doi.org/10.5061/dryad.cv35b) and reference sequences previously obtained from a draft genome of the common bottlenose dolphin (Tursiops truncatus; assembly turTru1, Jul 2008; database version 69.1) as described elsewhere (Hancock-Hansen et al., 2013; Van Cise et al., 2017). For each individual, SNP genotypes were called only if there were a minimum of 10 reads at each position, to minimize genotyping error (Fountain, Pauli, Reid, Palsbell, & Peery, 2016). SNPs within the same locus were combined into multi-SNP genotypes using PHASE (Morin et al., 2012; Stephens & Donnelly, 2003). Phasing was based on allele frequencies across all samples, with a cutoff threshold of 0.5 to minimize bias against rare heterozygotes (Garrick, Sunnucks, & Dyer, 2010), and the MCMC was run with a burn-in of 10,000, followed by 10,000 iterations, and thinned by 100 iterations.

2.4 | Data analysis: Phylogeography

The published literature uses the terms "Naisa type" and "Shiho type" to refer to two groups identified using mitochondrial and morphological data. Because multiple genetic data sets are used in this study (mitogenomes, nuclear SNPs, and control region sequences), we use specific nomenclature to orient the reader to the data set being used in each analysis. When discussing structure derived using mitogenome sequences, we refer to mitogenomic "clades". Similarly, when discussing nuclear SNP data we refer to nuclear "groups". Finally, we combine the mitogenome and nuclear SNP results to form a hypothesis of phylogeographic structure within short-finned pilot whales based on geographically defined strata, and refer to these as "strata", which we test using control region sequences. Control region sequences are used to test strata, rather than mitogenomes, in order to include a larger number of samples from the geographic range of the species, and to allow for comparison with published guidelines on taxonomic delimitation in marine mammals (Taylor, Archer, et al., 2017).

Mitogenome sequences were assigned haplotype labels using the STRATAg package (Archer, Adams, & Schneiders, 2017) implemented in the R computing environment. Tree topology was determined based on those haplotypes using a Bayesian maximum-clade-credibility (MCC) phylogenetic approach implemented in BEAST v. 1.8.4 (Drummond, Suchard, Xie, & Rambaut, 2012), rooted using a long-finned pilot whale mitogenome as the outgroup (GenBank Accession #HM060334.1). We used an HKY substitution model (Hasegawa, Kishino, & Yano, 1985) with gamma + invariant sites, which was selected using jModelTest (Darriba, Taboada, Doallo, & Posada, 2012; Guindon & Gascuel, 2003) for the complete mitochondrial genome haplotype alignment. We used an average substitution rate of 6.24 × 10^-9 substitutions/site/year, based on recent analyses of killer whales (Orcinus Orca) (Morin et al., 2015). Low mtDNA control region diversity (Van Cise et al., 2016) indicates a shallow tree; therefore, we do not expect variability in the substitution rate among branches. We therefore used a strict clock with a normal distribution and a standard deviation of 1 × 10^-7. Finally, we used a constant-size coalescent tree prior (Kingman, 1982) and 10 million MCMC steps sampled every 1,000 steps. Convergence of four replicate runs was checked using TRACER v1.6 (Ramout, Suchard, Xie, & Drummond, 2014) and AWTY (Warren, Geneva, & Lanfear, 2017). TREEANNOTATOR (v.1.8.1) in the BEAST software cluster (Drummond et al., 2012) was used to generate the maximum clade credibility tree after removal of the first 10% of trees. Additionally, a haplotype median joining
network (MJN; Bandelt, Forster, & Röhl, 1999) was generated using the program PopArt with default parameter settings (Leigh & Bryant, 2015).

We examined population structure in the nuclear genotypes using **STRUCTURE** (Pritchard, Stephens, & Donnelly, 2000), implemented in R using the **STRATAG** package. We used settings for correlated allele frequencies, 10,000 MCMC steps with a burn-in of 1,000 steps, and a k-range of one to six with five runs for each k. For each value of k, runs were combined into a single output using **CLUMPP** (Jakobsson & Rosenberg, 2007), and the optimum k value was selected by calculating the modal Δk using Evanno metrics (Evanno, Regnaut, & Goudet, 2005; Verity et al., 2016). In order to ensure that strong signals did not hide more localized population structure, additional **STRUCTURE** analyses were performed within mitogenome clades and within the geographic regions defined in the mitogenome population structure analyses below. We further used **STRATAG** to calculate the number of alleles in each of the nuclear groups when K = 2 (Naisa and Shio groups), as well as the proportion of private alleles in each. We also analyzed nuclear differentiation using a Discriminant Analysis of Principal Components (DAPC), implemented in R using the adegenet package (Jombart, 2008; Jombart & Ahmed, 2011). We first ran an unsupervised k-means DAPC to determine the number of clusters in the data by minimizing BIC. We then ran a supervised DAPC, choosing the optimum number of PCs with a 10-fold cross-validation test using a random selection of 90% of the data for training and 1,000 repetitions, to minimize root mean square error (RMSE) in classification when compared with **STRUCTURE** classifications.

### 2.5 Data analysis: Taxonomy and the subspecies hypothesis

We tested the subspecies hypothesis by calculating divergence and diagnosibility according to guidelines established by Taylor, Perrin, et al. (2017). According to these guidelines, when using mtDNA control region sequences to quantify differentiation, taxonomic categories for cetaceans (populations, subspecies, and species) are best characterized using Nei’s d_A (subspecies: >0.004, species: >0.02) and percent diagnosibility (PD) using a bootstrapped Random Forest analysis (subspecies and species: >95%). Morphological data were not available for all samples, and therefore could not be used to stratify control region sequences for hypothesis testing. Therefore, we used the independent, unsupervised clustering of SNP data into nuclear groups to stratify samples into three groups, which were then tested for divergence and diagnosibility using control region sequences. Using control region sequences for this analysis allows us to integrate a larger number of samples into this test, and to compare our results with established guidelines for taxonomic delimitation in marine mammals. This test was conducted in two steps: in the first, only samples that were grouped a priori based on **STRUCTURE** analysis of SNP data were included (n = 105); in the second, all samples from the current study and previous studies for which we had mtDNA control region sequences (Hill et al., 2015; Martien, Hill, et al., 2014; Oremus et al., 2009; Van Cise et al., 2017, 2016) were included, resulting in a total of 725 samples. The additional samples were assigned to a type based on sampling location, corresponding with the **STRUCTURE**-derived nuclear group stratification. In order to account for the potential effect of social structure on pairwise estimates of divergence and diagnosibility, we subsampled the full control region data set to include no more than three samples from each encounter with a group of pilot whales (n = 619), then re-ran these analyses and included the results in the Appendix S2.

Molecular diversity indices were calculated for all samples, and for each stratum, for mitogenomes (Theta [θ], mean nucleotide diversity [σ], haplotype diversity, and number of haplotypes) and SNP genotypes (average number of alleles per locus, expected and observed heterozygosity [H_e, H_o]), using the **STRATAG** package in R.

Pairwise differentiation was calculated to test the hypotheses that the **STRUCTURE**-derived stratification represents distinct populations, subspecies, or species, according to guidelines for subspecies delimitation presented by Taylor, Archer, et al. (2017), Archer, Martien and Taylor (2017), and Rosel et al. (2017). Because each pairwise comparison tested a unique hypothesis, corrections for multiple pairwise tests are inappropriate for this analysis and were therefore not conducted (Armstrong, 2014; Perneger, 1998). In order to compare our results with guidelines on subspecies and species (Taylor, Archer, et al., 2017), we extracted 345 bp of the mitochondrial control region sequence from all mitogenomes, and combined these with previously published control region sequences throughout the global distribution of short-finned pilot whales (Oremus et al., 2009; Van Cise et al., 2016). We estimated ΦST; net nucleotide divergence, or d_A (Nei, 1987); and PD based on a random forest classification algorithm following Archer, Martien, et al. (2017). ΦST was calculated using a Tamura–Nei model with invariant sites and without a gamma correction (Tamura & Nei, 1993), which was identified by ModelTest2 (Darriba et al., 2012) as the substitution model that best fit the data; p-values were calculated based on 1,000 permutations. d_A was calculated using p-distance without a correction factor, with pairwise deletion of sites with indels, using a bootstrap approach with 1,000 replications. We estimated the magnitude of nuclear FST among **STRUCTURE**-assigned nuclear groups using SNP genotypes, but did not test the significance of this estimate because the nuclear groups are not a priori hypotheses. Mitochondrial FST was not calculated because it has been found to be an unreliable indicator of taxonomic classification, exhibiting broad overlap in values among taxonomic classes within the order Cetartiodactyla (Rosel et al., 2017).

### 3 RESULTS

Full mitogenome sequences (16,390 bp) were successfully assembled for 181 samples. Ninety-seven unique mitogenome haplotypes were found. Sixty-two haplotypes had no unknown nucleotides, and all but four mitogenome sequences had 10 or fewer Ns; haplotypes mtGen13, mtGen80, mtGen81, mtGen89 had 21, 111, 38, and 12 Ns,
respectively. Control region sequences were extracted from these samples and aligned to previously sequenced control regions, resulting in 725 control region sequences from throughout the global distribution of short-finned pilot whales. Genotypes from 112 SNPs were generated at 47 unique loci for 245 samples (Tables S2 and S3); samples were only included if they had genotypes for at least 70% of the 112 SNPs in the data set. A total of 105 samples had both mitogenome sequences and SNP genotypes. Figure S1 shows the number of samples in each data set and overlap between data sets.

3.1 | Phylogeography

Mitochondrial nucleotide diversity and number of haplotypes was greatest within the Naisa nuclear group and the Naisa geographic stratum, while haplotype diversity was greatest within the Atlantic nuclear group and Atlantic stratum (Table 1). Expected and observed heterozygosity, and average number of alleles, were also greatest in the Naisa nuclear group, followed by the Atlantic nuclear group.

The global phylogenetic tree (Figure 2) and mitogenome haplotype MJN (Figure 3) both identify four mitogenomic clades; the previously described Shiho and Naisa types are contained within two of those clades. Mitogenome haplotype frequencies within each clade can be seen in Table S4. The Shiho clade is distinguished from the other three mitogenome clades by 15 substitutions (Figure 3). Shiho-clade short-finned pilot whales are primarily found in the eastern Pacific Ocean and northern Japan, largely separated from the Naisa clade (Figure 4a). The third Pacific clade has a geographic distribution that largely overlaps the Naisa clade in the Pacific and Indian Oceans, and is referred to as Clade 3 in this study. Clade 3 extends into the eastern Pacific, where it overlaps the distribution of the Shiho clade (Figure 4a). The fourth clade is found only in the Atlantic Ocean, and will therefore be called the Atlantic clade, although there are also three Atlantic Ocean sample haplotypes that were placed within the Naisa clade. Based on the combined BEAST analysis log files, we estimate a mean rate of 7.88 × 10⁻⁸ (95% HPD: 5.48 × 10⁻¹⁰–2.0 × 10⁻⁷) substitutions/site/year across the entire short-finned pilot whale mitogenome, and the median divergence time of the Shiho clade from all other short-finned pilot whales to be approximately 17.5 Kya (95% HPD: 3–176 Kya). Posterior support for each mitogenomic clade is shown in Figure 2; mitogenome haplotype labels and coalescent time distributions are shown in Figure S2.

Based on ΔK (Evanno et al., 2005; Verity et al., 2016), the structure analysis of nuclear SNPs indicated k = 3 as the best supported number of groups (Figure S3), and differentiated Naisa, Shiho, and Atlantic nuclear groups, corresponding closely with Naisa, Shiho, and Atlantic mitogenome clades, but did not support the differentiation of mitogenome Clade 3; all but one of the Clade 3 whales grouped with the Naisa nuclear group (Figure 5a). The unsupervised DAPC also returned an optimum group size of k = 3. Using 20 PCs (optimized by minimizing RMSE in classification), we achieved 100% classification agreement with the structure analysis (Figure 5b).

Some disparity was found between nuclear classifications and mitogenomic clades. One sample from the ETP had a Clade 3 mitogenome but clustered with the Shiho group in its nuclear DNA.

**Table 1** Molecular diversity indices for mitogenomes, mtDNA control region (345 bp), and nDNA SNP (n = 47) data sets

<table>
<thead>
<tr>
<th>Stratum</th>
<th>N</th>
<th>θ_\text{H}</th>
<th>π</th>
<th>Haplotype diversity</th>
<th>No. haplotypes</th>
<th>Average no. alleles</th>
<th>H_\text{O}</th>
<th>H_\text{E}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mitogenome sequences</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All samples</td>
<td>181</td>
<td>0.73</td>
<td>0.002</td>
<td>0.98</td>
<td>97</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Naisa clade</td>
<td>77</td>
<td>0.69</td>
<td>0.0008</td>
<td>0.92</td>
<td>35</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Shiho clade</td>
<td>43</td>
<td>0.68</td>
<td>0.0003</td>
<td>0.91</td>
<td>26</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Clade 3</td>
<td>36</td>
<td>0.71</td>
<td>0.0005</td>
<td>0.96</td>
<td>22</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Atlantic clade</td>
<td>25</td>
<td>0.69</td>
<td>0.0007</td>
<td>0.92</td>
<td>14</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td><strong>Control region sequences</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All samples</td>
<td>105</td>
<td>0.69</td>
<td>0.007</td>
<td>0.93</td>
<td>46</td>
<td>4.4</td>
<td>0.40</td>
<td>0.46</td>
</tr>
<tr>
<td>Naisa group</td>
<td>69</td>
<td>0.66</td>
<td>0.007</td>
<td>0.88</td>
<td>28</td>
<td>4.1</td>
<td>0.43</td>
<td>0.45</td>
</tr>
<tr>
<td>Shiho group</td>
<td>14</td>
<td>0.38</td>
<td>0.002</td>
<td>0.51</td>
<td>5</td>
<td>2.2</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>Atlantic group</td>
<td>22</td>
<td>0.72</td>
<td>0.003</td>
<td>0.97</td>
<td>17</td>
<td>2.7</td>
<td>0.36</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>Proposed type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All samples</td>
<td>725</td>
<td>0.62</td>
<td>0.008</td>
<td>0.83</td>
<td>64</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Naisa type</td>
<td>485</td>
<td>0.54</td>
<td>0.006</td>
<td>0.72</td>
<td>38</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Shiho type</td>
<td>190</td>
<td>0.28</td>
<td>0.002</td>
<td>0.38</td>
<td>13</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Atlantic type</td>
<td>50</td>
<td>0.60</td>
<td>0.002</td>
<td>0.081</td>
<td>17</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Note: The SNP data set includes 112 SNPs at 47 unique loci.
Abbreviations: N, mtDNA and SNP sample size; θ_\text{H}, Theta; π, nucleotide diversity; H_\text{O}, observed heterozygosity; H_\text{E}, expected heterozygosity.
Seven samples had Naisa clade mitogenomes but clustered with the Atlantic group in their nuclear DNA; of these, three were collected in the Atlantic Ocean and four were collected in the western/central Pacific and Indian Oceans. One sample, collected in the Bahamas, had an Atlantic clade mitogenome but clustered with the Naisa group in its nuclear DNA (Figure 4b). Additional STRUCTURE analyses (results not shown) did not indicate differentiation within the Shiho or Naisa nuclear groups, or mitogenome Clade 3.

Based on the concordance between mitogenomic clades and nuclear groups (Figures 2 and 4a,b), we define three distinct geographic strata within the short-finned pilot whale species (Figure 4a,b): an Atlantic Ocean stratum, a Naisa stratum (encompassing the western/central Pacific and Indian Oceans), and a Shiho stratum (occupying the eastern Pacific Ocean with a potentially relic population in northern Japan). Stratifying individual samples based on this hypothesis allows us to use geography as a proxy for genetic assignment in the absence of full mitogenome or nuclear SNP data, so that much larger data sets of mtDNA control region sequences can be used for phylogeographic and taxonomic analysis.

### 3.2 Taxonomy and the subspecies hypothesis

We estimated divergence and diagnosibility among the three geographic strata using 345 bp of the mtDNA control region, based...
on guidelines established to improve subspecies delineation using genetic data (Taylor, Archer, et al., 2017). Those guidelines suggested lower limits for two measures of mtDNA control region differentiation at the subspecies and species boundaries: $d_A$ (0.004, 0.02) and Random Forest Percent Diagnosibility, or PD (95% for both). In addition to these two metrics, we also report $\Phi_{ST}$ (Table 2), for which subspecies were found to generally fall between 0.2 and 0.6, but this measure is not recommended for use in the guidelines because it can result in overclassification (Rosel et al., 2017), therefore we did not use it in our taxonomic evaluation.

We examined pairwise net divergence ($d_A$) and PD of control region sequences for both the nuclear groups ($N = 105$) and the geographic strata ($N = 725$, Table 2). For the smaller data set based on the nuclear groups, the Naisa versus Shiho comparison met the subspecies threshold proposed by Taylor, Archer, et al. (2017) in their guidelines for both metrics ($d_A > 0.004$, PD > 95%). The Atlantic-Shiho nuclear group comparison met the threshold for PD but not $d_A$, while the Atlantic-Naisa comparison did not meet either threshold. When using the larger data set from the geographically defined strata, both the Naisa versus Shiho geographic strata comparison and the Atlantic versus Shiho geographic strata comparison met the...
subspecies thresholds for both PD and $d_A$, while the Atlantic versus Naisa geographic strata comparison met the $d_A$ threshold, but not the PD threshold. Using a subsampled data set to control for the potential effects of social structure did not appreciably change the results (Table S5).

Within the STRUCTURE-defined, $K = 2$ nuclear groups (Figure S4), the Naisa/Atlantic nuclear group had 104 private alleles (50%), while the Shiho nuclear group had seven private alleles (4%). In total, 54% of the alleles in the dataset were private to either the Naisa or Shiho nuclear group.

Although not supported by the nuclear SNP analyses, two mitogenomic clades (Naisa and Clade 3; Figures 2 and 3) were found in the western/central Pacific Ocean, with Clade 3 extending into the eastern Pacific Ocean. Using only the control region sequences, differentiation between these two mitogenome clades within the western/central Pacific/IO region met the threshold for subspecies ($d_A = 0.01$, Table 3). We additionally examined control region differentiation between the eastern Pacific Ocean and the western/central Pacific/IO within Clade 3, which was the only clade to span multiple geographic regions, and found significant $d_{ST}$ differentiation and high diagnosability (96%), but net divergence did not meet the subspecies threshold.

4 | DISCUSSION

4.1 | Phylogeography

Our results indicate that there are at least three divergent types of short-finned pilot whale throughout their global distribution (Figures 2, 4 and 5). Rather than directly conforming with ocean basins, the three types are distributed predominantly in the Atlantic Ocean (Atlantic type), western/central Pacific and Indian Oceans (Naisa type), and eastern Pacific Ocean (Shiho type). The Shiho clade diverged earliest, approximately 17.5 Kya, corresponding with the last glacial maximum (c. 18 Kya). The Atlantic clade was the next to diverge from Naisa/Clade 3. Although the timing of this split had little support in the mitogenome tree (0.30), it was further supported by nuclear DNA, which did not support a split between the Naisa mitogenome clade and Clade 3.

The distributions of these three types correlate with geographic and oceanographic boundaries that are found among multiple globally-distributed species (Figure 4). The Eastern Pacific Barrier is a known barrier to many shallow, coastally-distributed fishes, corals, and other invertebrates (Bowen et al., 2016; Chow et al., 2011; Rocha et al., 2007). Increasing evidence suggests it may also be an important barrier separating mobile trans-Pacific or globally-distributed species, such as Galapagos sharks (*Carcharhinus galapagensis*) (Pazmiño et al., 2018), tope sharks (*Galeorhinus galeus*) (Chabot, 2015), Risso’s dolphins (*Grampus griseus*) (Chen et al., 2018), and spinner dolphins (*Stenella longirostris*) (Leslie & Morin, 2018). Some evidence suggests that the barrier may be semi-permeable to some species (Lessios & Robertson, 2006), allowing occasional migration and mixing between the eastern and western Pacific Ocean. Similarly, the Eastern Pacific Barrier seems to be a semi-permeable barrier between Shiho- and Naisa-type short-finned pilot whales. The exact boundaries of their ranges in this area remain undefined due to lack of data, and it may be a region of occasional geographic overlap or temporal segregation. The
species has been found distributed throughout this oligotrophic region (Hamilton et al., 2009), though their density is lower there. The concordance across taxa of isolation and divergence along the East Pacific Barrier indicates this barrier, although not formed by a land mass, may be as important as more obvious barriers in driving evolutionary processes within marine taxa.

The Atlantic type is bordered to the west by the Isthmus of Panama, which separated the Atlantic Ocean from the eastern Pacific Ocean approximately 3.5 Mya (Lessios, 2008), long before the estimated radiation of short-finned pilot whales began. To the east, mixing between the Atlantic and Naisa types is limited by the cold Benguela Current on the southwest side of Africa.

FIGURE 5 (a) Assignment plot based on structure analyses, with $k = 3$. Mitogenome clade stratification is on the x-axis, and probability of assignment on the y-axis. (b) Supervised Discriminant Analysis of Principal Components, coloured according to the nuclear groups defined in the structure analysis in (a)
The long-finned pilot whale, the subarctic sister species to the short-finned pilot whale, inhabits the Benguela Current, while short-finned pilot whales prefer the warmer Agulhas Current, based on stranding records (Findlay, Best, Ross, & Cockcroft, 1992; van Bree, Best, & Ross, 1978). The Benguela Barrier has limited dispersal between the Atlantic and Indian Ocean species for approximately 2.5 My (Dwyer et al., 1995). Similar to the East Pacific Barrier, the Benguela Barrier is semi-permeable, preventing dispersal in temperate mobile species such as whale sharks (Castro et al., 2007), sailfish (Graves & McDowell, 1995) and blue marlins (Buonaccorsi, McDowell, & Graves, 2001), but permitting occasional migrations of tunas and other pelagic fishes (Viñas, Alvarado Bremer, & Pla, 2004). These occasional dispersal events are likely driven by southward incursions of the warm Agulhas Current from the southeast, providing a potential warm-water route for sporadic gene flow between the Atlantic Ocean and western/central Pacific and Indian Oceans (Castro et al., 2007). In most species, mixing v to be unidirectional, following the prevailing current westward from the Indian to the Atlantic Ocean (Bowen et al., 2016). Large, globally distributed whales also exhibit restricted gene flow between ocean basins, although they are not restricted to tropical or temperate waters (Baker et al., 1993).

The Indo-Pacific Barrier is a common barrier for many tropical and coastally-distributed species (Bowen et al., 2016), but our data do not show that this is a barrier to gene flow in short-finned pilot whales. It is possible, rather, that the complicated bathymetry in the region provides a rich prey base and habitat for short-finned pilot whales, which are often found along shelf breaks and slopes where they are thought to hunt deep water squid species. Similarly, we found no evidence of differentiation across the equatorial tropics.

There is evidence of limited historical or continued gene flow between the Atlantic and Naisa types, as well as the Naisa and Shihoo types (Figures 2 and 5). Our data support the migration of males and females from the Pacific or Indian Oceans to the Atlantic Ocean; it is possible that males and females migrated separately, but due to their social nature it is likely these animals migrated as social units across the Benguela Barrier. We also found evidence for male migration in the other direction, from the Atlantic Ocean to the western/central Pacific and Indian Oceans. Finally, we found evidence of female migration from the western/central Pacific and Indian Oceans to the eastern Pacific Ocean. Although these samples suggest the potential for historical or ongoing gene flow across barriers, it is important to note that there are several artifacts that might affect genetic clustering. We were able to rule out missing SNPs, high homozygosity, or large numbers of unknown mitogenome nucleotides as potential drivers of nongeographic clustering among samples. Genotype errors may also be caused by potential cross-contamination among samples sequenced in the same lane (Jun et al., 2012), miscalled genotypes, errors introduced during the phasing of genotypes into haplotypes, or errors introduced by the STRUCTURE algorithm (e.g., violation of model assumptions). Because the number of samples indicating gene flow between these types is small, we caution against drawing specific conclusions about gene flow among geographic regions without additional sampling.

Mitogenome Clade 3 was not supported in the nuclear DNA (Table 3 and Figure 5), and its distribution overlaps the Naisa clade throughout its range. In the Mariana Islands, Clade 3 and Naisa

### TABLE 2: Estimates of pairwise genetic differentiation between STRUCTURE-derived nuclear groups and geographically-defined hypothesized types

<table>
<thead>
<tr>
<th>Stratum</th>
<th>F&lt;sub&gt;ST&lt;/sub&gt;</th>
<th>Φ&lt;sub&gt;ST&lt;/sub&gt;</th>
<th>Φ&lt;sub&gt;ST&lt;/sub&gt; p-value</th>
<th>d&lt;sub&gt;A&lt;/sub&gt; 95% CI</th>
<th>PD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control region sequences</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naisa (69) versus Shihoo (22)</td>
<td>–</td>
<td>0.32</td>
<td>&lt;0.001</td>
<td>0.004</td>
<td>0.003–0.005</td>
</tr>
<tr>
<td>Atlantic (14) versus Shihoo (22)</td>
<td>–</td>
<td>0.10</td>
<td>0.04</td>
<td>0.003</td>
<td>0.002–0.003</td>
</tr>
<tr>
<td>Atlantic (14) versus Naisa (69)</td>
<td>–</td>
<td>–0.01</td>
<td>0.47</td>
<td>0.002</td>
<td>0.001–0.003</td>
</tr>
<tr>
<td><strong>Proposed type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naisa (485) versus Shihoo (190)</td>
<td>–</td>
<td>0.46</td>
<td>&lt;0.001</td>
<td>0.006</td>
<td>0.005–0.006</td>
</tr>
<tr>
<td>Atlantic (50) versus Shihoo (190)</td>
<td>–</td>
<td>0.31</td>
<td>&lt;0.001</td>
<td>0.004</td>
<td>0.003–0.004</td>
</tr>
<tr>
<td>Atlantic (50) versus Naisa (485)</td>
<td>–</td>
<td>0.15</td>
<td>&lt;0.001</td>
<td>0.005</td>
<td>0.004–0.005</td>
</tr>
<tr>
<td><strong>SNPs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naisa (69) versus Shihoo (14)</td>
<td>0.3</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Atlantic (22) versus Shihoo (14)</td>
<td>0.4</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Atlantic (22) versus Naisa (69)</td>
<td>0.1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Note: mtDNA control region differentiation was estimated using Φ<sub>ST</sub>, net divergence (d<sub>A</sub>), and percent diagnosibility (PD). The magnitude of nuclear SNP differentiation was estimated using F<sub>ST</sub>. Sample sizes for each stratum are shown in parentheses. Significant p-values are shown in bold. The lower bounds of the subspecies threshold for d<sub>A</sub> and PD are 0.02 and 95%, respectively.
TABLE 3  Estimates of pairwise genetic differentiation among mitogenomic clades within geographic regions with multiple clades, and vice versa. The magnitude of SNP differentiation was estimated using $F_{ST}$. Mitochondrial control region differentiation was estimated using $\Phi_{ST}$. Nei’s $d_A$, and PD

<table>
<thead>
<tr>
<th>Strata</th>
<th>$F_{ST}$</th>
<th>$\Phi_{ST}$</th>
<th>$\Phi_{ST}$ p-value</th>
<th>$d_A$</th>
<th>95% CI</th>
<th>PD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control region sequences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clades within geographic regions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W. Pacific/IO: Naisa (398) versus Clade 3 (62)</td>
<td>-</td>
<td>0.03</td>
<td><strong>0.020</strong></td>
<td>0.010</td>
<td>0.009–0.011</td>
<td>99.3</td>
</tr>
<tr>
<td>E. Pacific: Shiho (172) versus Clade 3 (17)</td>
<td>-</td>
<td>0.84</td>
<td>&lt;<strong>0.001</strong></td>
<td>0.006</td>
<td>0.006–0.006</td>
<td>99.5</td>
</tr>
<tr>
<td>Regions within Clades</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clade 3: E. Pacific (17) versus W. Pacific/Indian Ocean (62)</td>
<td>-</td>
<td>0.73</td>
<td>&lt;<strong>0.001</strong></td>
<td>0.002</td>
<td>0.002–0.003</td>
<td>96.0</td>
</tr>
<tr>
<td>SNPs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clades within geographic regions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W. Pacific/Indian Ocean: Naisa (52) versus Clade 3 (19)</td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Sample sizes for each stratum are shown in parentheses. Significant values are shown in bold. SNP comparisons with Clade 3 in the eastern Pacific could not be conducted due to small sample size within that region (n = 2). Abbreviations: IO, Indian Ocean; PD, percent diagnosability.

individuals have been found in the same social groups (Hill et al., 2018). Within Clade 3, the eastern Pacific Ocean regions and western/central Pacific and Indian Oceans regions were significantly differentiated, mimicking the patterns seen in the Naisa and Shiho types. These patterns could be caused by historically divergent clades with recent mixing or by lineage sorting within a widely distributed population, and may be better understood with additional nuclear and morphological data from Clade 3.

Genetic phylogeography has often been based on mitochondrial (mtDNA) control region diversity. Yet this single locus can, under certain conditions (e.g., low genetic diversity, large effective population size), misrepresent underlying patterns of isolation, divergence or speciation. Low mtDNA diversity may arise for a number of reasons, including recent population bottlenecks (Hoelzel, Fleischer, Campagna, Alvard, & Le Boeuf, 2002; Morin et al., 2018; Weber, Stewart, & Lehman, 2004), variation in mutation rates across the mitogenome (Aris-Brosou & Excoffier, 1996; Nabholz, Glémin, & Galtier, 2009; Nabholz, Glémin, Galtier, & Galtier, 2008), or selection on mtDNA (Foote et al., 2011; Finch, Zhao, Korkin, Frederick, & Eggert, 2014). In some social species, low mtDNA diversity may be caused by cultural hitchhiking, a phenomenon in which mtDNA variation changes through selection on maternally-transmitted cultural traits (Whitehead, 1998). Due to low mtDNA diversity in short-finned pilot whales, the use of additional lines of data (mitogenomes and SNPs) has improved our understanding of phylogeographic patterns and evolutionary divergence within the species.

4.2 | Taxonomy and the subspecies hypothesis

Following the guidelines for subspecies delineation summarized in Figure 3 of Taylor, Archer, et al. (2017), we find support for two subspecies: a Shiho subspecies in the eastern Pacific Ocean, and a Naisa subspecies encompassing the central/western Pacific and Indian Oceans as well as the Atlantic Ocean. We propose a nominate subspecies, G. macrorhynchus macrorhynchus, with the common name “Naisa short-finned pilot whale” (currently called “ma-gondo” in Japan), distributed throughout the central/western Pacific, Indian, and Atlantic Oceans, and an unnamed subspecies with the common name “Shiho short-finned pilot whale” (currently “tappa-naga” in Japan), found in the eastern Pacific Ocean and northern Japan (Figure S5). See the Appendix S2 for further considerations of the proposed common names. The holotype for G. macrorhynchus macrorhynchus would be that previously designated for G. macrorhynchus.

We recommend the unnamed subspecies be designated according to one of the previously synonymized names for G. macrorhynchus (see Appendix S2 for further consideration). Globicephala scammonii (Cope, 1869) is likely to be the earliest taxonomic designation for this subspecies, based on morphology and sample location, but this should be confirmed through genetic sequencing. We recommend that the holotypes for G. macrorhynchus (U.K. Natural History Museum, Accession #1846.8.9.2), G. scammonii (U.S. Natural History Museum, Accession #USNM.A.9074), and G. sieboldii (Naturalis Biodiversity Centre, Accession #RMNH.MAM.21648) be sequenced and compared with the Shiho and Naisa subspecies to resolve their taxonomic nomenclature.

There are two known regions of potential sympathy and/or introgression between the two subspecies: coastal Japan and the ETP. Off the coast of Japan, the two subspecies are spatio-temporally and ecologically isolated, with the Naisa subspecies using the warmer Kuroshio Current and the Shiho subspecies using the colder Oyashio Current (Kasuya et al., 1988). Less information is available from the ETP. Because of this, we tested mitochondrial control region differentiation between the Naisa and Shiho geographic strata in two ways—we first stratified Clade 3 samples according to geography,
to test the hypothesis of geographically-separated subspecies with some degree of admixture (Table 2). Second, we stratified the Clade 3 samples with the Naisa samples according to their mitogenomic classification, to test the hypothesis of genetically-differentiated subspecies with an area of sympatric distribution in the ETP and no genetic exchange (Table S6). Both stratification schemes support subspecies delimitation.

In addition to control region support, our analysis of nuclear SNPs independently clustered samples into Naisa and Shiho groups. The high proportion of private alleles (54%) indicates that contemporary gene flow between the two strata is very low, or possibly zero (Slatkin, 1985). Although this analysis is representative of the global range of short-finned pilot whales, there are geographic areas where sampling is scant or missing, such as the southern Atlantic Ocean, pelagic Indian Ocean, and the eastern central Pacific Ocean area of potential sympathy.

Additional data from other studies further support this recommendation. Morphological data collected off the coast of Japan show that the Shiho and Naisa types differ in skull shape, body length, and color pattern (Kasuya, 2017; Kasuya et al., 1988; Miyazaki & Amano, 1994). Skull size (length and width) may be considered a diagnostic difference between the two types (Kasuya, 2017; Miyazaki & Amano, 1994), but the limited sample size and geographic coverage outside Japan (Polisini, 1980) prohibit the use of this trait for taxonomic analysis. There is a considerably greater sample size for body length measurements, expanding the range of geographic coverage to include Hawaii and the ETP (Chivers et al., 2018). However, there is some overlap between the two types in the range of body length measurements; therefore body length cannot be considered a diagnostic trait (Cracraft, 1983; De Queiroz, 2007; Helbig, Knox, Parkin, Sangster, & Collinson, 2002; Sites & Marshall, 2004). However, length measurements do indicate a high level of concordance in geographic distribution between morphologically recognized forms and mitochondrial haplotypes (Chivers et al., 2018; Oremus et al., 2009; Polisini, 1980; Van Cise et al., 2016). Where genetic samples have been sequenced from individuals of known morphological form (Japan, Hawaii, and the eastern Pacific Ocean), concordance is 100% between the two (Oremus et al., 2009; Van Cise et al., 2016).

Finally, a few localized studies indicate the potential for additional differences between the two proposed subspecies. Where the two subspecies have allopatric distributions off the coast of Japan, there are differences in their life-history parameters, such as peak mating season (Kasuya, 2017). Similarly, a study of vocal repertoires in the eastern Pacific Shiho subspecies and the Naisa subspecies found in Hawaii indicated acoustic differentiation between the two (Van Cise et al., 2017).

A conservative taxonomic approach requires additional data, particularly from regions of sympatry, supporting complete diagnosability of the two subspecies in order to classify them as distinct species. Although we adhere to this conservative approach to species delimitation within this study, it is important to remember the risk involved in the under-classification of taxonomic units, especially with regard to conservation implications, as well as our scientific understanding of basic biological and evolutionary processes (Bowen et al., 2016; Daugherty, Cree, Hay, & Thompson, 1990; Leslie, 2015; Mace, 2004). In the case of short-finned pilot whales, and many other cetaceans, the difficulty in obtaining a sufficient data set of morphological or genetic samples covering the entire range of the species may perpetuate taxonomic under-classification, with consequences that range from failing to properly characterize the evolutionary trajectory within a specific taxon or failing to detect recent speciation events, to the extinction of under-classified species and loss of the associated evolutionary potential of that species (Allendorf & Luikart, 2011; Daugherty et al., 1990; Taylor, Perrin, et al., 2017; Wang, Frasier, Yang, & White, 2008). It is therefore our responsibility to consider, as we characterize and classify diversity, the potential for under- or over-classification of certain taxa due to logistical or biological constraints, as well as the trade-offs and consequences that may occur if our classification is not correct. In the case of Naisa and Shiho short-finned pilot whales, we recommend that priority be given to generating nuclear sequence data from areas of potential sympathy or introgression (i.e., coastal Japan and the ETP, see Figure 4), which can be used to assess gene flow and migration between the two taxa and determine whether there is support to formally elevate these two subspecies to species. Alternative methods for collecting morphological data, for example using drone photography to determine body length or melon shape, should also be explored.

Within the Naisa subspecies, two populations in the Atlantic Ocean and central/western Pacific and Indian Oceans may also be sufficiently distinct to be considered subspecies with further sampling. We suggest that delimitation of an Atlantic Ocean subspecies would require additional mitogenomic and nuclear data, or expanded morphological analyses, from the Atlantic and Indian Oceans.

ACKNOWLEDGEMENTS

We are grateful to the many people and organizations who collected and contributed samples for this study, and specifically to: Jooke Robbins, Provincetown Center for Coastal Studies; Jay Barlow, Susan Chivers, Tim Gerodette, Lisa Ballance, Robert Pitman, Louella Dolar, Marine Mammal Research Organization. Samples were collected under the following permits: NMFS MMPA permit 774-1714 issued to the SWFSC and CNMI-DFW permit, license No. 01721-10 (2010 samples); NMFS MMPA permit 14097 issued to the SWFSC and CNMI-DFW permit, license No. 02260-11 (2011 samples); NMFS MMPA permit 15240 issued to PIFSC and CNMI-DFW permit, license...
Nos. 02444-12 and 02694-13 (2012 and 2013 samples). We thank NOAA’s SWFSC Marine Mammal Genetics Group, especially Kelly Robertson, Gabriela Serra-Valente, and Vicki Pease for the curation and processing of samples used in this study. Samples provided by C.S. Baker were curated by Debbie Steel in Oregon and Rochelle Constantine in Auckland. We thank Bill Perrin and Bob Brownell for their expert advice regarding the taxonomic interpretation of results from this study. Funding for this project was provided by Commander, U.S. Pacific Fleet Environmental Readiness Division and NMFS Pacific Islands Fisheries Science Center; NMFS West Coast Region; Scripps Institution of Oceanography Edna Bailey Sussman Research Fellowship; and Woods Hole Oceanographic Institution.

AUTHOR CONTRIBUTIONS

A.M.V.C. contributed to research design, laboratory analyses, data analysis, and paper writing. P.A.M. contributed to research design and paper writing. Samples integral to the study were contributed by R.W.B., C.S.B., S.C., D.C., R.F., J.M., K.K.M., A.A.M.G., E.M.O., M.O., M.M.P., P.E.R., and B.L.T., all of whom also read and provided feedback on manuscript drafts. B.H.H. contributed to laboratory analyses and manuscript editing. Significant editorial contributions were made by C.S.B., K.K.M., P.E.R., and B.L.T.

DATA ACCESSIBILITY

All mitogenome, control region, and nuclear sequences are archived at NCBI GenBank. Mitogenome accession numbers are MK790686-MK790782. Control region accession numbers are FJ513329-FJ513341, KM624040-KM624062, DQ145032, and MH425574. Nuclear sequence accession numbers are MG023261-MG023309. SNP data are archived in the Dryad data repository: https://doi.org/10.5061/dryad.78521.

ORCID

Amy M. Van Cise https://orcid.org/0000-0002-0613-4072
Phillip A. Morin https://orcid.org/0000-0002-3279-1519

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of the article.