

In the format provided by the authors and unedited.

Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA

Eva Egelyng Sigsgaard^{1,2}, Ida Broman Nielsen¹, Steffen Sanvig Bach³, Eline D. Lorenzen², David Philip Robinson⁴, Steen Wilhelm Knudsen², Mikkel Winther Pedersen¹, Mohammed Al Jaidah⁵, Ludovic Orlando¹, Eske Willerslev^{1,6,7}, Peter Rask Møller², Philip Francis Thomsen^{1*}

*) pftthomsen@snm.ku.dk

1) Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, DK-1350 Copenhagen K, Denmark, 2)

Section for Evolutionary Genomics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, DK-1350 Copenhagen K,

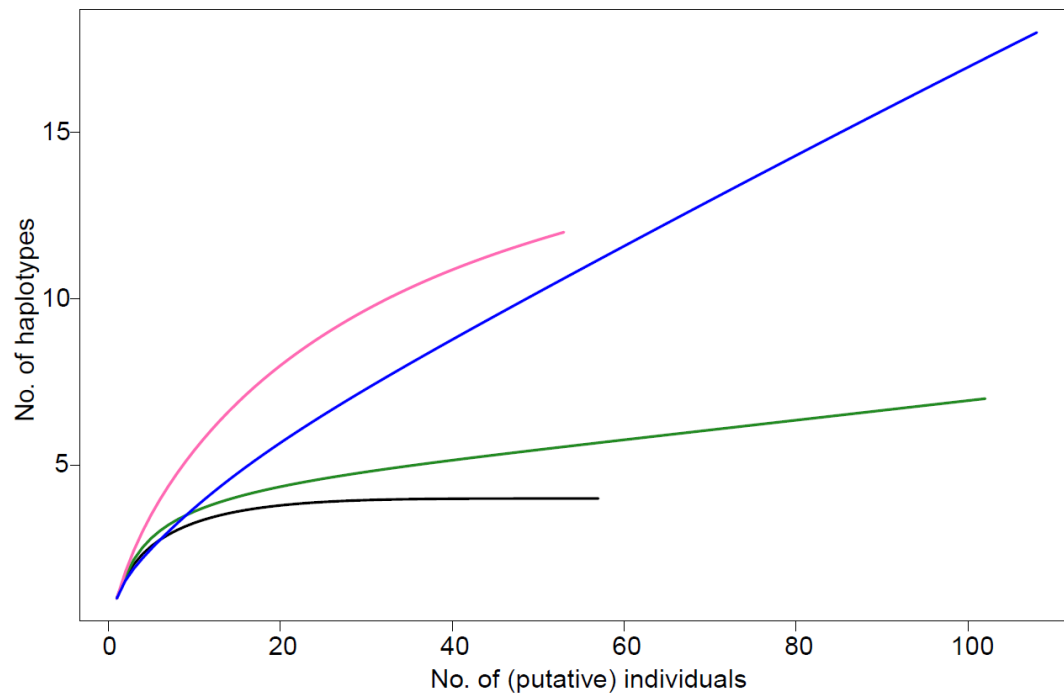
Denmark 3) Maersk Oil Research and Technology Centre, Al Jazi Tower, Building 20, Zone 60, Street 850, West Bay, Doha, Qatar, 4) School of Life

Sciences, Heriot-Watt University, Riccarton Campus, Edinburgh, EH14 4AS, UK, 5) Ministry of Municipality and Environment, Conference Centre Street, Al

Dafna 61, Doha, Qatar, 6) Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, UK, 7) Wellcome Trust Sanger Institute,

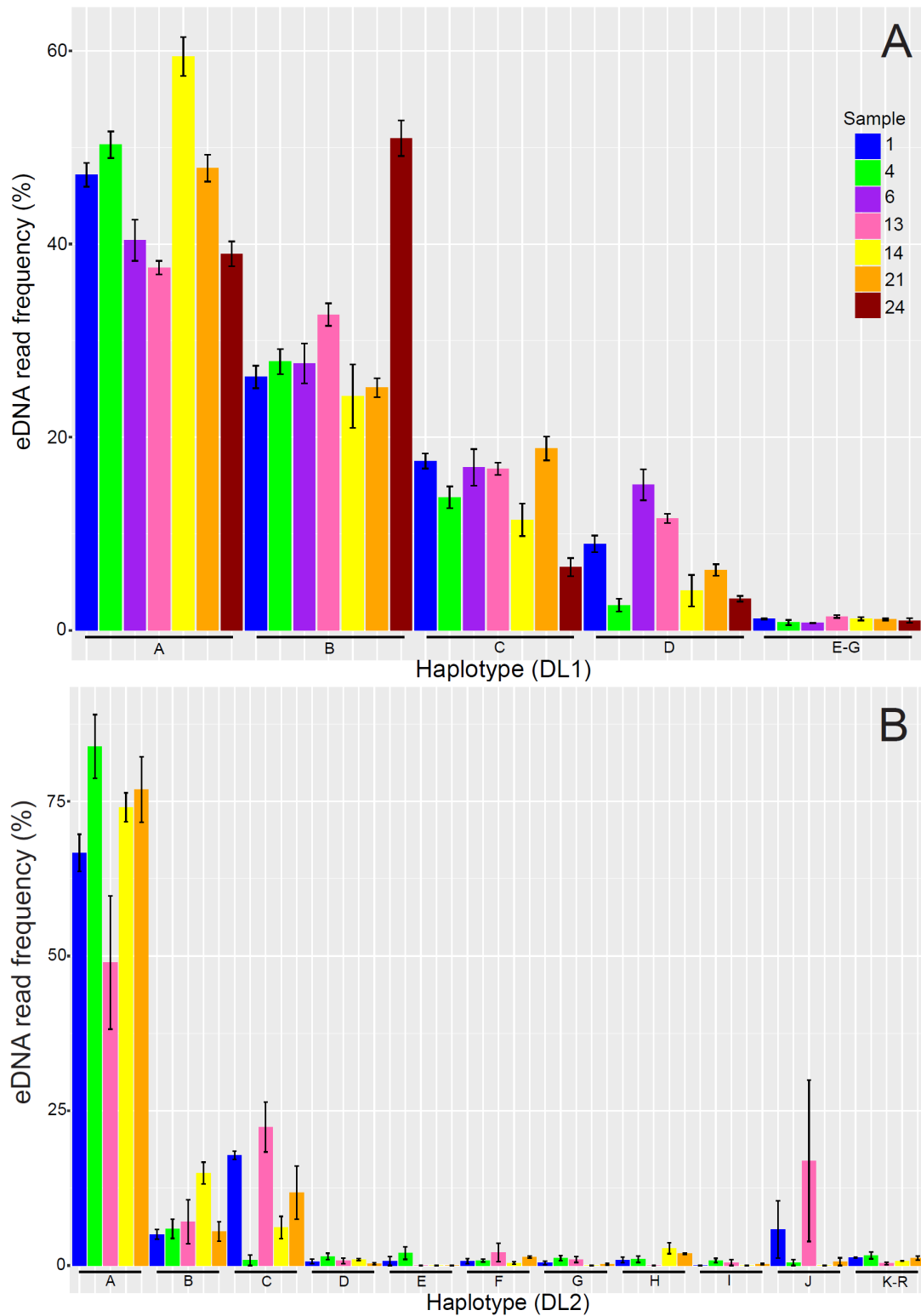
Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Supplementary Figures



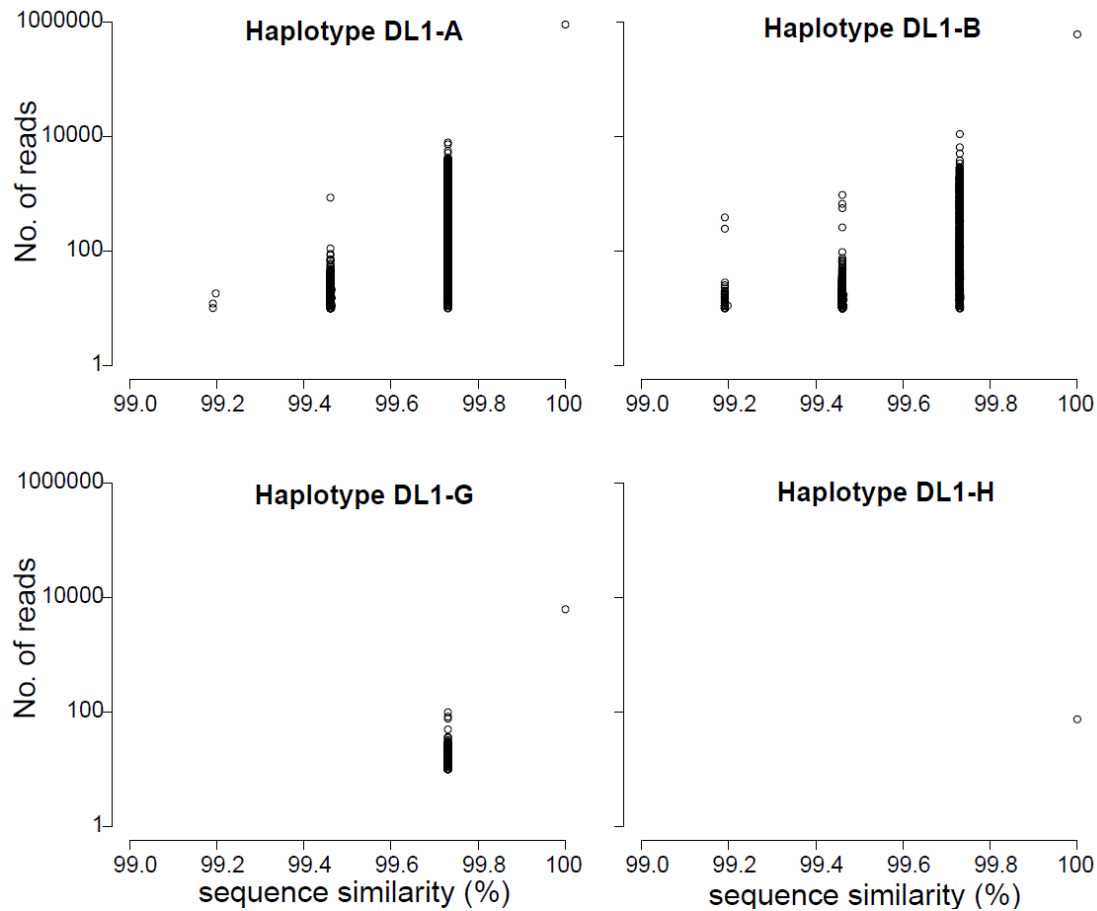
Supplementary Figure S1. Rarefaction curves

Rarefaction curves showing the number of DL1 and DL2 haplotypes expected in a certain number of whale shark tissue samples, and eDNA sequence reads from water samples, respectively, based on the results from this study. DL1 tissue (black); DL1 water (green); DL2 tissue (pink) and DL2 water (blue).



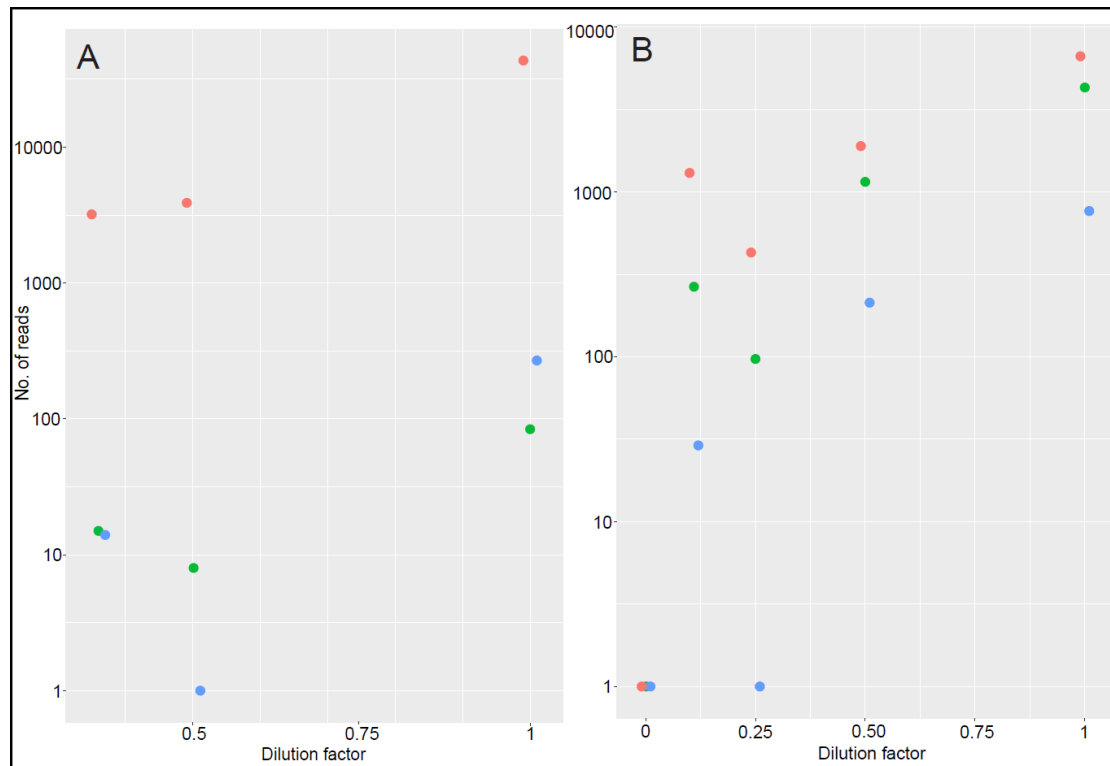
Supplementary Figure S2. Haplotype frequencies in water samples

Bar diagrams showing mean and standard error of haplotype frequencies across PCR replicates for each sequenced individual water samples and shown separately for each haplotype. Haplotypes in very low counts (DL1-E-G and DL2-K-R) are collapsed. A) DL1 and B) DL2.



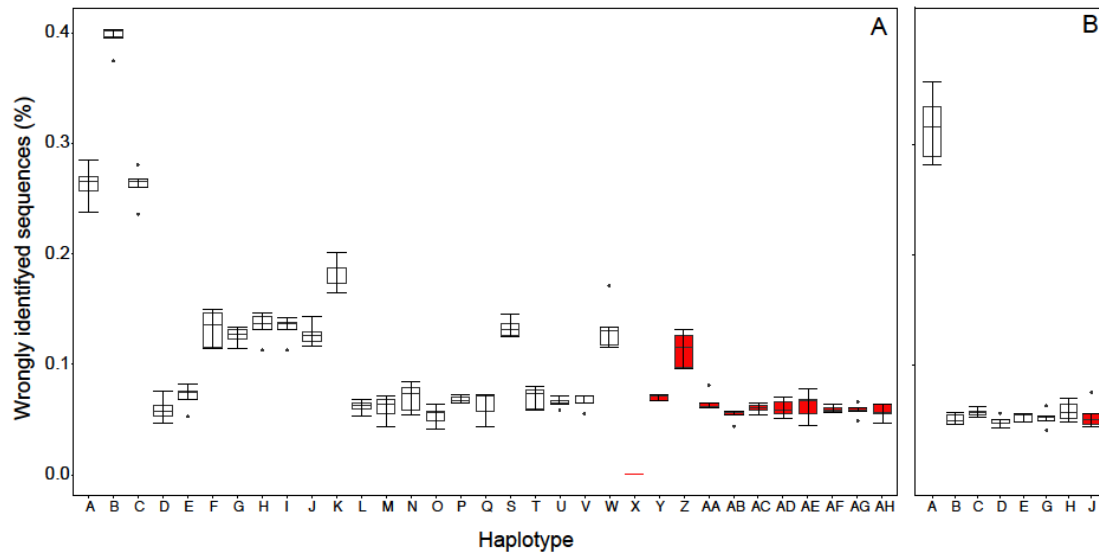
Supplementary Figure S3. Examples of sequencing error plots

Sequencing error plots for four DL1 haplotypes showing eDNA sequencing read counts as a function of similarity to the reference haplotype (from 99% to 100% similarity). DL1-H represents a haplotype, which was flagged as “dubious” and was later removed from the dataset as a likely result of PCR error.



Supplementary Figure S4. Mock positive control

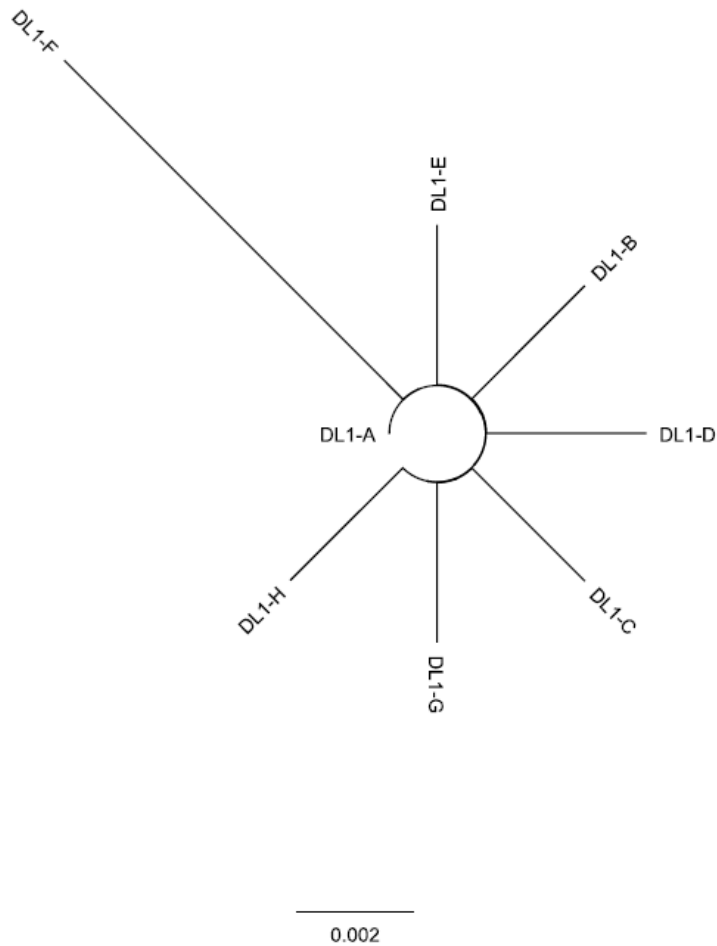
The number of reads obtained for the haplotypes added to the mock sample (positive control), as a function of their initial relative concentration. Data is shown for the original mock sample (green), re-sequencing of the original mock sample (blue), as well as a newly prepared mock sample (red). A) DL1 and B) DL2.



Supplementary Figure S5. Sequencing error modeling

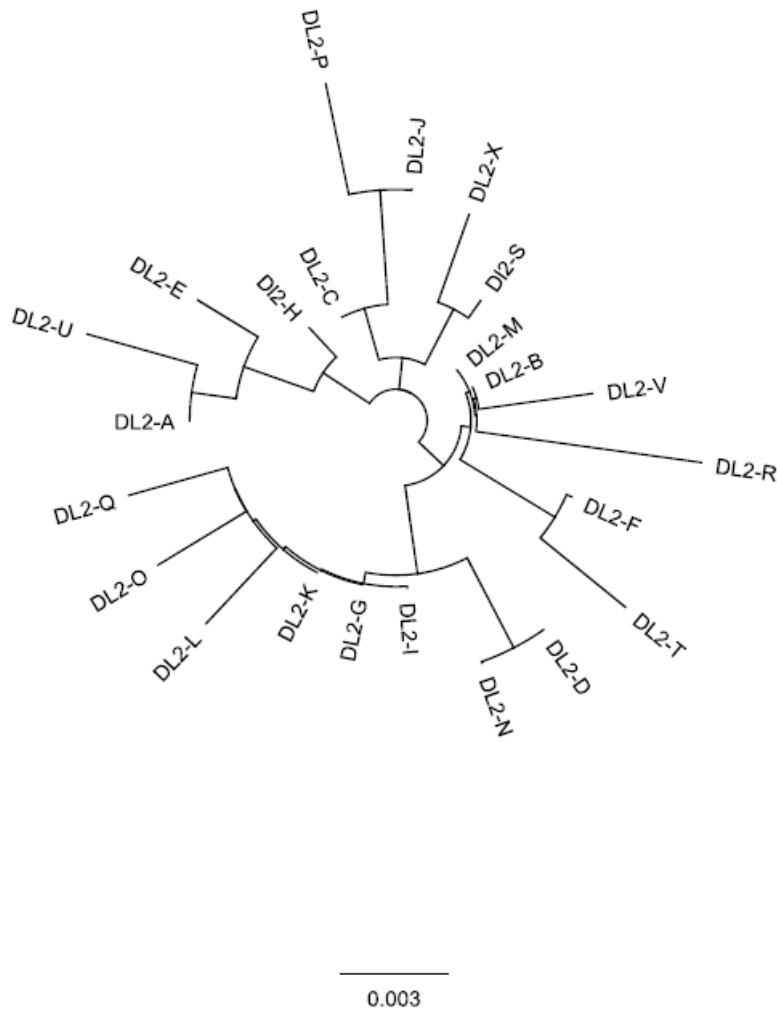
Results of sequencing error modeling using a uniform error rate of 0.3% per nucleotide. Boxplots show the percentage of sequences wrongly identified to a given haplotype (false positives) as a result of errors, across five model iterations. Red boxes show known haplotypes not found in the water samples.

A) DL2 and B) DL1.



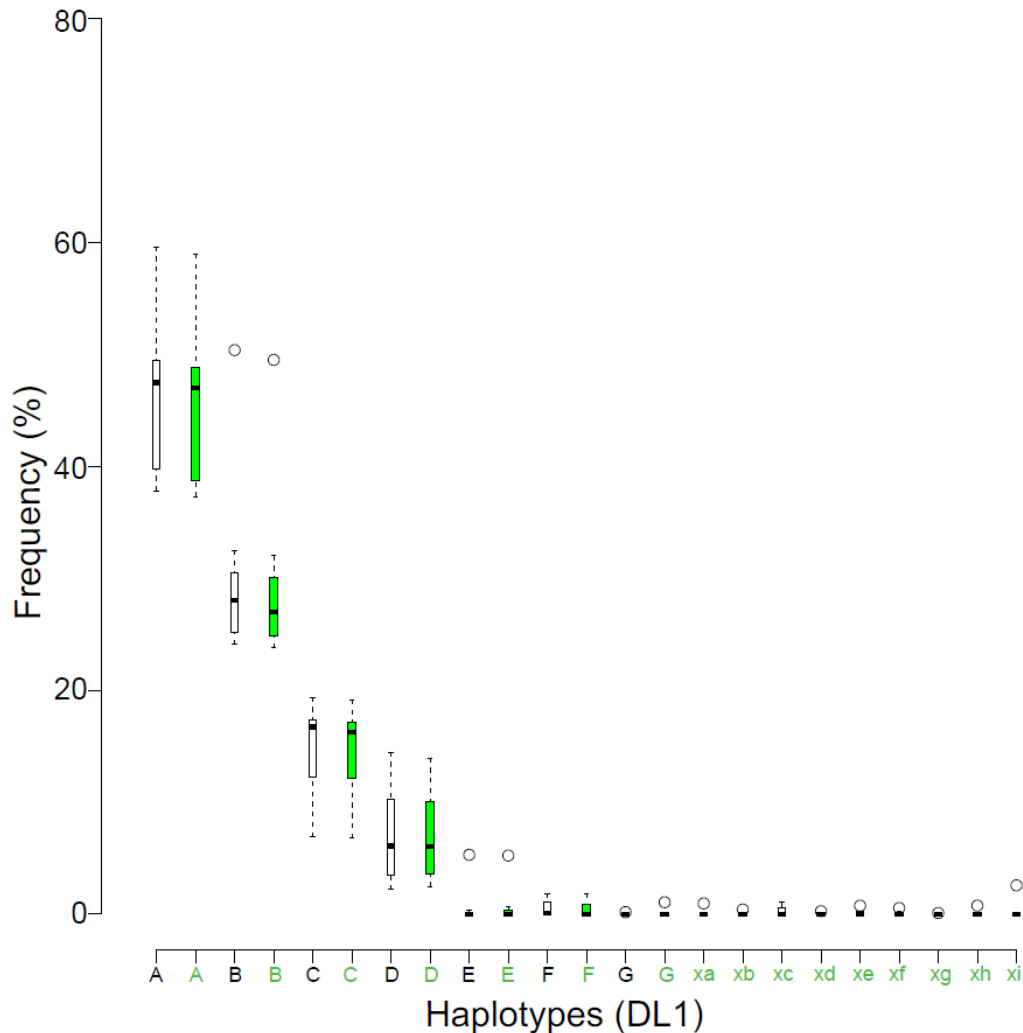
Supplementary Figure S6. Neighbour-joining tree of DL1 haplotypes

Phylogenetic tree of all DL1 haplotypes found in eDNA (before cleaning) and tissue samples from Qatar. The tree was constructed using the Tamura-Nei distance model and the neighbour-joining method in Geneious v. 7.1.9 (Biomatters Ltd.). Distances between haplotypes represent genetic distances in terms of the number of substitutions per site.



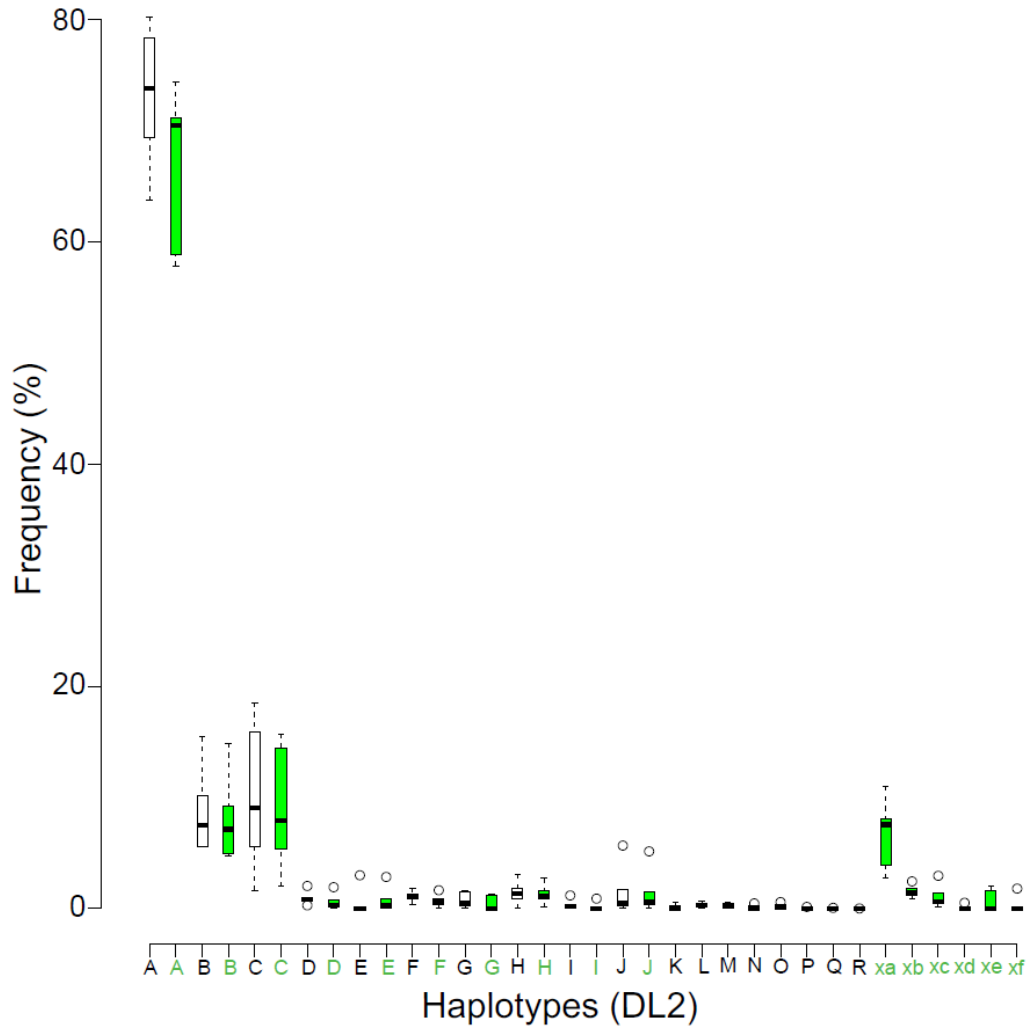
Supplementary Figure S7. Neighbour-joining tree of DL2 haplotypes

Phylogenetic tree of all DL2 haplotypes found in eDNA (before cleaning) and tissue samples from Qatar. The tree was constructed using the Tamura-Nei distance model and the neighbour-joining method in Geneious v. 7.1.9 (Biomatters Ltd.). Distances between haplotypes represent genetic distances in terms of the number of substitutions per site.



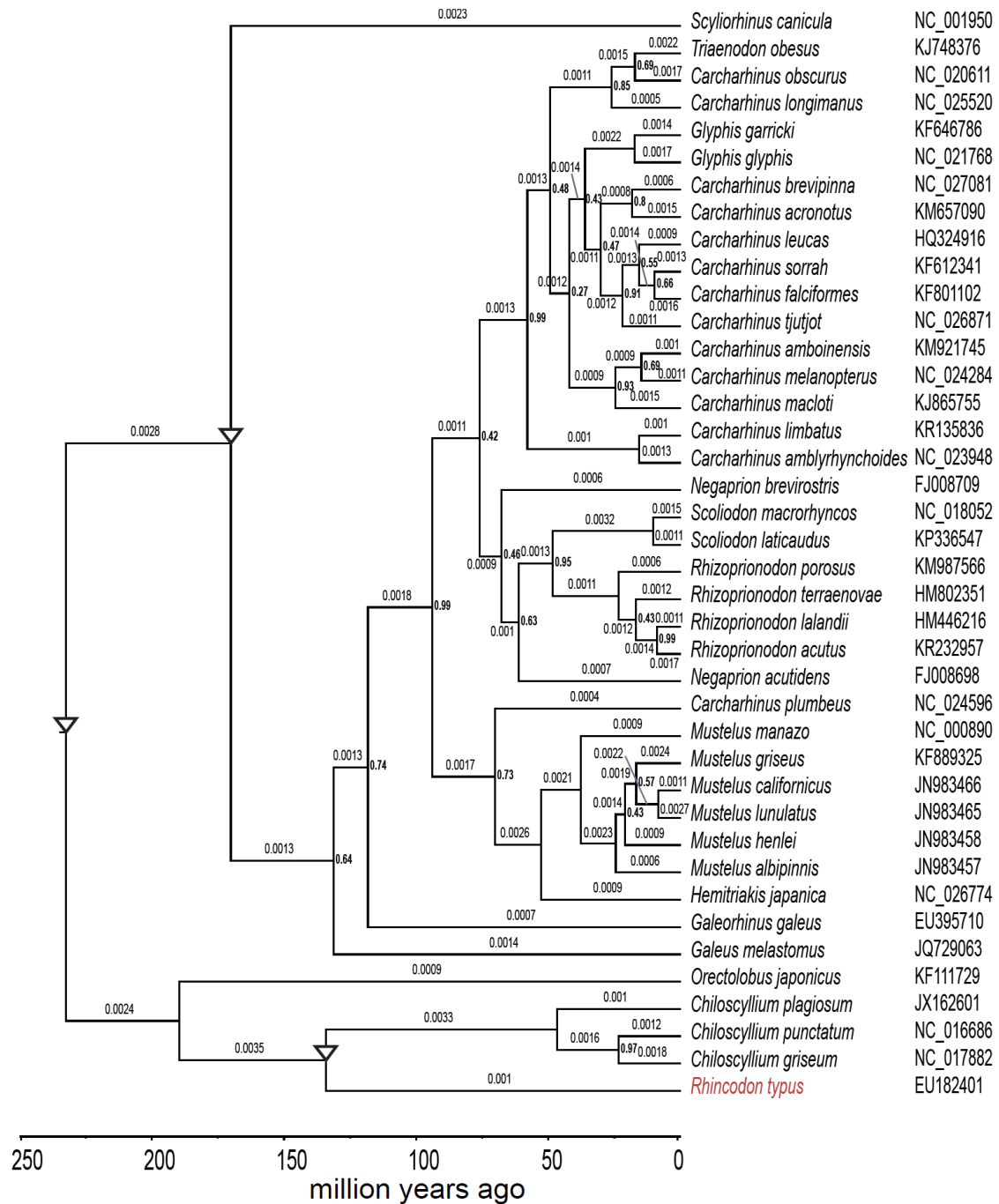
Supplementary Figure S8. DL1 haplotype frequencies in eDNA with and without use of a reference database

Frequencies of mitochondrial control region DL1 haplotypes obtained from seawater eDNA sequence reads (boxplots, $n=7 \times 3$ water samples) identified with (black) and without (green) use of a reference sequence database. Boxplot whiskers: most extreme data point ≤ 1.5 times the inter-quartile range from the box.



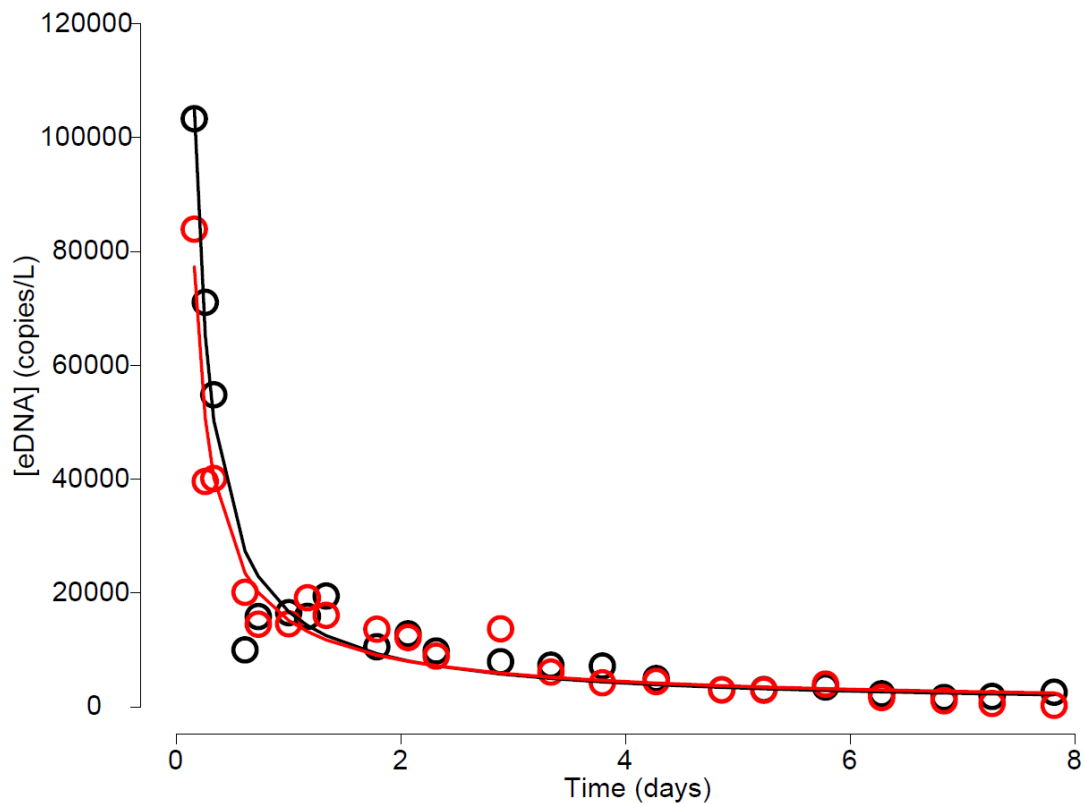
Supplementary Figure S9. DL2 haplotype frequencies in eDNA with and without use of a reference database

Frequencies of mitochondrial control region DL2 haplotypes obtained from seawater eDNA sequence reads (boxplots, $n=5 \times 3$ water samples) identified with (black) and without (green) use of a reference sequence database. Boxplot whiskers: most extreme data point ≤ 1.5 times the inter-quartile range from the box.



Supplementary Figure S10. Phylogenetic tree

Phylogenetic tree (maximum likelihood) based on the DL2 region of 40 species of Carcharhiniformes and Orectolobiformes, including the whale shark (red). Estimated mutation rates in the region are shown for each branch, and posterior probabilities are indicated on the nodes (probabilities of 1 not shown). Calibrated nodes are indicated by a triangle, and a time scale is provided. NCBI accession numbers for each taxon are shown on the right.



Supplementary Figure S11. Degradation model

Degradation of whale shark eDNA over time in two 90 L seawater samples collected in the study area, and placed in direct sunlight (red circles) or shade (black circles), respectively. Based on qPCR using a whale shark specific TaqMan assay targeting a 105 bp cytochrome b (*CYTB*) gene product. Regression lines show the fit of an exponential decay model to the data from each treatment.

Supplementary Experimental Procedures

Study site

Seawater samples were collected in the central Arabian Gulf off the coast of Qatar (Figure 1B, Supplementary Table S1). With an average depth of 30 m, the central Gulf is a eutrophic, warm (up to 35°C), high-saline (up to 40 ppt) sea^{1–3}, with low fish diversity for a tropical area, but with a high abundance of many of the fish species present⁴.

Whale shark aggregation at Al Shaheen

The presence of whale sharks in the Arabian Gulf has been known for many years⁵. However, it was not until 2010 that large whale shark aggregations were observed in May to September in the Al Shaheen area off Qatar. It is believed that the sharks are gathering near the Al Shaheen oil field to feed on fish spawn, as has been described for aggregations in Belize⁶ and Yucatan, Mexico⁷. Robinson et al. (2013) suggested that at least one of the spawning fish species is the mackerel tuna *Euthynnus affinis*. This species is closely related to one of the main prey species of whale sharks in the Yucatan area, little tunny *Euthynnus alletteratus*⁷.

Seawater eDNA sampling

On the 27th-28th of May, 2013, and 19th-20th of May, 2014, PRM, IBN, EES, PFT, SSB and staff from Qatar Ministry of Environment collected water samples from the Ministry of Environment's boat *R/V Saqt Al Khaleej*, at 15 locations, most of which are located in the Al Shaheen oil field, 90 km north east of Qatar (Figure 1B, Supplementary Table S1). In addition, SSB collected water samples on the 28th of May and 27th of June, 2014, at one of the previously sampled locations (Loc. 4, Supplementary Table S1). Two locations were sampled both years (Loc. 4 and Loc. 5, see Supplementary Table S1). Three samples of 500 mL were taken at each site, except at locations 9 and 10, where two 500 mL samples were collected. Samples were taken approximately 10 cm below the surface and were filtered on the same day, either immediately after sampling (all samples from 2014), or upon return to Doha (all samples from 2013). Samples were kept on ice until filtering). Filtering was done through sterile 0.22 µm SterivexTM-GP filters (Merck Millipore, Germany) using 60 mL syringes (Soft-Ject®, HSW, Tuttlingen, Germany). We have good experience with the SterivexTM filters, which are enclosed and

therefore much less exposed to contamination in the field or in the lab than other filters. In addition to the 500 mL samples, a large water sample of 6*30 L was collected in 2014 at Loc. 13 (Supplementary Table S1) for an eDNA degradation experiment (described below). The water temperature was 30 °C during sampling in May 2014. Filters were stored on ice during transport from the study site to Doha and then kept frozen at -18 °C until processing at the laboratory in Copenhagen (Supplementary Table S1).

DNA extraction from seawater samples

Extraction was performed using the Qiagen DNeasy® Blood and Tissue kit (modified spin column protocol). After cleaning the outside of each filter with bleach, 720 µL ATL-buffer and 80 µL proteinase K was added through the inlet, and the filters were incubated for two hours at 56 °C with agitation. The buffer mix was transferred to 2 mL Eppendorf tubes using 3 mL disposable Luer-Lock™ Syringes and 600 µL of each sample was then pipetted to new 2 mL tubes to obtain equal volumes for all samples. A volume of 600 µL AL-buffer and 600 µL 96% ethanol was then added to each tube, and the mixture was centrifuged 600 µL at a time in DNeasy Mini spin columns for 1 minute at 7900 rpm. The flow-through was discarded. Columns were then washed with 500 µL AW1-buffer and centrifuged as above, followed by washing with 500 µL AW2- buffer and centrifugation at 11,600 rpm for 2 minutes. Lastly, DNA was eluted using 110 µL ddH₂O, with incubation for 5 minutes at 37 °C and centrifugation for 1 minute at 11,600 rpm. The flow-through was transferred back to the spin column, and incubation and centrifugation was repeated. Extraction controls were included throughout the extraction process and these were tested for amplification in all further analyses.

Whale shark DNA reference database

A reference database of whale shark sequences for comparison with obtained eDNA sequences was built using whale shark tissue samples from Qatar, provided by The Ministry of Environment, State of Qatar, DPR and Jennifer V. Schmidt (Dept. of Biological Sciences, University of Illinois, Chicago, USA). Samples were obtained from live sharks using a biopsy spear. 3 g of tissue was retrieved and fixed in 96% ethanol. Each shark was photographed for subsequent identification and included in the WildBook database (www.whaleshark.org) (Table 2). Individuals without photo were identified using microsatellite data (available upon request from DPR). When possible, the sex of the shark was

determined based on the presence (males) or absence (females) of claspers, and the length of the shark was estimated by eye, comparing the shark to other snorkelers or to the boat.

A total of 110 tissue samples (Supplementary Table S1) were initially extracted using the Qiagen DNeasy® Blood and Tissue kit, using the manufacturer's protocol with the following modifications: i) samples were incubated in lysis buffer for 3 hours and ii) elution was done in ddH₂O, after incubation in the spin-column for 5 minutes at 37 °C. PCR amplification of the mitochondrial control region was performed using the WSCR1-F and WSCR1-R primers designed by Castro et al. (2007) set up in 25 µL reactions using 18.4 µL ddH₂O, 2.5 µL GeneAmp® 10X PCR Buffer I, 1 µL of each primer (10 µM), 1 µL dNTPs (2.5 mM), 1 µL purified DNA and 0.1 µL AmpliTaq Gold® polymerase. Cycling parameters were 95 °C for 5 minutes, 35 cycles of 94 °C for 2 minutes, 54 °C for 2 minutes, and 72 °C for 2 minutes, and a final elongation step of 72 °C for 7 minutes. PCR products were visualized on a 2% agarose gel stained with GelRed™ (Biotium Inc.). Extraction blanks showed no bands on the gels. PCR amplicons were sent to Macrogen Europe and Sanger sequenced using the WSCR1-F and WSCR1-R primers, as well as the primers WSCR2-F and WSCR2-R⁸, generating four sequences. The four sequences obtained from each sample were carefully checked for quality and aligned in Geneious v. 7.1.7 (Biomatters Ltd.), and further trimmed manually. The consensus sequences were aligned and checked for bases that occurred only once or twice at a certain position in the alignment. These base calls were inspected in the raw sequences to determine whether they appeared to be errors or true single nucleotide polymorphisms (SNPs). After trimming and careful inspection, sequences for both targeted mitochondrial d-loop fragments, hereafter referred to as DL1 (d-loop fragment 1) and DL2 (d-loop fragment 2) (see “PCR amplification of d-loop fragments from eDNA”) were obtained from 49 individuals. For a further eight individuals the DL1 alone was obtained, and for a further four cases only the DL2 sequence was obtained (Supplementary Table S1). In addition to the sequences from Qatar, all whale shark control region sequences from National Center for Biotechnology Information (NCBI) (GenBank: EU182401 through EU182444, GU289922, KC633221, KF679782), as well as sequences from Vignaud et al. (2014) (available at [doi:10.5061/dryad.489s0](https://doi.org/10.5061/dryad.489s0)) were included in the reference library (Supplementary Table S1-S2). The Vignaud et al. (2014) sequences cover the DL2 region, but not DL1. As an approximate test of whether the genetic diversity of the Qatar aggregation had been sampled exhaustively, rarefaction plots were made for the DL1 and DL2 haplotypes obtained from the local tissue samples (Supplementary Figure S1). This was done using the “vegan” library

function “rarecurve” in R v. 3.2.3. For eDNA, the read frequencies were scaled to a total of 100 individuals. Haplotypes that appeared in less than 1% of the total reads in the final sequences were scaled to represent one individual (meaning that the total number of individuals can be above 100).

Haplotype designation

DnaSP v. 5.10.1⁹ was used to designate haplotypes based on separate alignments of DL1 and DL2 sequences from the whale shark reference library. In several of the sequences from Vignaud et al. (2014), the first 38 base pairs of sequence are unknown. These sequences were left out of further analyses, as potentially important variation in this first part of the DL2 region would otherwise have to be ignored. In the following haplotype designation, positions with an unknown base in one or more sequences were removed from the analysis (these positions were all invariable when excluding the “Ns”), while gaps representing insertions/deletions were included.

PCR amplification of d-loop fragments from eDNA

Two primer sets, RhitypDL1 (RhitypDLR1 (forward primer): 5'-CCACATTTCTATAACATATTA-3' and RhitypDLL1 (reverse primer): 5'-TATTGACGGCAGATGTCGAG-3'), and RhitypDL2 (RhitypDLR2: 5'-TGCATGGTTTTATGTACGTCAGT-3' and RhitypDLL2: 5'-TGGATTAATGCAGGTTTTTACAAAC-3'), targeting a ca. 412 and a 476-493 bp region of the d-loop region in the whale shark mitochondrial genome, respectively, were developed to investigate genetic diversity in the water samples. The region targeted by the DL2 primers varies in length due to several insertions/deletions, and is the most variable of the two target regions.

The eight extracted and pooled water samples that were taken where whale sharks were visibly present (samples Qat.01, 04, 06, 13, 14, 16, 19, and 20, see Supplementary Table S1) were PCR amplified in six replicates with each of the primer sets (Supplementary Figure S2). Both forward and reverse primers were tagged using eight nucleotide long oligos all differing by a minimum of three nucleotides¹⁰ to enable tracking of sequences to individual samples and PCR reactions after sequencing. Primers were synthesized by Biomers (Ulm, Germany) and were High Performance Liquid Chromatography (HPLC) purified. A unique combination of tags was used for each PCR reaction making it possible to track PCR products to the individual PCRs. Before the final PCR, a subset of these tag combinations was tested on whale shark tissue derived DNA, and subsequently on a small

scale (one PCR replicate per sample) on the extracted water samples. For the DL1 primer set, reactions were prepared in 25 μL volumes of 10 μL TaqMan Environmental MasterMix 2.0, 10 μL ddH₂O, 1 μL of each primer (10 μM) and 3 μL sample of extracted eDNA from filtered water. The same mix was used for the DL2 primer set except that 4 μL was added of each primer, and only 4 μL of ddH₂O was used. TaqMan Environmental MasterMix 2.0 is developed for qPCR analysis, and therefore contains among other things ROX dye. Furthermore, the mix contains dUTPs, to enable removal of carryover contamination with the enzyme Uracil-N-Glycosylase (UNG). However, this mix works very well for PCR amplification of environmental samples in our experience, and we have no reason to believe that the presence of ROX dye or of uracil in the amplicons should negatively affect downstream sequencing. Thermocycling parameters for the DL1 primer set were: 95 °C for 5 minutes, 40 cycles of 94 °C for 30 s, 55 °C for 30 s, 72 °C for 60 s, and a final elongation of 2 minutes at 72 °C, while for DL2 the following settings were used: 95 °C for 5 minutes, 45 cycles of 94 °C for 30 s, 50 °C for 30 s, 72 °C for 60 s, and then 2 minutes at 72 °C. In the final setup, each tagged primer was used only once, and a negative control was included for every tag combination. Extraction controls were run in six replicates with each primer set, using a tag combination that had not been used for the samples. After PCR, DNA presence was determined on 2% agarose gel stained with GelRedTM (Biotium Inc.). As some PCR reactions were unsuccessful, amplification of the samples in question was repeated using new tag combinations. For DL1, six positive replicates were obtained for seven of the tested samples (all samples except for Qat.16). These seven samples were processed further for sequencing. For DL2, amplification success was lower and a final set of three replicates of each of five samples (Qat.01, 04, 13, 15, and 19. Supplementary Table S1) was obtained.

The final amplicons were pooled for library building such that replicate number one of each sample was added to one pool, replicate number two to another pool and so on, resulting in six pools for DL1 and three pools for DL2. Each pool thus contained only one PCR replicate of each of the seven (DL1) or five (DL2) included samples. This was done so that; i) each replicate of a sample would receive a different library index, providing a way of double-checking identifications of individual PCRs based on primer tags, and ii) samples were spread evenly across libraries, as sequencing success and depth can vary from library to library. Pooling was done using equal volumes, as band strengths of positive replicates were similar. From each pool, 30 μL was run on a 2% agarose gel stained with GelRedTM (Biotium Inc.), and the band corresponding to the expected fragment length of was cut out using a

sterile scalpel blade. DNA extraction of the gel cuts was done using the QIAQuick Gel Extraction Kit (Qiagen) following the manufacturer's protocol, except that elution was done with 40 μ L EB buffer which was incubated in the spin-column for 10 minutes at 37 °C. Importantly, while the DL2 primers offer a high resolution and were useful for the current study, we recommend that new markers are developed for future population level eDNA studies of the whale shark, due to the relatively low amplification success and difficulties with sequencing, including incomplete sequences and a low output. However, we cannot conclude that these issues are related to the DL2 primers.

Library building and Illumina sequencing

Sequencing libraries were prepared using the NEBNext[®] DNA Library Prep Master Mix Set for 454[®] (New England Biolabs Inc.), with the following modifications of the protocol. For the end-repair reaction, samples were incubated at 12 °C for 20 minutes, and 15 minutes at 37 °C. Elution was done in 17 μ L EB buffer during purification after end-repair, and in 23 μ L after adaptor ligation, and samples were incubated with the elution buffer for 15 minutes at 37 °C. The concentration of adaptors was adjusted to the concentration of target PCR product, which was determined on an Agilent 2100 Bioanalyzer (Agilent Technologies). After fill-in, samples were incubated at 65 °C for 20 minutes, and at 80 °C for 20 minutes.

Index PCR was performed using 30 μ L ddH₂O, 10 μ L DNA, 5 μ L 10X High Fidelity buffer, 1.5 μ L MgSO₄ (50 mM), 1 μ L dNTPs (2.5 mM), 1 μ L of each primer (10 μ M), and 0.5 μ L Invitrogen[™] Platinum[™] *Taq* DNA polymerase High Fidelity (5 U/ μ L) (ThermoFisher Scientific). Thermocycling was set to 94 °C for 1 minutes, 10 cycles (DL1) or 15 cycles (DL2) of; 94 °C for 30 s, 55 °C for 30 s, and 68 °C for 30 s, and finally 5 minutes at 68 °C. Libraries were sent to MacroGen Europe for sequencing in multiplex (250 bp paired-end for DL1 and 300 bp paired-end for DL2) on the Illumina MiSeq platform. A spike-in of PhiX was used to increase complexity in the runs.

Bioinformatic analysis of Illumina data

Sequences were analyzed using OBITools¹¹. For DL1, paired reads were assembled with the “*illumina-paired-end*” command using a minimum alignment score of 40. DL2 sequences were found to drop significantly in quality after the first ~200 bp and were therefore joined end to end, as they could not be aligned correctly. In order to remove the low quality sequence in the middle part of joined reads,

two conserved motifs; “TTTCGT” (position 216–221 in DL2) and “ACTCATTAAT” (position 630–639 in DL2), at either end of the problematic middle stretch were used to remove this region from the sequences using a custom python script (written by Emil Vissing. See “Bioinformatic code”). For both DL1 and DL2, only sequences with a count of at least 10 were retained. The number of initial paired-end raw reads was 12,787,089 for DL1 and 47,034,448 for DL2, and the number of final reads after all trimming was 3,325,512 for DL1 and 445,866 for DL2. Final reads after trimming were 371 bp (DL1) and 313–347 bp (DL2). The sequences were assigned in OBITools by searching against a local ecoPCR database containing the reference whale shark DNA sequences of DL1 and DL2 (see “Whale shark DNA reference database”). Of these, 2,007,666 reads of DL1 and 381,797 reads of DL2 had a 100% match to a known haplotype in the ecoPCR database. Final sequences that did not display a characteristic error pattern of a pronounced relationship between read count and match (99%–100%) to the given haplotype (Supplementary Figure S3) were flagged as “doubtful”. This included one doubtful haplotype for DL1 (DL1-H) and eight doubtful haplotypes for DL2 (DL2-P to DL2-X) from the final data. The authenticity of these doubtful haplotypes was finally evaluated through comprehensive error modelling over a range of possible error rates (see “False positive estimation and final data cleaning: Dealing with sequencing and PCR errors”).

Mock sequencing control

A mock sample (positive control) of DNA from six known whale shark individuals in varying relative concentrations (ws_105: 1, ws_26: 1:2, ws_29: 1:4, ws_24: 1:10, ws_89: 1:100, ws_93: 1:1000.

Supplementary Table S1) was amplified and prepared as above and sequenced along with the water samples. DNA concentrations in the tissue extractions were estimated with qPCR using the RhitypCB qPCR assay described under “Quantitative PCR (qPCR): Whale shark vs. mackerel tuna eDNA” (data not shown). We obtained few reads from the mock sample, but the relationship between haplotypes was retained overall (Supplementary Figure S4), except for a single haplotype (DL2-F) appearing at high frequency, which was not added to the mock. This unexpected haplotype could stem from i) PCR and sequencing errors in an authentic added haplotype, or ii) contamination of the tissue extractions used for the mock sample, originating anywhere from the field to stages of the laboratory workflow preceding the tagged PCR. To investigate this further, and in an attempt to obtain a larger data set for the mock sample, the sample was resequenced (in a new library containing only this sample) together

with a new mock sample, prepared in the same way as the first one. The unexpected haplotype, DL2-F, also appeared in the resequenced mock sample and in the newly prepared mock sample, again at a high frequency. Based on i) negative PCR and extraction blanks (see “Sequencing of negative controls”), ii) the consistent appearance of the unexpected haplotype in the mocks, and iii) the high frequencies of the unexpected haplotype, which exceeds any frequency expected from errors (see “False positive estimation and final data cleaning: Dealing with sequencing and PCR errors”), we conclude that the most plausible source of this sequence is a cross-contamination between tissue samples in the field. Some level of cross-contamination is inevitable when dealing with a large number of tissue samples. However, we believe that contamination between tissue samples is unlikely to have impacted the conclusions of the study, as a sample containing high levels of non-target DNA would have displayed overlapping peaks in the Sanger sequencing data, and should therefore have been detected during trimming and quality checking of the sequences. The resequenced as well as the new mock sample both displayed a positive relationship between the relative amount of DNA added of a haplotype and the number of reads obtained matching that haplotype (Supplementary Figure S4). The DL2 haplotype corresponding to the tissue extraction that had been diluted the most (1,000 times) was not retrieved from any of the mock samples.

Sequencing of negative controls

Negative PCR controls (one for each tag combination used for the water samples) and extraction blanks were pooled and built into two libraries (DL1 and DL2) as described above, and were sequenced along with the new mock sample and the new library built on the original mock sample (see “Mock sequencing control”). Sequencing was done on the Illumina MiSeq platform (250 bp paired-end) at the National High-throughput DNA Sequencing Center. The original PCR blanks for DL1 were unfortunately lost, and new blanks were therefore prepared in the same way and using the same primer aliquots as when the original PCRs were set up. To increase sequence complexity, a library built on DNA amplified with generic metazoan primers¹² was added to the sequencing pool. 2,513,933 aligned sequences (from 5,045,684 raw paired-end reads) were obtained for the DL1 PCR and extraction blanks. Of these 2,128,147 sequences were correctly tagged, but none were a 100% match to the whale shark reference database (in fact all sequences were less than 50% similar to a known whale shark haplotype). For the DL2 negative controls, 9,586,552 raw reads corresponding to 4,793,276 joined

reads were obtained, of which 3,585,780 were correctly tagged. Similarly to DL1, none of the DL2 sequences matched a haplotype in the database (maximum 70% identity).

False positive estimation and final data cleaning: Dealing with sequencing and PCR errors

Sequencing errors can potentially lead to false positives¹³ especially in studies where the variable distance in the genomic barcode can be as short as one mutation. However, the actual proportion of false positives is often neglected and false positives are instead accounted for by setting a certain number of reads as the minimum needed for a sequence to be accepted as a true positive. To investigate the proportion of false positives generated by sequencing errors we performed an *in silico* simulation inferring sequencing errors to the whale shark haplotypes from the true dataset. Firstly, using bowtie2¹⁴, we aligned all PhiX reads from the spike-in to our libraries to the complete genome of Enterobacteria phage PhiX174 (downloaded from NCBI: [gi|9626372|ref|NC_001422.1|](https://www.ncbi.nlm.nih.gov/nuclorf/gi|9626372|ref|NC_001422.1|)). The alignment was parsed to Tablet¹⁵ for visual exploration and calculation of the mismatch percentage (0.3%), which we define as the sequencing error rate. The distribution of the sequencing error rate per position over the read for both mate1 and mate2 was computed, visualized, and used to determine the closest ‘relative abundance model’ setting for the modelled dataset in Grinder¹⁶. We generated a fasta file for each haplotype from the true dataset containing 100,000 copies of the haplotype sequence, roughly equivalent to the highest read counts found in a single library (DL1: 61,351 reads of DL1-A and DL2: 135,309 reads of DL2-A). The fasta files were parsed to Grinder for sequencing simulation inferring a ‘uniform’ relative abundance model, limiting the error type to ‘substitutions’, and randomly inferring a sequencing ‘error rate of 0.3%’. The simulation was set up with 5 iterations for all haplotypes. Each sequence in the simulated fasta files was then aligned to the complete whale shark reference database and taxonomically assigned using the method applied to the true dataset, but skipping initial trimming and quality filtering (thus starting with the obiuniq step. See “Bioinformatic analysis of Illumina data”). From this we calculated for each DL1 and DL2 haplotype the likelihood of false positives (see Supplementary Figure S5), which we define as the number of sequences falsely assigned to that haplotype. Interestingly, the results showed an increased likelihood for certain haplotypes to arise as false positives (up to fourfold), including DL1-A, DL2-A, DL2-B, and DL2-C. These differences in likelihood can be explained by the genetic distances between the haplotypes and the fact that multiple closely related haplotypes can have a high affinity for becoming the same

phylogenetically more basal haplotype (see Supplementary Figures S6 and S7). The large differences in false positive rates observed between haplotypes underline the importance for analyses of these rates when interpreting eDNA metabarcoding data.

To explore the effect of increasing sequencing error rates we repeated the above iterations using a range of error rates (0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.2, 1.4 and 1.6%), and found that rates above 0.3% yield fewer false positives.

The cut-off generated by the 0.3% sequencing error model for each haplotype was used to remove haplotype sequences within each PCR replicate. This resulted in the cleaning and exclusion of three of the “flagged” haplotypes (see “Bioinformatic analysis of Illumina data”), namely DL2-S, DL2-V, and DL2-X.

An additional cleaning was performed to remove errors generated during PCR amplification. In the mock sample (see “Mock sequencing control”), a number of known haplotypes that had not been added to the sample, but were closely related to the added haplotypes (1 or 2 nucleotide differences) appeared at a low count. While these spurious sequences could stem from low-level contamination, it is also quite possible that they originated from PCR errors. The most abundant of these low count sequences appeared at a frequency of 1.3% compared to the total count of all added haplotypes differing by 1 or 2 nucleotides from the spurious sequence. This frequency was comparable to the average frequency of sequences that were unknown (no 100% matches to the reference database) but differed by one nucleotide from an added haplotype (range and median for added haplotypes in: original DL2 mock: 0.03–3.41 (0.36); new DL2 mock: 0.68–10.51 (2.68); and new DL1 mock: 0.03–0.69 (0.63)). Based on this, any haplotype in the sample PCRs occurring at a frequency of less than 1.3% of the total count of closely related (1–2 nucleotide differences) haplotypes were removed. This cleaning removed three additional haplotypes from the data compared to the sequencing error cutoff, namely DL1-H, DL2-U and DL2-T.

The total amount of sequences removed from the final data set, based on both PCR and sequencing error rates, made up 8,753 (0.44%) and 12,715 (3.33%) reads for DL1 and DL2, respectively.

The final number of haplotypes retained in the data for all analyses were thus 7 for DL1 (A–G) and 18 for DL2 (A–R) (Figures 1C–D, Supplementary Figure S2).

We also performed a data analysis and cleaning using only the sequences of the six individuals in the mock sample. Cleaning was done based on the error rate observed in the mock sample (1.3%),

assuming only that the most abundant sequence from a PCR was authentic, and requiring presence in at least two PCRs (Supplementary Figures S8 and S9). All seven DL1 haplotypes were retained, while 10 of the 18 DL2 haplotypes identified using the reference database were retained. Several unknown putative haplotypes were retained (nine for DL1; DL1xa-DL1xi and six for DL2; DL2xa-DL2xf). N_f was estimated at 63,400 females (95% CI: 38,525–162,899) based on 100 individuals.

Population comparisons, mutation rate and effective population size

The mean read frequencies of each DL1 and DL2 haplotype in the water samples were compared to the corresponding haplotype frequencies in local tissue samples, and to samples collected by Castro et al. (2012) and Vignaud et al. (2014) in the Western Indian, Eastern Indian, Northwest Pacific, Northeast Pacific and Atlantic Oceans. These comparisons were done via boxplots (Figures 1C and 1D) and a principal component analysis performed with “prcomp” in R v. 2.15.2¹⁷ (Figure 1E). In order to make the read frequencies from eDNA more comparable to frequencies based on counts of individuals, the frequencies were scaled to a total of 100 individuals. Haplotypes that appeared in less than 1% of the total reads in the final sequences, but were deemed authentic based on error modelling (see “False positive estimation and final data cleaning: Dealing with sequencing and PCR errors”), were scaled to represent one individual; the minimum possible number of source individuals for a haplotype. The scores of the first principal component for each geographical location were compared to the distance to the Gulf by the shortest sea route (estimated using the distance measurement tool in Google Earth (Map data: © 2016 Google)) using Pearson’s correlation test in R v. 2.15.2¹⁷. The Al Shaheen aggregation was also compared to other aggregations using F_{st} values calculated in Arlequin v. 3.5.2¹⁸.

To estimate the mutation rate in the DL2 target region of the whale shark mitogenome we aligned a whale shark DL2 haplotype (DL2-D, e.g. GenBank: EU182401) to the corresponding region in forty species from Orectolobiformes and Carcharhiniformes, using Geneious v. 7. Three species of Hemiscylliidae; *Chiloscyllium punctatum* (GenBank: NC016686), *C. plagiosum* (GenBank: JX162601) and *C. griseum* (GenBank: NC017882), were included to represent evolutionary close relatives to the whale shark¹⁹. While nuclear and mitochondrial DNA markers in combination suggest that Stegostomatidae is a close sister group to Rhincodontidae²⁰, Hemiscylliidae (represented here by *Chiloscyllium*) appear to be the closest relatives of the whale shark, when relationship is inferred from mitochondrial DNA alone (Alam et al. 2014). The whale shark displayed a long insert in the alignment,

which was absent from all other species of Carcharhiniformes and Orectolobiformes. As this insert, which may stem from a single mutation, could have a disproportionately large impact on the analysis, it was removed from the alignment. The final DL2 alignment comprised 290 bp. We used DAMBE v. 5.6.14²¹ to estimate the proportion of invariant sites and reached an estimate of 0.111 based on the neighbor-joining method. The level of substitution saturation was then estimated by Xia et al.'s test (2003) using the estimated proportion of invariant sites and including only fully resolved sites. The ISS value was significantly lower than the ISS.c value, indicating a low level of saturation, and the alignment was therefore assumed to be informative.

The alignment was tested in jModelTest 2.1.7 v 20150530²², where likelihood scores were computed for 11 substitution schemes, including the possibility for equal or unequal base frequencies, invariable sites, and rate variation among sites, giving a total of 88 tested models. The models were tested using a maximum likelihood base tree and the best fit of either an NNI (nearest neighbor interchange)^{23,24} or SPR (subtree pruning and regrafting)²⁵ tree search. Evaluation of the models was done using the Bayesian information criterion, which indicated that a TrN93 model²⁶ with the Gamma model of heterogeneity was best described the data. In BEAUti v1.8.2²⁷, Orectolobiformes and Carcharhiniformes were each restricted as monophyletic, and were also set as being monophyletic together. The taxon set comprising Hemiscyllidae and Rhincodontidae was also restricted as monophyletic. An uncorrelated lognormal relaxed clock was used to allow time-dependency to vary across branches²⁸, and to obtain a better estimate of the timescale for divergence²⁹ and the mutation rate. The tree prior was set to “speciation” following the Birth-Death Model with incomplete sampling³⁰ and an exponential distribution of the prior sample probability, with the initial and mean values set to 0.1, to reflect that the species included in the analysis likely represent less than 10% of all extant Orectolobiformes and Carcharhiniformes. Prior distributions of the time to most recent ancestor (TMRCA) were modeled using an exponential distribution for all taxon sets, as this distribution is considered suitable for fossil calibration when information is too limited to set a lognormal distribution with confidence³¹. Three nodes were calibrated using first appearance dates (FADs). These three nodes were defined by the taxon set including Hemiscyllidae and the whale shark, the taxon set including the Carcharhiniformes and the taxon set including both Carcharhiniformes and Orectolobiformes – i.e. the root of the topology. The TMRCA prior for the taxon set including Hemiscyllidae and the whale shark was calibrated using an offset of 125 mio. years, representing the FAD for Hemiscyllidae based on †*Chiloscyllium* sp. from

the Aptian stage³², and a mean of 16.9 mio. years. A 95% confidence interval (CI) for the prior distribution of 125–175.6 mio. years was thereby obtained, the latter being the FAD for Orectolobiformes, based on †*Palaeobrachaelurus bedfordensis* from the Aalenian period³³.

The TMRCA prior for Carchariniformes was calibrated using an offset of 167.7 mio. years, the FAD of Scyliorhinidae, which is the oldest of the six included families. This FAD is based on †*Palaeoscyllium tenuidens* and †*Eypea leesi* from the Bathonian stage³⁴. The distribution mean was set to 2.6 mio. years to obtain a 95% CI for the prior distribution of 167.7–175.6 mio. years, where the 175.6 mio. years represents the FAD for Orectolobiformes.

Lastly, the TMRCA prior for the taxon set comprising Carchariniformes and Orectolobiformes – the root of the tree – was calibrated using an offset of 175.6 mio. years and a mean of 31.5 mio. years, to obtain a 95% CI for the prior distribution of 175.6–270 mio. years, where the 270 mio. years is based on the oldest known neoselachian remains, a specimen of †‘*Synechodus*’ *antiquus* (uncertain family), from the Artinskian stage³⁵.

An exponential distribution was used to describe the expected probability distribution of the uncorrelated lognormal relaxed clock mean. The initial and mean clock rate prior was set to 0.005 per mio. years, giving a 95% CI of 0–0.015 substitutions per site per mio. years, to assure that sampled substitution rates encompassed the range of substitution rates observed in the mtDNA of sharks, which exhibit some of the lowest rates of divergence in mtDNA measured in vertebrates³⁶. The MCMC analysis was run in BEAST v1.8.2.²⁷ for 40 mio. generations with parameters being logged every 1000 generations. TRACER v1.6.0³⁷ was used to examine the resulting log files and inspect the effective sample sizes and likelihood stabilization of the model parameters. Effective sample sizes (ESS) were all above 200. The resulting trees were summarized in TreeAnnotator v1.8.2 (BEAST package) using a burn-in of the first four mio. states, mean node heights, and a posterior probability limit of zero (such that posterior summaries were calculated for all nodes), and was then visualized in FigTree v1.4.2³⁸. To tests the effect of the priors³⁹ we ran the Bayesian analysis on an empty alignment created in BEAUti using the same priors as when nucleotide data was included. For all parameters, excluding likelihood parameters, ESS values were above 200, and marginal probability distributions were unimodal, indicating that there was no cross-influence of priors.

Based on the nucleotide diversity (π)⁴⁰ of the scaled haplotype frequencies from eDNA and the estimated mutation rate (μ) for DL1 and DL2, an effective female population size (N_f) was estimated as

$N_f = \pi/2\mu$ ^{41–44}. This equation assumes that the sampled locus contains an infinite number of sites, and therefore that no two mutations ever occur at the same site⁴².

In addition, it is assumed that sequences have diverged only by neutral mutation in an isolated population of constant effective size. A mutation rate of 0.1% (95% CI: 0.04% to 0.16%) per mio. years was estimated for DL2 (Supplementary Figure S10). The eDNA from the water samples returned a π of 0.00358 (s.d.: 0.00048) for DL2, which gave an estimated N_f of 71,600 (95% CI: 43,618–183,526) females. Tissue samples returned a π of 0.692 (s.d.: 0.00042) for DL2, which gave an estimated N_f of 138,400 (95% CI: 85,087–351,654) females with the estimated mutation rate for the region (Supplementary Figure S10). The male:female ratio of the tissue-sampled whale sharks is approximately 2:1 (Supplementary Table S1), which is more equal than what has been observed for most other aggregations, which are largely dominated by juvenile males^{3,45,46}.

Quantitative PCR (qPCR): Whale shark vs. mackerel tuna eDNA

qPCR analyses of whale shark and mackerel tuna eDNA were performed on all samples taken between the 27th of May 2013 and the 20th of May 2014 (17 samples in total. Supplementary Table S1). Prior to qPCR analysis, equal aliquots of the extracted subsamples from each 3*500 mL sample set, were pooled into a single sample. qPCR was performed on the pooled samples using species-specific primer sets and TaqMan hydrolysis probes (with 6-carboxyfluorescein (FAM) as the fluorophore and Black Hole Quencher (BHQ1) as the quencher) developed for this study. The whale shark assay (RhitypCBR: 5'-CCTGTTGGGTTGTTTGAACC-3', RhitypCBL: 5'-TACCCGCTTCTTTGCATTTC-3', and RhitypCB.probe: 5'-FAM-CTTTCTCTTGCCATTCTAATTGCAGA-BHQ1-3') and the mackerel tuna assay (EutaffCBR: 5'-ATTTGAATTCAGCCCGATTG-3', EutaffCBL: 5'-TCTTCGCCTTCCACTTCCTA-3', and EutaffCB.probe: 5'-FAM-TCCCCTTCGTTATCGCGGCC-BHQ1-3') target a 105 bp and 110 bp product of the mitochondrial cytochrome b (*CYTB*) gene, respectively. Amplifications were performed on a Stratagene Mx3005P in 25 μ L reactions of 10 μ L TaqMan® Environmental Master Mix 2.0 (Life Technologies), 10 μ L ddH₂O, 1 μ L of each primer (10 μ M), 1 μ L probe (2.5 μ M) and 5 μ L of extracted eDNA from filtered water samples. Thermocycling parameters were as follows: 5 minutes at 50 °C, 10 minutes at 95 °C, and 50 cycles of 30 s at 95 °C and 2 minutes at 60 °C. Four reactions were performed for each sample. Eight negative PCR controls, and standard dilutions of 10-10⁷ copies per μ L in three replicates each, were included in every qPCR run.

Standards were prepared from a purified PCR product amplified from tissue derived DNA with the primers used for qPCR. Amplifications were performed using the same reaction mix as for the whale shark tissue samples (see “Whale shark DNA reference database”). Thermocycling parameters were: 95°C for 5 minutes, 40 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 30 s, and finally 72°C for 2 minutes. PCR products were checked on 2% agarose gels stained with GelRed™ (Biotium Inc.), and were then purified using the MinElute Purification kit (Qiagen). Elution was done in 50 µL EB buffer which was incubated for 10 minutes at 37 °C in the spin column before centrifugation. The concentration of amplified DNA was measured on a Qubit 2.0 Fluorometer (Life Technologies), and serial dilutions were then prepared. A new standard series was prepared for each qPCR run.

Extraction controls were tested for amplification with both the mackerel tuna and the whale shark primer/probe set, using the same reaction mix and cycling settings as described above. The qPCR system for whale shark was tested for negative amplification on DNA from zebra shark (*Stegostoma fasciatum*, ZMUC P06273, no voucher), milk shark (*Rhizoprionodon acutus*, ZMUC P06274-76, no voucher, photo available) and blacktip reef shark (*Caraarhirus melanopterus*, ZMUC P06277, no voucher, photo available)). Institutional abbreviations associated with voucher specimens follow Fricke and Eschmeyer (2015). The qPCR system for mackerel tuna was tested on kingfish (*Scomberomorus commerson*, ZMUC P74254, no voucher, photo available)), yellowfin tuna (*Thunnus albacares* ZMUC P74256), Atlantic bonito (*Sarda sarda*, ZMUC P74231), chub mackerel (*Scomber japonicus*, ZMUC P74255, no voucher, photo available)), and Atlantic mackerel (*Scomber scombrus*, ZMUC P74261). Thermocycling settings were as described above.

Cloning and sequencing were done to verify qPCR results. Due to limited cloning success when using qPCR products directly, purified qPCR products were PCR amplified and purified a second time, prior to cloning and sequencing. PCR was done as described for the qPCR standards. DNA presence and length were checked on 2% agarose gel stained with GelRed™ (Biotium Inc.), followed by purification as above. PCR products were cloned using the TOPO® TA Cloning® Kit for Sequencing (Life Technologies) and commercially sequenced at Macrogen Europe. Sequences were trimmed of primers, quality checked, and BLAST searched against the NCBI “nt” database using Geneious 6.1.7 (Biomatters Ltd.).

Seventeen water samples were used for analysis of whale shark and mackerel tuna eDNA concentrations. Positive amplification of whale shark eDNA was detected in all samples from May

2013, and in samples 13, 15, and 16 from May 2014 (Supplementary Table S1). Starting quantities of DNA in positive samples varied from less than ten to thousands of target copies per μL of extracted sample, corresponding to a range of hundreds to hundreds of thousands of DNA copies per liter of sampled water (in the experiment, one DNA copy per μL of extracted sample corresponds to 220 DNA copies per liter of water). For both years, the highest quantities of DNA were found where whale sharks were observed (samples Qat.01, 04 and 06 for 2013 and samples Qat.13, 15 and 16 for 2014.

Supplementary Table S1), with average concentrations varying between $\sim 10,000$ and $\sim 540,000$ copies/L. DNA from whale shark was also found in samples taken where no sharks were observed (samples Qat.02, 03, 05, 07–10, May 2013. Supplementary Table S1), with an average copy number per liter of ~ 100 to $\sim 7,000$ copies. The number of positive qPCR replicates per sample varied, with samples Qat.01 to Qat.06 and samples Qat.13, 15, and 16 amplifying in all replicates, while samples Qat.07 to Qat.10 amplified in two to three replicates.

Positive amplification of mackerel tuna eDNA was detected in all samples, with starting quantities of DNA varying from a few to tens of thousands of target copies per μL of extracted sample, corresponding to hundreds to millions of target copies per liter of sampled water. The highest concentrations of mackerel tuna DNA were found where whale sharks were observed (samples Qat.01, 04, 06, 13, 15 and 16. Supplementary Table S1), with concentrations ranging between $\sim 320,000$ and $24,000,000$ copies/L. In the remaining samples, DNA quantities of ~ 30 to $\sim 190,000$ copies/L were found.

The efficiency of standard curves was 92% and 98%, respectively, for the two qPCRs with mackerel tuna primers. For the qPCRs targeting the whale shark, standard curve efficiencies were 93% and 87%.

Gel electrophoresis verified the length of the amplified products, and sequencing confirmed that amplicons were a 100% match to whale shark and mackerel tuna, respectively. No amplification was detected in the extraction controls or negative qPCR controls, and the zebra shark and blacktip reef shark also showed no amplification with the whale shark qPCR system. The milk shark did show positive amplification with the whale shark qPCR assay, but reached the cycling threshold (C_t) at a much later point than the whale shark ($C_t = 22.60$ for the whale shark and 36.06 for the milk shark).

Statistical analysis was performed using R v. 2.15.2. The average eDNA concentration (copies/L) was calculated for each 3×500 mL sample, based on the four qPCR replicates. Samples were grouped based on whether whale sharks were observed during sample collection or not, and a Wilcoxon rank sum test

was used to test whether the amount of eDNA found in samples differed between the two groups. This test was selected as sampling was based on whale shark observations and not done randomly. Furthermore, visual inspection of the data did not suggest a normal distribution. The correlation between the amount of mackerel tuna and whale shark eDNA was tested using Pearson's correlation test, and a linear model was fitted to the data (Figure 1F). Data was log transformed, and the data for 2013 and 2014 were merged and analyzed as one data set in the statistical analysis. Replicates showing no amplification were treated as zero when calculating the average concentrations, as advocated by Ellison et al. (2006).

Seawater eDNA degradation experiment

In order to obtain an estimate of the degradation rate of whale shark eDNA in warm, saline waters, an experiment was set up using the 6*30 L water samples collected in May 2014 (Supplementary Table S1). In Doha, the samples were transferred to two large plastic buckets (3*30 L in each). The uncovered buckets were placed outdoors, one in direct sunlight, and the other in permanent shade. For the next eight days, a 500 mL sample was taken regularly from each of the buckets following the sampling procedure described above. On the first two days, samples were taken at short intervals of approximately two to six hours. On the third day, three samples were taken, and for the remainder of the experiment, a sample was taken every morning and evening. In total, 22 samples were taken from each bucket. Immediately before sampling, the water in the buckets was mixed thoroughly, and on at least two sampling events per day, the water temperature was measured. Water temperature measurements in the experiment ranged from 29 °C to 43 °C in the sunlight treatment and from 29 °C to 40 °C in the shade treatment, with mean temperatures being 36.0 °C and 35.7 °C, respectively (median: 36 °C and 35.5 °C). After sampling, Sterivex filters were immediately stored at -18 °C until further processing.

Quantitative PCR analysis of the samples from the degradation experiment was performed using the whale shark primer/probe assay described above, and employing the same reaction mix and thermocycling settings used for the field samples. Two qPCR runs – one for each treatment (sunlight and shade) – were performed on consecutive days, using a standard dilution series of 2, 10, 10², 10³, 10⁴ and 10⁵ copies/μL, which was prepared as described earlier. All samples and standards were run in three replicates, except for the two standards of lowest concentration, which were run in four replicates

to reduce stochasticity. In both qPCR runs, sample Qat.13 was included as a reference, representing eDNA concentrations in the field, in the area where the samples for the experiment were collected. This sample was taken approximately two hours before the 6*30 L samples were taken, but within the same aggregation of sharks. Six negative PCR controls were included in each run.

Similarly to the whale shark versus mackerel tuna qPCR experiment (see “Quantitative PCR (qPCR): Whale shark vs. mackerel tuna eDNA”), the average eDNA concentration (copies/L) for each 500 mL sample was calculated based on the three qPCR replicates. A Spearman’s rank sum correlation test was performed to test for the effect of time on eDNA concentration. For each treatment an exponential decay model was then fitted to the data using the “nls” function in R. The model was of the form $N_t = N_0 e^{-\beta t}$, where N_t represents the eDNA concentration at time = t hours after field sampling, N_0 is the initial DNA concentration at the time of field sampling, and β is the decay constant. This model was chosen, as DNA degradation has been found to conform well to an exponential decay model in studies on aDNA⁴⁷, iDNA⁴⁸, and eDNA from water⁴⁹. However, as this model did not fit the data well, based on inspection of the regression curve, the time variable was log transformed, giving a model of the form $N_t = N_0 e^{-\beta \log(t)}$ (Supplementary Figure S11). The residuals of the nls models were tested for deviations from normality using the Shapiro-Wilk normality test and for randomness with the Runs test. The former test showed a significant deviation from normality for the shade treatment ($W = 0.77$, $p < 0.001$). The Runs test (tests the null hypothesis that the residuals increase or decrease in value randomly) gave no indication of non-randomness for either treatment (Standard Normal = -2.55, $p = 0.01$ for the light treatment and Standard Normal = -1.65, $p = 0.1$ for the shade treatment).

Since statistical testing for a difference between treatments (sunlight/shade) was not appropriate due to the lack of replication, scatter plots for each of the two treatments were overlaid and inspected visually for any obvious differences in degradation pattern.

Efficiencies of qPCR standard curves were 93% ($R^2 = 0.997$) and 90% ($R^2 = 0.985$) for the sunlight and shade treatment, respectively. Whale shark eDNA concentration decreased rapidly over time in both treatments. From starting concentrations of ~100,000 copies/L ($t = 3.8$ hours after the water was sampled in the field), concentrations dropped an order of magnitude in the first 48 hours of the experiment. Spearman’s rank sum correlation test confirmed the strong decrease in whale shark eDNA concentration over time in both treatments ($p = 5.7 \times 10^{-15}$ and $p = 8.5 \times 10^{-12}$, respectively). After 4.9 days, the concentration of whale shark eDNA was below 20 copies/ μ L (4400 copies/L) in both

treatments. However, whale shark eDNA remained detectable throughout the experiment. In the final sample ($t = 7.8$ days) eDNA concentrations were estimated to be <2 copies/ μL (<440 copies/L) and <15 copies/ μL (~ 2600 copies/L) in the sunlight and shade treatments, respectively. The exponential decay model fit the data well for both treatments based on the R-squared ($R^2 = 0.95$ for the sunlight treatment, and $R^2 = 0.96$ for the shade treatment). The model of the sunlight data predicted a concentration at $t = 0$ of $\sim 255,000$ copies/L, while for the shade treatment model, the prediction for $t = 0$ was $\sim 405,000$ copies/L. The estimated decay constant of the model was $\beta = 0.89$ for sunshine. A similar value of $\beta = 1.00$ was estimated for shade. There was no apparent difference in degradation pattern between the two treatments based on the overlaid scatterplots (Supplementary Figure S11).

Cost and effort analysis

The approximate cost and effort spent to obtain the d-loop sequences from tissue samples and water samples, respectively, was estimated and compared. Only samples that were used in the final data analyses were included in the estimates, and the cost of primer design (for eDNA) and primer testing (tissue and eDNA) was not included. The cost and time spent on transport by boat to and from the study site, was assessed for both tissue and water sampling. For tissue, the resources spent on biopsy darting, photography, and photo identification, was added to this expense, while in the case of eDNA, costs of sampling including filters, syringes etc. was estimated and included in the expenses. Finally, the cost of lab work and bioinformatics analyses to obtain sequences from tissue and eDNA, respectively, was added to the estimates. For estimating salary expenses, we used an hourly salary of €25, which is equivalent to the salary of a scientific assistant in our lab. The final cost estimates were €15,171 for eDNA and €19,255 for tissue.

Bioinformatic code

Custom python script for removing middle stretch of DL2 sequences:

```
{
import re, time

master_string =

print("\nMake sure the python file is in the same folder with the data file")
```

```

filename = input("\n\nPlease input file name (only the full file name - not path): ")

if (filename == "1"):
    filename = 'filename'

try:
    with open(filename, 'r') as datafile:
        data = datafile.readlines()

        for each in data:
            master_string += each

        with open(filename.replace(' ', '_')[:-6] + "_modified_" + time.strftime("%d%b-%H%M%p") +
".fasta", "w") as myfile:
            for line in master_string.split('>'):
                myfile.write(">" + (re.sub(r'(ttcgt).*(actcataat)', r'\1\2', line, 0, flags=re.DOTALL)))

except FileNotFoundError:
    print("Oops! file not found!")
    exit()
}

```

Permissions

Permissions for obtaining samples for this study is given in Memorandum of Understanding between The Ministry of Environment, State of Qatar, and Maersk Oil Research and Technology Centre, Qatar (no. MoU-MoE-MORTC-2012-03-12).

SUPPLEMENTARY REFERENCES

1. John, V. C., Coles, S. L. & Abozed, A. I. Seasonal cycles of temperature, salinity and water masses of the western Arabian Gulf. *Oceanol. Acta* **13**, 273–281 (1990).
2. Sheppard, C. *et al.* The Gulf: a young sea in decline. *Mar. Pollut. Bull.* **60**, 13–38 (2010).
3. Robinson, D. P. *et al.* Whale Sharks, *Rhincodon typus*, Aggregate around Offshore Platforms in Qatari Waters of the Arabian Gulf to Feed on Fish Spawn. *PLoS ONE* **8**, e58255 (2013).
4. Carpenter, K. E. *Living marine resources of Kuwait, Eastern Saudi Arabia, Bahrain, Qatar, and the United Arab Emirates.* (Food & Agriculture Org., 1997).
5. Blegvad, H. & Løppenthin, B. Fishes of the Iranian Gulf. (1944).
6. Heyman, W. D., Graham, R. T., Kjerfve, B. & Johannes, R. E. Whale sharks *Rhincodon typus* aggregate to feed on fish spawn in Belize. *Mar. Ecol. Prog. Ser.* **215**, 275–282 (2001).
7. de la Parra Venegas, R. *et al.* An Unprecedented Aggregation of Whale Sharks, *Rhincodon typus*, in Mexican Coastal Waters of the Caribbean Sea. *PLoS ONE* **6**, e18994 (2011).
8. Castro, A. L. F. *et al.* Population genetic structure of Earth's largest fish, the whale shark (*Rhincodon typus*). *Mol. Ecol.* **16**, 5183–5192 (2007).
9. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452 (2009).
10. Meyer, M., Stenzel, U. & Hofreiter, M. Parallel tagged sequencing on the 454 platform. *Nat. Protoc.* **3**, 267–278 (2008).
11. Boyer, F. *et al.* obitools: a unix-inspired software package for DNA metabarcoding. *Mol. Ecol. Resour.* **16**, 176–182 (2016).
12. FOLMER, O. DNA primers for amplification of mitochondrial cytochrome oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol* **3**, 294–299 (1994).
13. Ficitola, G. F. *et al.* Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Mol. Ecol. Resour.* **15**, 543–556 (2015).
14. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
15. Milne, I. *et al.* Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* bbs012 (2012).

16. Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P. & Tyson, G. W. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* **40**, e94–e94 (2012).
17. Team, R. C. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 2013. (ISBN 3-900051-07-0, 2014).
18. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
19. Alam, M. T., Petit III, R. A., Read, T. D. & Dove, A. D. M. The complete mitochondrial genome sequence of the world's largest fish, the whale shark (*Rhincodon typus*), and its comparison with those of related shark species. *Gene* **539**, 44–49 (2014).
20. Vélez-Zuazo, X. & Agnarsson, I. Shark tales: a molecular species-level phylogeny of sharks (Selachimorpha, Chondrichthyes). *Mol. Phylogenet. Evol.* **58**, 207–217 (2011).
21. Xia, X. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. *Mol. Biol. Evol.* **30**, 1720–1728 (2013).
22. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772–772 (2012).
23. Robinson, D. F. Comparison of labeled trees with valency three. *J. Comb. Theory Ser. B* **11**, 105–119 (1971).
24. Moore, G. W., Goodman, M. & Barnabas, J. An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. *J. Theor. Biol.* **38**, 423–457 (1973).
25. Swofford, D. L. & Begle, D. P. *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1*, March 1993. (Center for Biodiversity, Illinois Natural History Survey, 1993).
26. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526 (1993).
27. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
28. Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**, e88 (2006).
29. Duchêne, S., Lanfear, R. & Ho, S. Y. The impact of calibration and clock-model choice on molecular estimates of divergence times. *Mol. Phylogenet. Evol.* **78**, 277–289 (2014).

30. Stadler, T. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *J. Theor. Biol.* **261**, 58–66 (2009).
31. Ho, S. Y. & Phillips, M. J. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst. Biol.* syp035 (2009).
32. Batchelor, T. J. & Ward, D. J. Fish remains from a temporary exposure of Hythe Beds (Aptian–Lower Cretaceous) near Godstone, Surrey. *Mesoz. Res.* **2**, 181–203 (1990).
33. Thies, D. Jurazeitliche Neoselachier aus Deutschland und S-England. (1983).
34. Underwood, C. J. & Ward, D. J. Neoselachian sharks and rays from the British Bathonian (Middle Jurassic). *Palaeontology* **47**, 447–501 (2004).
35. Ivanov, A. Early Permian chondrichthyans of the middle and south Urals. *Rev. Bras. Paleontol.* **8**, 127–138 (2005).
36. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci.* **90**, 4087–4091 (1993).
37. Rambaut, A., Suchard, M. A., Xie, D. & Drummond, A. J. *Tracer v1. 6*. (2014).
38. Rambaut, A. FigTree, a graphical viewer of phylogenetic trees. *See* [Http://treebioedacuk/software/figtree](http://treebioedacuk/software/figtree) (2007).
39. Sanders, K. L. & Lee, M. S. Evaluating molecular clock calibrations using Bayesian analyses with soft and hard bounds. *Biol. Lett.* **3**, 275–279 (2007).
40. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
41. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893 (1969).
42. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
43. Li, W.-H. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics* **85**, 331–337 (1977).
44. Nei, M. & Li, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76**, 5269–5273 (1979).
45. Rowat, D. & Brooks, K. S. A review of the biology, fisheries and conservation of the whale shark *Rhincodon typus*. *J. Fish Biol.* **80**, 1019–1056 (2012).

46. Berumen, M. L., Braun, C. D., Cochran, J. E., Skomal, G. B. & Thorrold, S. R. Movement patterns of juvenile whale sharks tagged at an aggregation site in the Red Sea. *PloS One* **9**, e103536 (2014).
47. Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. Lond. B Biol. Sci.* **279**, 4724–4733 (2012).
48. Schnell, I. B. *erholm et al.* Screening mammal biodiversity using DNA from leeches. *Curr. Biol.* **22**, R262–R263 (2012).
49. Thomsen, P. F. *et al.* Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples. *PLoS ONE* **7**, e41732 (2012).