Fin whale (Balaenoptera physalus) mitogenomics: A cautionary tale of defining sub-species from mitochondrial sequence monophyly

Cabrera, Andrea A.; Hoekendijk, Jeroen P. A.; Aguilar, Alex; Barco, Susan G.; Berrow, Simon; Bloch, Dorete; Borrell, Asunción; Cunha, Haydée A.; Dalla Rosa, Luciano; Dias, Carolina P.

Published in:
Molecular Phylogenetics and Evolution

Link to article, DOI:
10.1016/j.ympev.2019.02.003

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Fin whale (*Balaenoptera physalus*) mitogenomics: A cautionary tale of defining sub-species from mitochondrial sequence monophyly


**Abstract**

The advent of massive parallel sequencing technologies has resulted in an increase of studies based upon complete mitochondrial genome DNA sequences that revisit the taxonomic status within and among species. Spatially distinct monophyly in such mitogenomic genealogies, i.e., the sharing of a recent common ancestor

**Keywords:** Fin whale, *Balaenoptera physalus*
among con-specific samples collected in the same region has been viewed as evidence for subspecies. Several recent studies in cetaceans have employed this criterion to suggest subsequent intraspecific taxonomic revisions. We reason that employing intra-specific, spatially distinct monophyly at non-recombining, clonally inherited genomes is an unsatisfactory criterion for defining subspecies based upon theoretical (genetic drift) and practical (sampling effort) arguments. This point was illustrated by a re-analysis of a global mitogenomic assessment of fin whales, *Balaenoptera physalus* spp., published by Archer et al. (2013), which proposed to further divide the Northern Hemisphere fin whale subspecies, *B. p. physalus*. The proposed revision was based upon the detection of spatially distinct monophyly among North Atlantic and North Pacific fin whales in a genealogy based upon complete mitochondrial genome DNA sequences. The extended analysis conducted in this study (1676 mitochondrial control region, 162 complete mitochondrial genome DNA sequences and 20 microsatellite loci genotyped in 380 samples) revealed that the apparent monophyly among North Atlantic fin whales reported by Archer et al. (2013) to be due to low sample sizes. In conclusion, defining sub-species from monophyly (i.e., the absence of para- or polyphyly) can lead to erroneous conclusions due to relatively “trivial” aspects, such as sampling. Basic population genetic processes (i.e., genetic drift and migration) also affect the time to the most recent common ancestor and hence the probability that individuals in a sample are monophyletic.

1. Introduction

Genealogies estimated from mitochondrial DNA (mtDNA) sequences have been employed towards resolving inter- and intraspecific taxonomic relationships for more than three decades (Avise, 1989; Ball and Avise, 1992; Burbright et al., 2000). Taxonomic assessments aimed below the nominal species level usually focus on the spatial distinctiveness of monophyletic clades in genealogies estimated from mtDNA sequences, i.e., the presence of phylogeographic structure (Avise et al., 1979, 1987; Ball and Avise, 1992). The presence of spatially confined monophyletic mitochondrial clades has typically been inferred as evidence for a high level of reproductive isolation and consequently some degree of evolutionary distinctiveness. Evolutionary significant units (ESUs) serve as an illustrative example (Ryder, 1986; Bernatchez, 1995). ESUs are generally viewed as distinct components of intraspecific genetic diversity (Ryder, 1986; Bernatchez, 1995). Moritz (1994) proposed that ESUs were defined by their reciprocal monophyly, specifically in genealogies estimated from mtDNA sequences. In addition, allele frequencies at nuclear loci should be “significantly” different between reciprocally monophyletic mtDNA sequence clades. When monophyly in a mtDNA genealogy is employed as the defining criterion, a key question becomes whether this level of phylogeographic structure always equates to isolation and evolutionary distinctiveness. In other words, does the absence of monophyly implies a recent common ancestry and evolutionary indistinctiveness? Paetkau (1999) pointed to the fact that the effective population size and the time to the most recent common ancestor (TMRCA) are positively correlated. This fundamental relationship implies that isolated populations at low effective population sizes will become monophyletic faster than populations with large effective population sizes. This difference has immediate ramifications in those cases when mtDNA monophyly is employed as the main, or sole, criterion in defining ESUs (e.g., Banguera-Hinestroza et al., 2002; Archer et al., 2013).

Similarly, low sampling effort may lead to erroneous conclusions when defining ESUs from spatial monophyly in mitogenomic genealogies. Apparent monophyly could be simply a product of insufficient sampling, i.e., an insufficient number of specimens to capture all mtDNA clades (Funk and Omland, 2003). Intraspecific genealogies inferred from mtDNA sequences often contain multiple well-supported clades. However, the relative proportions of such clades typically vary across space. Consequently, insufficient sampling in all, or some regions, may result in failure to sample DNA sequences belonging to uncommon clades, erroneously leading to the conclusion of monophyly (Funk and Omland, 2003).

Initially most phylogeographic studies were based solely upon genealogies inferred from mtDNA sequence variation (Ball and Avise, 1992; Burbright et al., 2000; Pons et al., 2006). The mtDNA genome (mitogenome) was viewed as especially suitable for this kind of assessments due to its haploid, often maternal and clonal inheritance, which alleviates potential issues in inferring genealogies from recombining nuclear loci. However, several studies have demonstrated that inferring intraspecific isolation from mtDNA sequences only, could be misleading, ironically because of the maternal inheritance, which prevents detection of male mediated gene flow (Prager et al., 1993; Palumbi and Baker, 1994). Consequently, many studies have since complemented mtDNA sequences with nuclear, biparentally-inherited DNA sequences in phylogeographic analyses aimed at assessing evolutionary distinctiveness, such as ESUs as proposed by Moritz (1994).

The relatively recent development of affordable massive parallel sequencing technologies (Funk et al., 2012) has led to a resurgence in phylogeographic studies based solely on mtDNA sequences, albeit of the complete mitogenome as opposed to a few hundreds of base pairs (Morin et al., 2004, 2010; Archer et al., 2013; Meng et al., 2013) remains considerably lower compared to contemporaneous studies based upon Sanger (1981) DNA sequencing of smaller mtDNA regions and nuclear loci (Pastene et al., 2007; Dalbert et al., 2013; Jackson et al., 2014). These two aspects, relying solely on mitogenome sequence data (Zachos et al., 2013) and low sample sizes, implies that the detection of monophyly is prone to the caveats that haunted earlier, similar studies based upon shorter mtDNA sequences, such as the mtDNA control region (CR). Studies based upon complete mitogenome sequences typically yield very high support for the basal nodes, leading to the impression of high accuracy. However, high accuracy in a single-locus genealogy does not necessarily imply that the genealogy accurately reflects the population/subspecies history as has been pointed out by numerous authors in the past (Pamilo and Nei, 1988; Page and Charleston, 1997; Leaché, 2009).

A case in point is *Cetacea* (whales, dolphins and porpoises), a group of highly derived mammals, which has recently been subjected to several re-assessment of species/subspecies status based upon the estimation of intraspecific genealogies from complete mitogenome sequences (Morin et al., 2010; Vilstrup et al., 2011; Archer et al., 2013). The large body sizes, wide ranges and limited availability of osteological specimens in most cetacean species has made it difficult to apply traditional, non-molecular approaches to define intra-specific taxonomic entities and explains the popularity of molecular-based taxonomic assessments in cetaceans. Most baleen whale (Mysticeti) species have global distributions and migrate seasonally between low latitude winter breeding grounds and high latitude summer feeding grounds (Ingebrigtsen, 1929; Dawbin, 1966; Jonsgård, 1966; Katona and Whitehead, 1981). As a result, most baleen whale populations roam across entire ocean basins making it challenging to delineate intra-
specific evolutionary units. Two aspects are generally assumed, \textit{a priori}, to confine baleen whale distributions and restrict gene flow. The antitropical distribution of most baleen whale species presumably acts as a reproductive barrier between the two hemispheres, despite the (proximate) low latitude locations of winter breeding grounds, because the breeding season for each hemisphere is separated by half a year (Davis et al., 1998). In addition, most ocean basins are separated by the continents, which prevent inter-oceanic dispersal as well. Consequently, it is generally assumed that gene flow between con-specific baleen whale populations in different ocean basins is very limited (Valsecchi et al., 1997; Bérubé et al., 1998; Pastene et al., 2007; Jackson et al., 2014).

Accordingly, current, recognized baleen whale species and subspecies designations typically correspond to ocean basins or hemispheres. For instance, the right whales are comprised of "Eubalaena glacialis", in the North Atlantic; "E. australis", in the Southern Hemisphere; and "E. japonica", in the North Pacific (Rice, 1998; Rosenbaum et al., 2000). Similarly Northern Hemisphere blue whales, "Balaenoptera musculus", are classified as "B. m. musculus" and Southern Hemisphere blue whales as "B. m. intermedia", in addition to the pygmy blue whale, "B. m. brevicauda" (Rice, 1998).

The fin whale, "Balaenoptera physalus" spp. (Linnaeus, 1758), is a common and globally distributed baleen whale (Gambell, 1985). Fin whales in the Northern Hemisphere are classified as belonging to the subspecies "B. p. physalus" and fin whales in the Southern Hemisphere to "B. p. quayi" (Fischer, 1829). The fin whale subspecies designations were based upon differences in the vertebrate characteristics (Lönnberg, 1931) as well as traits correlated with body size (Tomilin, 1946 cited by Rice, 1998). Employing this classification, North Pacific and North Atlantic fin whales both belong to the same subspecies, despite the observation that gene flow between the two ocean basins is unlikely, at least since the rise of the Panama Isthmus approximately 3.5 million years ago (Coates et al., 1992). Recently, Archer et al. (2013) employed complete mitogenome sequences from North Atlantic, North Pacific and Southern Hemisphere fin whale specimens to assess the current subspecies status of Northern Hemisphere fin whales. Archer et al. (2013) concluded that North Atlantic and some North Pacific fin whale specimens constituted separate subspecies. This conclusion was based upon the observation of a single monophyletic clade that contained all North Atlantic specimens (a sample of 14 specimens), and the presence of several monophyletic clades containing solely North Pacific specimens (Fig. 3a).

The results of Archer et al.’s (2013) mitogenomic analysis appeared at odds with previous phylogeographic assessments by Bérubé et al. (1998, 2002). Bérubé and colleagues based their assessments upon DNA sequences from the highly variable mtDNA CR. Their study identified two mtDNA CR haplotypes in North Atlantic specimens that clustered together with mtDNA CR haplotypes identified among North Pacific specimens. Bérubé and co-workers inferred this result as evidence for recent gene flow between the North Atlantic and North Pacific (Bérubé et al., 1998, 2002) likely in a stepping stone manner via the Southern Ocean. In order to resolve the discrepancy between the above-mentioned studies and the support for the taxonomical revision proposed by Archer et al. (2013), this study extended the sample size of North Atlantic Ocean (including the Mediterranean Sea) fin whales from the 34 mtDNA CR sequences analyzed by Archer et al. (2013) to a total of 828 mtDNA CR sequences. The complete mitogenome was sequenced in a subset (n = 6) of North Atlantic specimens, including mtDNA CR haplotypes that clustered with mtDNA CR haplotypes detected in specimens sampled outside the North Atlantic. In addition, 20 microsatellite loci were genotyped in 358 specimens from the North Atlantic, North Pacific and Southern Hemisphere. The re-estimation of the genealogy based upon the complete mitogenome sequences from this study showed all ocean basins to be polyphyletic. In other words, these results did not support the current nor the proposed subspecies if monophyly.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{sampling_location.png}
\caption{Sampling location from the North Atlantic, North Pacific and Southern Hemisphere fin whales. Notes: Approximate sampling locations in the North Atlantic (NATL, red), North Pacific (NPAC, blue) and the Southern Hemisphere (SHEM, green). The numbers represent the total number of mitochondrial control region DNA sequences from each region. Numbers in black denote the sample sizes from the Mediterranean Sea (MED, included in NATL) and the Sea of Cortez (SOC, included in NPAC). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)}
\end{figure}
in a genealogy estimated from mtDNA sequences is employed as the sole or main defining criterion. The basal topology of the genealogies estimated from the mitogenome and mtDNA CR sequences were qualitatively similar as expected given that the mitogenome represents one linked locus. The assignment test based on the genotype of 20 microsatellite loci revealed that the North Atlantic specimens from the two different mitochondrial DNA clades all belonged to the same North Atlantic gene pool.

The findings of this study highlight the implications of insufficient sampling when attempting to identify monophyletic clades from mtDNA sequences. However, the results did not negate the possibility that fin whales from different ocean basins could potentially represent different subspecies, although the analysis from this study revealed recent gene flow between fin whales from different ocean basins and hemispheres.

More generally, employing monophyly in genealogies based upon DNA sequences from non-recombining genomes to classify subspecies ignores fundamental population genetic processes as well as key practical issues. These caveats make the approach less valid than its current widespread use suggests. Although these caveats have been highlighted earlier (Paetkau, 1999; Funk and Omland, 2003), the approach has nevertheless regained momentum given the ease of applying massive parallel sequencing technologies to unparentally inherited, non-recombining genomes, such as the mitogenome.

2. Materials and methods

2.1. Sample collection

Tissue samples were obtained from fin whales in the North Atlantic Ocean basin and the Mediterranean Sea (henceforth referred to collectively as the North Atlantic); the North Pacific Ocean basin and the Sea of Cortez (henceforth referred to collectively as the North Pacific) as well as the Southern Hemisphere between 1982 and 2014 (Fig. 1). Most tissue samples were collected as skin biopsies from free-ranging fin whales as described by Palsbøll et al. (1991). The tissue samples originating from Iceland and Spain were collected during whaling operations prior to the international moratorium on commercial whaling. Some samples collected in Greenland originated from local subsistence whaling and some samples collected in US waters from dead, beached individuals. All samples were collected in agreement with national and international regulations. Samples were preserved in 5 M NaCl with 20% dimethyl sulfoxide and stored at −20 degrees Celsius (Amos and Hoelzel, 1991).

2.2. Mitochondrial DNA sequence data

The mtDNA sequence data were either generated during this study or obtained from the study by Archer et al. (2013) who deposited the data in the Dryad repository (https://doi.org/10.5061/dryad.084g8). The experimental methods used to generate the published data were described by Archer et al. (2013).

The DNA sequence data generated for this study were obtained in the following manner. Total-cell DNA was extracted from tissue samples either by phenol/chloroform extraction as described by Sambrook and Russell (2001) or using the Qiagen DNAEasy™ Blood and Tissue Kit columns (QIAGEN Inc., Valencia, CA, USA) following the manufacturer’s instructions. Samples were sexed using the ZFY/ZFX multiplexing system as described by Bérubé and Palsbøll (1996b); Bérubé and Palsbøll (1996a). MtDNA CR DNA sequencing was performed as previously described either by (i) Palsbøll et al. (1995), but replacing the original reverse primer with BP16071, (Drouot et al., 2004); or (ii) Bérubé et al. (2002). The complete mitogenome was sequenced in eight selected specimens. These specimens were selected from genealogy estimated from all mtDNA CR haplotypes. Six specimens were selected among the 33 North Atlantic specimens with mtDNA CR haplotypes that clustered with specimens sampled in other ocean basins. The remaining two specimens were selected among specimens from the North Atlantic and North Pacific Ocean, respectively, with mtDNA CR haplotypes that clustered with other specimens sampled in the same ocean within monophyletic clades. A total of 35 nested primer pairs were employed (Supplementary Materials, Table S1) to amplify and determine the DNA sequence of the complete fin whale mitogenome from partially overlapping ~500 base pair (bp) fragments. PCR (Mullis and Faloona, 1987) amplifications and DNA sequencing were performed under conditions identical to those described above for the mtDNA CR sequencing albeit at different annealing temperatures (Supplementary Materials, Table S1). Phred quality scores were employed to assess the sequence quality and discard low-quality sequences (phred score below 20). The quality of the previously published data is described by Archer et al. (2013).

2.3. Microsatellite genotyping

The genotype was determined at 20 diploid, autosomal microsatellite loci in samples from four different regions; the North Atlantic (including the Mediterranean Sea) (n = 288), the Eastern North Pacific (n = 25), the Sea of Cortez (n = 46) and the Southern Hemisphere (i.e., the Antarctic, n = 21). The specific microsatellite loci were: AC087, CA234 (Bérubé et al., 2005), EV00, EV037, EV094 (Valacchi and Amos, 1996), GATA028, GATA098, GATA417 (Palsbøll et al., 1997), GATA25072, GATA43950, GATA547654, GATA6063318, GATA91083 (Bérubé et al., in prep), GT011 (Bérubé et al., 1998), GT023, GT211, GT271, GT310, GT575, (Bérubé et al., 2000) and TAA023 (Palsbøll et al., 1997) with tetra, tri- or dimer repeat motifs (Supplementary Materials, Table S2). Individual PCR amplifications were performed in 10 μL volumes, each containing ~2–10 ng of extracted DNA, 0.2 μM of each oligo-nucleotide primer (Supplementary Materials, Table S2) and 1X final QIAGEN Microsatellite PCR Multiplex Mix™ (Qiagen Inc.). Thermo-cycling was carried out on a MJ Research PTC-100™ Thermal Cycler (BioRad Inc.). The PCR amplification consisted of an initial step of five minutes at 95 degrees Celsius, followed by 35 cycles; each of 30 s at 95 degrees Celsius, 90 s at 57 degrees Celsius and 30 s at 72 degrees Celsius. The final step was 10 min at 68 degrees Celsius. PCR reactions were diluted 60 times with MilliQ water and then 1 μl of diluted PCR reaction was added to 9 μl of GeneScan-500™ ROX (Applied Biosystems Inc.) and deionized formamide (GeneScan-500™ ROX 1 μL: 70 μL) prior to electrophoresis on an ABI 3730 Genetic Analyzer™ (Applied Biosystems Inc.). The length of each amplification product was determined using GeneMapper™ ver. 4.0 (Applied Biosystems Inc.).

2.4. Assembly and analysis of the mitochondrial DNA sequences

MtDNA sequences were aligned and assembled against the fin whale mitogenome sequence deposited in GenBank™ (accession # NC001321) by År纳斯 et al. (1991) using SEQMAN™ (ver. 5.05, DNASTAR Inc.) with default parameter settings. All DNA sequences were trimmed to equal length, i.e., 16,423 and 285 bp for the mitogenome and mtDNA CR DNA sequences, respectively.

Estimation of mtDNA haplotype genealogies and divergence times: The genealogies of the mtDNA CR and complete mitogenome haplotypes as well as divergence times, were estimated employing the software BEAST (ver. 1.8.2, Drummond and Rambaut, 2007; Drummond et al., 2012) largely following the approach by Archer et al. (2013). However, in contrast to Archer et al. (2013), only a single copy of each haplotype from each ocean basin (both complete and CR mtDNA sequences) was included in each data set. Insertion and deletions were coded as a fifth character. Genealogies were rooted with the homolog DNA sequence from humpback whale, Megaptera novaeangliae, (GenBank™ accession # NC006927, Sasaki et al. (2005)) using the alignment reported by Archer et al. (2013). The most probable nucleotide
The three independent estimations, smooth and unimodal distribution within the prior range and an effective sample size above 100,000 for all parameter estimates were also considered as indications of convergence (Supplementary Materials, Table S3).

2.5. Multi-locus genotype assignments

The likelihood of multi-locus microsatellite genotypes given the observed allele frequencies in each putative source population was estimated using the probability of identity (Paetkau and Strobeck, 1994) as implemented in GENECLASS v. 2.0 (piry et al., 2004). The null-distribution was estimated from 10,000 multi-locus genotypes drawn at random with replacement from the observed data. The type 1 error rate was set to the default value at 0.01. The observed allele frequencies in each putative source population was estimated from 266 multi-locus microsatellite genotypes for the North Atlantic (including the Mediterranean Sea), 24 for the Eastern North Pacific, 45 for the Sea of Cortez and 20 for the Southern Hemisphere (i.e., the Antarctic).

3. Results

The final data sets comprised 1676 fin whale mtDNA CR DNA sequences and 162 fin whale complete mitogenome sequences (Table 1). Among the final 1676 mtDNA CR DNA sequences, 428 DNA sequences were obtained from Archer et al. (2013) and 1248 were generated for this study. A total of 410 mtDNA CR sequences from the 782 North Pacific were collected in the Sea of Cortez, a population with low mtDNA sequence diversity (Berube et al., 2002). Among the 828 North Atlantic mtDNA CR sequences, 115 were collected in the Mediterranean Sea (Fig. 1). A total of 161 haplotypes were detected among the 1676 fin whale mtDNA CR sequences, (Table 1) and 147 haplotypes among the 162 complete mitogenome sequences of which 154 were published by Archer et al. (2013) and eight generated during this study (Table 1).

Both genealogies estimated from the complete mitogenome and mtDNA CR haplotypes published by Archer et al. (2013) identified a single monophyletic clade containing all, and only, North Atlantic specimens (denoted NA clade in Figs. 2a and 3a). In contrast, haplotypes detected in North Pacific and Southern Hemisphere specimens were polyphyletic (Figs. 2a and 3a). In contrast, the genealogies estimated from the complete mitogenome and mtDNA CR haplotypes including all data, i.e., both the data generated for this study and those published by Archer et al. (2013), partitioned the North Atlantic specimens in two major clades; one clade (denoted NA clade in Fig. 2b) comprised solely of North Atlantic specimens and another clade comprised DNA sequence haplotypes detected in specimens from the North Pacific and Southern Hemisphere, in addition to the North Atlantic

Table 1

Mitochondrial DNA control region (mtDNA CR) and complete mitochondrial DNA genome (mitogenome) sequences and haplotypes per ocean basin.

<table>
<thead>
<tr>
<th></th>
<th>North Atlantic</th>
<th>North Pacific</th>
<th>Southern Hemisphere</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEQ</td>
<td>HAP</td>
<td>SEQ</td>
<td>HAP</td>
</tr>
<tr>
<td><strong>mtDNA CR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archer et al. (2013)*</td>
<td>34</td>
<td>13</td>
<td>346</td>
<td>35</td>
</tr>
<tr>
<td>This study</td>
<td>794</td>
<td>80</td>
<td>436</td>
<td>14</td>
</tr>
<tr>
<td>Combined data</td>
<td>828</td>
<td>80</td>
<td>782</td>
<td>39</td>
</tr>
</tbody>
</table>

|                  |     |     |     |     |     |     |     |     |
| **Complete mitogenome** |     |     |     |     |     |     |     |     |
| Archer et al. (2013)* | 14  | 12  | 97  | 89  | 43  | 42  | 154 | 143 |
| This study       | 7   | 4   | 1   | 1   | 0   | 0   | 8   | 5   |
| Combined data    | 21  | 16  | 98  | 89  | 43  | 42  | 162 | 147 |

* The data was recreated from the sample information file (http://datadryad.org/bitstream/handle/10255/dryad.48318/Bphy%20sample%20info.csv?sequence = 1) deposited by Archer et al. (2013) in the Dryad data repository. The file contains GenBank accession numbers for each sample entry (either only mtDNA CR sequence (n = 274), or the mtDNA CR sequence extracted from the complete mitogenome sequence (n = 154). SEQ: number of sequences, HAP: number of unique haplotypes.

* North Atlantic including Mediterranean Sea.

** North Pacific including Sea of Cortez.
Fig. 2. Bayesian genealogy estimated from North Atlantic, North Pacific and Southern Hemisphere fin whale mitochondrial control region (mtDNA CR) haplotypes. Notes: Genealogies were estimated from (a) 82 mtDNA CR haplotypes reported by Archer et al. (2013) and (b) 161 mtDNA CR haplotypes reported by Archer et al. (2013) combined with additional mtDNA CR haplotypes reported in this study. Colors represent the three ocean basins/regions: the Southern Hemisphere (green, denoted SHEM), the North Pacific (blue, denoted NPAC) and the North Atlantic (red, denoted NATL), respectively. Numbers at basic nodes denotes the posterior probability of the specific node (only the support for basic nodes is reported). A humpback whale (Megaptera novaeangliae) mtDNA CR haplotype (Genbank NC_006927) was employed to root the tree (not shown). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Fig. 3. Bayesian genealogy estimated from North Atlantic, North Pacific and Southern Hemisphere fin whale mitochondrial genome (mitogenome) haplotypes. Notes: Genealogies were estimated from (a) 142 mitogenome haplotypes reported by Archer et al. (2013) and (b) 146 mitogenome haplotypes reported by Archer et al. (2013) combined with additional mitogenome haplotypes reported in this study. Colors represent the three ocean basins/regions: the Southern Hemisphere (green, denoted SHEM), the North Pacific (blue, denoted NPAC) and the North Atlantic Ocean (red, denoted NATL), respectively. Numbers at basic nodes denotes the posterior probability of the specific node (only the support for basic nodes is reported). A humpback whale (*Megaptera novaeangliae*) mitogenome haplotype (Genbank NC_006927) was employed to root the tree (not shown). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
(haplotypes NATL_011 and NATL_012, Fig. 2b). The genealogy estimated from the novel and previously published complete mitogenome haplotypes was similar to the genealogy inferred from the mtDNA CR sequences (Figs. 2b and 3b). In agreement with the genealogy estimated from the mtDNA CR sequences, North Atlantic specimens were partitioned into two different clades; one clade containing solely North Atlantic specimens (NA clade, Fig. 3b) and another clade containing specimens from the North Pacific, Southern Hemisphere and North Atlantic (Fig. 3b). The latter clade contained three haplotypes (haplotypes NATL_011.01, NATL_011.02 and NATL_012.01, Fig. 3b) represented by all six North Atlantic specimens from which complete mitogenome DNA sequences were obtained during this study (Fig. 3b).

The TMRCA estimated from all complete mitogenome haplotypes included this study (Fig. 3b) was estimated at 1.9 million years and the 95% HPD (highest probability density) interval from 1.1 to 2.8 million years (Table 2). The divergence time of the three North Atlantic complete mitogenome haplotype, which clustered outside the NA clade (Fig. 3b) detected during this study was estimated at 0.095 million years and the 95% HPD interval from 0.04 to 0.17 million years. The TMRCA for all the complete mitogenome haplotypes detected in the North Atlantic was estimated at 0.99 million years and the 95% HPD from 0.54 to 1.4 million years (Table 2). This estimate was 0.45 million years older than that reported by Archer et al. (2013). The TMRCA for all the mtDNA CR haplotypes included in this study (Fig. 2b) was estimated at 4.3 million years and the 95% HPD interval from 1.97 to 6.8 million years. In the case of the North Atlantic fin whales, the TMRCA was estimated at 4.2 million years and the 95% HPD interval from 1.96 to 6.8 million years.

The population origin of 25 fin whale specimens (22 sampled in the North Atlantic, one in the North Pacific, one in the Sea of Cortez and one in the Southern Hemisphere) was inferred from the assignment tests based upon diploid genotypes at 20 microsatellite loci. Among the 22 fin whale specimens from the North Atlantic, 20 samples had mtDNA CR haplotypes that were assigned to clades that also contained specimens from the North Pacific and Southern Hemisphere (i.e., outside clade “NA”, specifically haplotypes NATL_011 and NATL_012, Fig. 2b). The remaining two North Atlantic specimens had mtDNA CR haplotypes that were assigned to the main North Atlantic clade (clade “NA”, specifically haplotypes NATL_003 and NATL_016, Fig. 2b). All specimens were assigned to the population in which they were sampled with probabilities > 1.0 (Table 4).

The immigration rate (i.e., \( N_m \)) estimated from the mtDNA CR sequences from the Southern Hemisphere into the North Pacific was estimated at 0.36 (95% credible interval: 0–3.41, Table 3) which is equivalent to a single immigrant per 2.8 generations. The immigration rates from the Southern Hemisphere into the North Atlantic and from the North Atlantic into the North Pacific were estimated at 0.37 (95% credible interval: 0–2.79, Table 3) and 0–1.56, Table 3), respectively. Lastly the immigration rate from the North Pacific into the North Atlantic was estimated at 0.0029 (95% credible interval: 0–1.97, Table 3).

### 4. Discussion

The initial reason for undertaking this study was the discrepancies between Archer et al. (2013) findings and the earlier work published by Bérubé et al. (2002, 1998). However, there was a more general concern about the recent resurgence in mitogenomic-based studies employing monophyly to delineate intraspecific evolutionary distinct units.

Meiotic recombination facilitates the incorporation of population-wide variation into each haploid genome complement (Pamilo and Nei, 1988). Accordingly, population-specific monophyly at recombining loci requires substantial reproductive isolation during a considerable time. How long depends upon the effective population size (see Hudson and Turelli, 2003). The situation is different for a uniparentally inherited, non-recombining genome where each lineage contains only the variation of its own lineage rather than the population at large, which is why non-recombining loci are sensitive to the sampling effects illustrated in this study. This difference highlights the need for a sufficiently exhaustive sampling scheme. It appears that sampling was the cause for the monophyly of North Atlantic fin whales observed by Archer et al. (2013). Archer et al. (2013) included a total of 34 North Atlantic fin whale specimens (including Mediterranean Sea specimens) in their analysis, represented by 13 haplotypes. The extended sample in this study comprised 828 North Atlantic fin whale mtDNA CR sequences representing 80 haplotypes. Among the 828 sequences, 33 (i.e., ∼4%) had one of the two haplotypes that clustered outside the main North Atlantic clade (NA clade in Fig. 2b). The relative scarcity of North Atlantic fin whale mtDNA haplotypes clustering outside the North Atlantic could be the result of recent dispersal into the North Atlantic. Consequently, North Atlantic fin whales with these mtDNA haplotype may simply be immigrants and not actually part of the North Atlantic gene pool per se. However, the analyses of the biparentally inherited microsatellite loci in this study suggested that these individuals were part of the North Atlantic gene pool and unlikely to originate from the North Pacific (both the Sea of Cortez and the eastern North Pacific) nor the Southern Hemisphere. The probability of all these samples’ multi-locus genotypes was higher in the North Atlantic and Mediterranean Sea than in the other ocean basins (Table 4).

The divergence times among the basic nodes estimated during this study, (Table 2) were considerably longer than the estimates reported by Archer et al. (2013). These differences were mainly due to polyphyly detected in the expanded sample from the North Atlantic. As a result, the TMRCA among North Atlantic specimens in this study was much longer. Another consequence of a polyphyletic North Atlantic samples was that the intra-oceanic and global TMRCA were similar. The

### Table 2

<table>
<thead>
<tr>
<th>Mitochondrial control region</th>
<th>TMRCA 95% HPD</th>
<th>TMRCA 95% HPD</th>
<th>TMRCA 95% HPD</th>
<th>TMRCA 95% HPD</th>
<th>Substitution rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>North Atlantic</strong></td>
<td>Archer et al. (2013)</td>
<td>2.0 ± 3.5</td>
<td>3.3 ± 4.5</td>
<td>3.0 ± 5.0</td>
<td>3.6 ± 6.0</td>
</tr>
<tr>
<td><strong>North Pacific</strong></td>
<td>Combined data</td>
<td>4.2 ± 6.6</td>
<td>3.5 ± 6.6</td>
<td>3.3 ± 5.6</td>
<td>4.3 ± 6.6</td>
</tr>
<tr>
<td><strong>Southern Hemisphere</strong></td>
<td>Combined data</td>
<td>0.45 ± 0.68</td>
<td>1.9 ± 1.8</td>
<td>0.87 ± 1.8</td>
<td>1.9 ± 2.8</td>
</tr>
<tr>
<td><strong>All three ocean basins</strong></td>
<td>Combined data</td>
<td>0.99 ± 1.4</td>
<td>1.9 ± 1.8</td>
<td>0.87 ± 1.8</td>
<td>1.9 ± 2.8</td>
</tr>
</tbody>
</table>

Notes: TMRCA: the time to the most recent common ancestor, 95% HPD: 95% interval of the highest posterior density. Times are in million years, and the substitution rate is in substitution per site per million years.

* North Atlantic excluding Mediterranean Sea.

** North Pacific excluding Sea of Cortez.
The non-trivial nature of defining species and subspecies is a monophyly criterion as discussed above.

Recent, or low ongoing, migration between the Southern Hemisphere and the two Northern Hemisphere to two different subspecies. This taxonomic classification assigns whales from the Southern Hemisphere, i.e., Antarctic, NPA denotes the Eastern North Pacific Ocean and SOC denotes the Sea of Cortez. The n value in parenthesis represents the sample size for each putative source population. The assigned population was based on the most likely population and its relative score. CR: mtDNA control region.

A. A. Cabrera, et al.
Molecular Phylogenetics and Evolution 135 (2019) 86–97
questioned by Perrin et al. (2009), who suggested that the different latitudinal origin of the holotypes might explain the observed size difference (Perrin et al., 2009). However, this explanation is difficult to evaluate since the holotype that served as the basis for the differences in the vertebral described by Lönberg (1931) was not collected and hence is unavailable. Alternatively, if the Northern Hemisphere populations were founded from the Southern Hemisphere the observed polyphyly could be due to incomplete lineage sorting (Avise et al., 1998) as suggested by Pastene et al. (2007) in the case of common minke whales, Balanoptera acutorostrata, in the Atlantic Ocean. This appears to be the inference drawn by Archer et al. (2013), who emphasizes that the three well supported North Pacific groups (Fig. 3a) observed in their mitogenome-based genealogy could be due to incomplete lineage sorting. However, whether the patterns observed in the taxon (i.e., monophyletic, polyphyletic or paraphyletic groups) represent sub-species, as opposed to incomplete lineage sorting, population structure and/or incomplete sampling remains unclear.

Unsurprisingly no qualitative differences between the topologies inferred for the basal part of the mitochondrial genealogies were detected when increasing the data from only 285 bp of mtDNA CR sequence to the complete mitogenome DNA sequences. The general support for individual nodes, especially the most recent nodes, increased with the number of base pairs per haplotype and hence node support was substantially higher in genealogy estimated based upon complete mitogenome haplotypes. However, in most cases, the basic nodes are the target of interest in analyses of intraspecific variation aimed at assessing sub-species or ESUs. This observation, together with the obvious need for an increased sampling coverage, suggests that it might be worthwhile to first sequence a limited number of mitogenomes from the extreme parts of the species’ distribution. The mitogenome sequences can then serve as a backbone to identify and subsequently specifically target informative regions, which likely can be sequenced efficiently and at low costs using “standard” Sanger sequencing as proposed by Coulson et al. (2006). Such a strategy, as opposed to next-generation sequencing of the entire mitogenome in all specimens, would facilitate large sample sizes presumably with minimal loss of phylogenetic signal in terms the basic nodes.

In conclusion, the present study found that some of the spatially distinct mtDNA haplotype monophyly in North Atlantic fin whales reported by Archer et al. (2013) was due to an insufficient sample size. Although untested in this study, the same could be the case for some of the monophyly detected in other ocean basins or regions. Since monophyly essentially relies upon “absence of evidence” for poly or paraphyly proving monophyly, especially below the species level is difficult and prone to biases. As pointed out by Crandall et al., 2000 identifying sub-species or ESUs solely from genetic data is an over-simplification and should be complemented with ecological, behavioural and morphological data. In principle, genetic data are well-suited to assess divergence times and the degree of reproductive isolation (when gene low is low) but the choice of suitable statistics and appropriate threshold values is no simple task.

Acknowledgements

We are grateful to Hanne Jørgensen, Anna Sellas, Mary Beth Rew and Christina Færch-Jensen for technical assistance. We thank Drs. P. E. Rosel and K. D. Mullin (U.S. National Marine Fisheries Service Southeast Fisheries Science Center) and members of the U.S. Northeast and Southeast Region Marine Mammal Stranding Network and its response teams, including the International Fund for Animal Welfare, the Marine Mammal Stranding Center, Mystic Aquarium, the Riverhead Foundation for Marine Research and Preservation (K. Durham) and the Marine Mammal Stranding Program of the University of North Carolina Wilmington for access to fin whale samples from the western North Atlantic. We thank Gislín Vikingsson for providing samples. We are indebted to Dr. Eduardo Secchi for facilitating data sharing. Data collection in the Southern Ocean was conducted under research projects Baleias (CNPq grants 557064/2009-9 and 408096/2013-6), INTERBIOTA (CNPq 407889/2013-2) and INCT-APA (CNPq 574018/2008-5), of the Brazilian Antarctic Program and a contribution by the research consortium ‘Ecology and Conservation of Marine Megafauna – EcoMega-CNPq’. MAS was supported through a FCT Investigator contract funded by POPH, QREN European Social Fund, and Portuguese Ministry for Science and Education (IF/00943/2013). RP was supported by an FCT postdoctoral grant (SFRH/BPD/108007/2015). We acknowledge funds provided by FCT to MARE, through the strategic project UID/MAR/04292/2019. Data collection in the Azores was funded by TRACE-PTDC/MAR/74071/2006 and MAPCET-M2.1.2/F/012/2011 [FEDER, COMPETE, QREN European Social Fund, and Proconvergência Açores/EU Program]. AAC was supported by University of Groningen. VR-L was funded by Consejo Nacional de Ciencia y Tecnología (CONACYT). Fin whale illustration herein is used with the permission of Frédérique Lucas. We acknowledge the Center for Information Technology of the University of Groningen for IT support and access to the Peregrine high-performance-computing cluster.

Contributions to the paper

PJP, MB, JP AH and AAC conceived and designed the study. AA, SGB, SB, DB, AB, HAC, LDR, PG, SL, FL, VM, SM, NO, CP, SP, RP, CR, JR, CR, RS, MAS, JU, GV, FWW provided data or sample material, MB, JP AH, CPD, WH, VR-L conducted laboratory analyses. AAC conducted the data analysis with contributions from JP AH and inputs from ES. TO conducted the Web of Science review. AAC, JP AH, PJP and MB drafted the manuscript. All authors read, edited, commented on and approved the final manuscript.

Declaration of interest

None.

Data accessibility

Mitochondrial DNA sequence data, microsatellite genotypes and input data files for the analyses conducted in this manuscript have been deposited in Datadryad.org under accession: https://doi.org/10.5061/dryad.ql528n0.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2019.02.003.

References


