Supplementary text

Influence of past climatic change on phylogeography and demographic history of narwhals, *Monodon monoceros*

Marie Louis, Mikkel Skovrind, Jose Alfredo Samaniego Castruita, Cristina Garilao, Kristin Kaschner, Shyam Gopalakrishnan, James S. Haile, Christian Lydersen, Kit M. Kovacs, Eva Garde, Mads Peter Heide-Jørgensen, Lianne Postma, Steven H. Ferguson, Eske Willerslev, Eline D. Lorenzen

DOI: 1098/rspb.2019-2964

Materials and methods

DNA extraction, amplification and sequencing

We extracted DNA from tissue samples using the Qiagen Blood and Tissue Kit following the manufacturer's protocol with minor modifications. The volume of proteinase K was increased to 50 µL and the incubation time was extended to 24 hours. Genomic DNA was diluted to 15 ng/µL. Libraries were prepared and sequenced according to two different protocols.

For 84 samples, DNA was fragmented using the Covaris M220 Focused-ultrasonicator to create ~350-550 base pair (bp) fragment lengths. Libraries were built from the fragmented DNA extracts using Illumina NeoPrep following the NeoPrep Library Prep System Guide applying default settings. PCR amplification, quantification, and normalization were all carried out by the NeoPrep Library Prep System. The libraries were screened for size distribution on an Agilent 2100 Bioanalyzer and pooled in equimolar ratios before sequencing on an Illumina HiSeq 2500 with 80bp SE technology.

For 37 samples, the DNA was sheared to an average size of ~350-550 bp using a Bioruptor run with 4 cycles of 25 seconds on, and 90 seconds off. Further cycles were added if the fragments were too long. Paired-end sequencing libraries were built on the sheared DNA

extracts using the BEST protocol (i.e. Blunt-End Single-Tube library building for modern and ancient DNA) [1]. This method involved three steps, i) end repair with an incubation of 30 min at 20°C followed by 30 minutes at 65°C and cooling at 4°C, ii) ligation with an incubation of 30 min at 20°C followed by 30 minutes at 65°C and cooling at 4°C, iii) fill-in buffer with an incubation of 20 min at 65°C followed by 20 minutes at 80°C and cooling at 4°C. Libraries were cleaned using SPRI bead purification following Rohland and Reich (2012) [2]. Libraries were subsequently double-index amplified with unique 6bp sequences for 15 cycles using TaqGold Polymerase (5U/µl), 10X Buffer, 25 mM MgCl2, 25 Mm dNTPs, in 50 µL or 100 µL reactions. The libraries were purified using SPRI beads. The DNA concentration of the libraries was measured using a TapeStation. Libraries were pooled approximately equimolarly and sequenced on an Illumina HiSeq Xten with 150bp PE technology.

Bio-informatics

Sequencing reads were processed using Paleomix v1.2.13.1 [3]. The first step involved trimming residual adapter sequence contamination from FASTQ reads as well as low-quality stretches at read ends (i.e. consecutive stretches of N's and of bases with a quality score of 3 or lower) using AdapterRemoval/v2.2.2 [4]. Sequence reads that were ≤25 bp following trimming were discarded. Read quality was inspected using FastQC. The remaining reads were mapped to the published narwhal mitogenome reference (Genbank Accession Number: NC_005279) [5] using bwa v0.7.15 with the backtrack algorithm [6] requiring a mapping phred quality score of 30. Read groups were added using picard-tools v2.6.0 (http://broadinstitute.github.io/picard/faq.html). The same program was used to merge the bam files from each individual from different lanes and remove duplicates reads. Indel realignment was performed using GATK v4.0.4.0 [7]. Coverage was estimated using bedtools v2.26.0 [8]. Consensus sequences were built using the FastaAlternateReferenceMaker function in GATK.

Nucleotides were called "N" if there were < 3 reads, < 10 reads when there was one or more nucleotide variations, or > 10 reads where a single nucleotide did not represent > 80% of the reads. Sequences were aligned using the ClustalW algorithm in MEGA X [9] and visually inspected. Insertions and deletions were inspected manually and were masked, as it was difficult to assess whether they were real or reflected sequencing errors due to the difficulties of the Illumina sequencing technology to correctly sequence the accurate number of nucleotides, when there are poly single-nucleotide regions.

Diversity statistics - cetaceans

Data from the eight other cetacean species were compiled from Genbank, or alignments were received from the authors. For belugas, we used an unpublished dataset (Skovrind et al, unpublished data). We only considered population-scale studies, although these could represent both range-wide and local datasets (minimum sample size in the comparative data sets was 8). We re-estimarted π even if it was already available, to ensure missing data and gaps were treated consistently. Due to the high amount of missing data in some of the published datasets, we estimate nucleotide diversity excluding sites with gaps and missing data only in each pairwise comparison. This did not change the estimate for the narwhal (0.001) due to the very low number of sites with missing data. Note that our estimates may differ from the published literature, due to differences in how missing data were treated and the fact that sites with insertions and deletions were not considered in each pairwise comparison in our study. Full mitogenomes only were downloaded from Leslie et al. 2018 [10], which explains the difference in sample size for the spotted dolphins, as Leslie et al. also included partial mitogenomes (n=76 in Leslie et al. 2018 while n=70 in our analysis).

Phylogenetic analysis

Odontocete phylogeny

We included 15 species in the toothed whale phylogeny, which included species within the *Delphinidae* (n=4), *Phocoenidae* (n=3), *Monodontidae* (n=2), *Ziphiidae* (n=4) and *Physeteroidae* (*Physeteridae* + *Kogiidae*) (n=2, supplementary table S2). The four river dolphin families were excluded due to their paraphyletic topology [11] and fast clock rates [12].

Four fossil calibrations were used to calibrate the toothed whale tree:

(i) *Ferecetotherium kelloggi* [13] was used to set the minimum age of crown Odontoceti to 23 Mya. This calibration point has been used by Dornburg et al. 2012 [12] and Galatius et al. 2018 [14]. The calibration was applied as a log normal distribution with an offset of 23 Mya and a mean age of 35 Mya (HPC 90% = 27.7-47.1). This mean corresponded to the consensus age of Odontoceti estimated from molecular analyses [15–18].

(ii) *Kentriodon pernix* [19] was used to calibrate the minimum age of Ziphiidae + Delphinidae to 18 Mya as recommended by Lambert et al. 2017 [20]. This calibration point was set as a log normal distribution with an offset of 18 Mya and a mean of 18.5 Mya (HPC 90% = 18-19.5 Mya).

(iii) *Globicetus hiberus* [21] and *Archaeoziphius microglenoideus* [22] supported a minimum age of crown Ziphiidae of approximately 13.2 Mya. This age constraint, which was recommended by Geisler et al. 2011 [18] and Lambert et al. 2017 [20], is here applied as a log normal distribution with an offset of 13.2 Mya and a mean of 13.8 Mya (HPC 90% = 13.2-15.0 Mya). Similar node calibrations for crown Ziphiidae have previously been used by Galatius et al. 2018 [14].

(iv) *Salumiphocaena stocktoni [23]* was used to calibrate the minimum age of *Monodontoida*e (*Monodontidae + Phocoenidae*) to 7.5 Mya as recommended by Lambert et al. 2017 [20] and Geisler et al. 2011 [18]. This calibration point was set as a log normal distribution with an offset of 7.5 and a mean of 9.9, (HPC 90% = 7.8-15.0 Mya), similarly to Steeman et al. 2009 [16] and Dornburg et al. 2012 [12].

Complete mitogenome sequences of the 14 other toothed whales (supplementary table 2) were downloaded from NCBI and gene regions were extracted using published annotations. All regions including protein-coding genes (n=13), the control region (n=1), rRNAs (n=2) and tRNAs (n=22) were individually aligned using Mafft 7.3 [24]. The alignments were checked and if necessary manually corrected to match the reading frame for the protein coding genes. The three codon positions of each gene were split into separate partitions, which resulted in a total of 64 data subsets. Dambe [25] was used to test for substitution saturation. The best substitution model for each data subset was determined using PartitionFinder 2.1.1 [26] (supplementary table S3a).

The phylogenetic analysis of odontocetes was run in Beast v2.5.1 [27] using the tRNAs, rRNAs, first and second codon positions of the protein coding regions. The third codon positions of the protein coding regions and the control region were excluded due to substitution saturation and alignment difficulties, respectively. A calibrated Yule model [28] with an estimated birth rate was used as the tree model prior. All regions were set to have the same topology and clock rate but the substitution models for each region were set according to PartitionFinder and

substitution rates were estimated. The gamma category count was set to 4. The heterogeneity of clock rates among odontocetes was accommodated by applying a relaxed log normal clock with an estimated rate [12]. The mean of the clock rate prior was set to 0.004 as estimated by an initial run using a strict clock. The birth rate prior was set as a uniform distribution ranging from 0 to 1000. Following Heath (2015) [29] the prior on the transversion rate of A<-->G mutations was given a gamma distribution with a beta value of 2 and alpha values of 0.5 giving a mean of 1 (C<-->T mutations were fixed to 1). The priors on the transition rates were set the same way, but alpha values were set to 0.25 giving mean values of 0.5. The Markov chain Monte Carlo (MCMC) was run twice with 50,000,000 steps logged every 5,000 steps. The tree files and log files of the two runs were combined using LogCombiner v2.5.1 [30] with a burnin of 10%. We assessed stationarity by examining ESS values in Tracer v1.7.1 [31] (ESS values above 200 for all parameters) and convergence by comparing posterior distributions between the two chains. TreeAnnotator v2.5.1 [32] was used to create a maximum clade credibility tree with mean node heights and a posterior probability limit of 0.9 based on the combined trees. The tree was plotted using FigTree 1.4.3 available from (http://tree.bio.ed.ac.uk/software/figtree/).

Narwhal phylogeny

The phylogenetic analysis of narwhals included the 64 unique mitogenome haplotypes identified. Protein-coding genes (n=13), tRNAs (n=22), rRNAs (n=2) and the control region for the 64 narwhal haplotypes were separated and aligned using the mafft algorithm [24] based on the published annotation of the reference narwhal mitogenome [5]. The alignments were checked and, if necessary, manually corrected to match the reading frame of the protein coding genes.

All regions were set to have the same topology and clock rate, but the substitution models for each region were set according to PartitionFinder, and substitution rates were estimated. We used a coalescent constant population model as data was from a single species, and applied a strict clock as we expect little heterogeneity in clock rate within narwhals. We set the prior on the clock rate as a uniform distribution with lower and upper limits of 0 and 1, respectively. We applied a calibration to the root age of the tree (TMRCA) that is the mean and 95% credibility intervals of the oldest divergence within narwhals as estimated in the odontocete phylogeny (using a log-normal distribution with a mean of 0.13, a standard deviation of 0.29 and

an offset of 0). The Markov chain Monte Carlo (MCMC) was run twice with 50,000,000 steps logged every 5,000 steps. The tree files and log files of the two runs were combined using LogCombiner v2.5.1 [30] with a burnin of 10%. We assessed stationarity by examining ESS values in Tracer v1.7.1 [31] (ESS values above 200 for all parameters) and convergence by comparing posterior distributions between the two chains. We used TreeAnnotator v2.5.1 [32] to create a maximum clade credibility tree with mean node heights and a posterior probability limit of 0.9 based on the combined trees. The tree was plotted using FigTree 1.4.3 available from (http://tree.bio.ed.ac.uk/software/figtree/).

Bayesian skyline analysis

The substitution models for each partition were defined in the narwhal phylogeny. A strict clock was used and the clock rate was estimated. We calibrated the age of the root using the mean and 95% highest posterior density (HPD) estimate from the phylogenetic tree of narwhals (using a log-normal distribution with a mean of 0.11, a standard deviation of 0.3 and an offset of 0). We also ran the analysis calibrating the clock rate (using the value obtained in the phylogenetic tree of narwhals) without setting a prior on the age of the root and got similar results (results not shown). Two chains of 50,000,000 states with sampling every 5,000 states were run and burnin was 10%. We assessed stationarity by examining ESS values in Tracer v1.7.1 [31] (ESS values above 200 for all parameters) and convergence by comparing posterior distributions between the two chains. We used a generation time of 30 years to scale the population size estimates [33].

Species distribution models

We used AquaMaps to predict suitable available habitat for narwhals for three climatic periods, the LGM, representing a glacial period, the present, representing an interglacial period, and for year 2100 [34–36]. AquaMaps is a bioclimatic model that combines existing point occurrence data with independent knowledge about the distribution and habitat usage of a species to estimate its environmental preference with respect to certain parameters and generate large-scale predictions of natural occurrence. These predictions are visualized in a grid of 0.5 degree latitude by 0.5 degree longitude cells by comparing the environmental envelopes (i.e. habitat usage of the species) with local environmental parameters, to determine the relative suitability of given geographical areas for a species. Prediction values range between 0 (not suitable) to 1 (highly suitable) and are the product of the suitability scores assigned for each environmental

parameter. In this study, we generated environmental envelopes using distribution data inferred from updated distribution maps published in the Global Review of Monodontids (GROM) report [37]. Ideally, raw sighting data points should have been used. However, we did not have access to this data for all regions, and therefore used distribution maps recently produced by a panel of narwhal experts [37]. Distance to the land and mean annual values for depth, sea surface temperature, salinity and sea ice concentration were extracted for summer, winter and year-round distribution of narwhals, as the species is migratory and distributions are seasonal. We generated environmental envelopes and computed predictions of habitat suitability for the LGM and the present for summer, winter and year-round in the northern hemisphere, using all parameters. We generated predictions for the winder only for 2100. We restricted the predictions to the Atlantic side as narwhals are not distributed in the Pacific. We generated predictions for different combinations of the above parameters. Maps in the main text were plotted with an Arctic projection using QGIS 3.2 considering a probability threshold of 0.6, which represents habitats of high suitability [34]. The maps in the supplementary material show all habitat suitability probabilities. The size and the mean latitude of the suitable available habitat were calculated and compared between the three periods.

Results

Bioinformatics

The mitogenome sequence had 16,383 sites, including 16,030 sites with no missing data or gaps across all individuals. 353 sites had missing data; 15,879 sites were invariables and 151 sites were variable. Over all 121 individuals and the full 13,838 bp sequence, representing 1,982,343 nucleotides, 16,438 nucleotides were called Ns, representing 0.83% of the nucleotides. Missing data mainly occurred at the ends of the mitogenome sequences and in the control region for the samples with the lowest coverage. We called 489 nucleotides as Ns as there was < 10 reads with one or more nucleotide variations, or > 10 reads with a single nucleotide not representing > 80% of the reads. This mainly occurred in the control region and might represent heteroplasmy.

Diversity statistics

Haplotype diversity did not significantly differ among localities (table S4). Nucleotide diversity was significantly different in eight comparisons (table S4). The value in East Greenland was

significantly higher than in four other localities, and the value in Svalbard was significantly lower than in three other localities. Nucleotide diversity in Northern Hudson Bay was significantly higher than in Eclipse Sound.

Bayesian skyline analysis

Our Bayesian skyline analysis using third codon positions only of the protein coding regions indicates a more recent increase in *Nef,* starting ~6 kya (supplementary figure S6), although confidence intervals in the two analyses largely overlap. The discrepancy between results based on the two different data sets might reflect purifying selection on the first and second codon positions of the protein coding regions, but this hypothesis is not statistically supported.

Species distribution models

All combinations of parameters and season/period show the same results: a shift northwards and an increase in available suitable habitat between the LGM and the present (figure 4a-c, supplementary figures S7-11). Results using sea ice concentration, depth and sea surface temperature envelopes were conservative (figure 4, supplementary figures S7-8), and the increase in habitat was lower than when including salinity (supplementary figures S9-11). In addition, these three parameters (sea ice concentration, depth and temperature) are critical for the species and we therefore chose to present the results for this combination of parameters. We present the results for the winter habitat suitability. As narwhals are intensively foraging during the winter [38], this season is critical for the species. In addition, the predicted distribution in the present time bin matched the known occurrence of narwhals during winter relatively well, providing confidence in the estimate for the LGM projection. We would like to stress that densities are not taken into account, and therefore predictions do not reflect fine habit preferences, and hence we are limited by the amount of available parameters from the Pleistocene (LGM). We use mean annual average values as this is what is available for the LGM time bin. This analysis therefore aims to see whether there has been a change in the size and distribution of suitable habitat, and we do not aim to interpret the results further.

‘

Literature

1. Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding MHS, Samaniego JA, Wales N, Sicheritz-Pontén T, Gilbert MTP. 2018 Single-tube library preparation for degraded DNA. *Methods Ecol. Evol*. **9**, 410–419. (doi:10.1111/2041-210x.12871)

2. Rohland N, Reich D. 2012 Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res*. **22**, 939–946. (doi:10.1101/gr.128124.111)

3. Schubert M *et al.* 2014 Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc*. **9**, 1056–1082. (doi:10.1038/nprot.2014.063)

4. Schubert M, Lindgreen S, Orlando L. 2016 AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88.

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

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

7. McKenna A *et al.* 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

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

9. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018 MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549.

10. Leslie MS, Archer FI, Morin PA. 2019 Mitogenomic differentiation in spinner (*Stenella longirostris*) and pantropical spotted dolphins (*S. attenuata*) from the eastern tropical Pacific Ocean. *Mar. Mamm Sci*. **35**, 522–551. (doi:10.1111/mms.12545)

11. Nikaido M *et al.* 2001 Retroposon analysis of major cetacean lineages: the monophyly of toothed whales and the paraphyly of river dolphins. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 7384–7389.

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

13. Mchedlidze GA. 1970 *Nekotorye obshchie cherty istorii kitoobraznykh*. Metsniereba, Tbilisi.

14. Galatius A, Olsen MT, Steeman ME. 2018 Raising your voice: evolution of narrow-band high-frequency signals in toothed whales (Odontoceti). *Biol. J. Linn. Soc. Lond.* **126**, 213–224.

15. McGowen MR, Spaulding M, Gatesy J. 2009 Divergence date estimation and a comprehensive molecular tree of extant cetaceans. *Mol. Phylogenet. Evol.* **53**, 891–906.

16. Steeman ME *et al.* 2009 Radiation of extant cetaceans driven by restructuring of the oceans. *Syst. Biol.* **58**, 573–585.

17. Xiong Y, Brandley MC, Xu S, Zhou K, Yang G. 2009 Seven new dolphin mitochondrial genomes and a time-calibrated phylogeny of whales. *BMC Evol. Biol.* **9**, 20.

18. Geisler JH, McGowen MR, Yang G, Gatesy J. 2011 A supermatrix analysis of genomic, morphological, and paleontological data from crown Cetacea. *BMC Evol. Biol.* **11**, 112.

19. Kellogg R. 1927 *Kentriodon Pernix: A Miocene Porpoise from Maryland*. U.S. Government Printing Office.

20. Lambert O, Bianucci G, Urbina M, Geisler JH. 2017 A new inioid (Cetacea, Odontoceti, Delphinida) from the Miocene of Peru and the origin of modern dolphin and porpoise families. *Zool. J. Linn. Soc.* **179**, 919–946.

21. Bianucci G, Miján I, Lambert O, Post K, Mateus O. 2013 Bizarre fossil beaked whales (Odontoceti, Ziphiidae) fished from the Atlantic Ocean floor off the Iberian Peninsula. *Geodiversitas* **35**, 105–153.

22. Lambert O, Louwye S. 2006 *Archaeoziphius microglenoideus*, a new primitive beaked whale (Mammalia, Cetacea, odontoceti) from the Middle Miocene of Belgium. *J. Vert. Paleontol.* **26**, 182–191.

23. Wilson LE. 1973 *A Delphinid (Mammalia, Cetacea) from the Miocene of Palos Verdes Hills, California*. University of California Press.

24. Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.

25. Xia X, Xie Z. 2001 DAMBE: software package for data analysis in molecular biology and evolution. *J. Hered.* **92**, 371–373.

26. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2016 PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772–773.

27. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014 BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537.

28. Heled J, Drummond AJ. 2012 Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.* **61**, 138–149.

29. Heath TA. 2015 Divergence time estimation using BEAST v2.2.0. In *Source URL: http://treethinkers. org/tutorials/divergence-time-estimation-using-beast/: Tutorial written for workshop on applied phylogenetics and molecular evolution, Bodega Bay California*, pp. 1–44.

30. Rambaut A, Drummond AJ. 2014 LogCombiner v2.1.3. *Institute of Evolutionary Biology, University of Edinburgh, UK*

31. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018 Posterior summarization in

Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904.

32. Rambaut A, Drummond AJ. 2013 TreeAnnotator v1. 7.0. *Available as part of the BEAST package at http://beast.bio.ed.ac.uk*

33. Garde E, Hansen SH, Ditlevsen S, Tvermosegaard KB, Hansen J, Harding KC, Heide-Jørgensen MP. 2015 Life history parameters of narwhals (*Monodon monoceros*) from Greenland. *J. Mamm*. **96**, 866–879. (doi:10.1093/jmammal/gyv110)

34. Kaschner K, Tittensor DP, Ready J, Gerrodette T, Worm B. 2011 Current and future patterns of global marine mammal biodiversity. *PLoS One* **6**, e19653.

35. Ready J, Kaschner K, South AB, Eastwood PD, Rees T, Rius J, Agbayani E, Kullander S, Froese R. 2010 Predicting the distributions of marine organisms at the global scale. *Ecol. Modell.* **221**, 467–478.

36. Kaschner K, Kesner-Reyes K, Garilao C, Rius-Barile J, Rees T, Froese R. 2016 AquaMaps: Predicted range maps for aquatic species. World wide web electronic publication, www.aquamaps.org, Version 08/2016.

37. NAMMCO. 2018 Report of the Global Review of Monodontids. 13-16 March 2017, Hillerød Denmark.

38. Laidre KL, Heide-Jorgensen MP. 2005 Winter feeding intensity of narwhals (*Monodon monoceros*). *Marine Mammal Science*. **21**, 45–57. (doi:10.1111/j.1748-7692.2005.tb01207.x)