nature ecology & evolution

BRIEF COMMUNICATION

PUBLISHED: 21 NOVEMBER 2016 | VOLUME: 1 | ARTICLE NUMBER: 0004

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

Eva Egelyng Sigsgaard[1,2], Ida Broman Nielsen[1], Steffen Sanvig Bach[3], Eline D. Lorenzen[2], David Philip Robinson[4], Steen Wilhelm Knudsen[2], Mikkel Winther Pedersen[1], Mohammed Al Jaidah[5], Ludovic Orlando[1], Eske Willerslev[1,6,7], Peter Rask Møller[2] and Philip Francis Thomsen[1]*

Population genetics is essential for understanding and managing marine ecosystems, but sampling remains challenging. We demonstrate that high-throughput sequencing of seawater environmental DNA can provide useful estimates of genetic diversity in a whale shark (*Rhincodon typus*) aggregation. We recover similar mitochondrial haplotype frequencies in seawater compared to tissue samples, reliably placing the studied aggregation in a global genetic context and expanding the applications of environmental DNA to encompass population genetics of aquatic organisms.

Population genetic information is essential for the informed management and conservation of endangered species, but for rare oceanic species sampling remains a challenge. The whale shark is an iconic, but endangered, oceanic species, mainly due to overexploitation[1]. Despite its large size, much of the whale shark's biology remains unknown[2]. For instance, although studies have documented coastal aggregations of whale sharks around the world[2], little is known about offshore aggregations[3]. Population studies have primarily depended on tissue sampling and tagging, which are expensive and potentially harmful[4].

Here, we investigated the use of seawater environmental DNA (eDNA) to obtain genetic information at the population level. Environmental DNA from water samples has been used to detect and quantify aquatic macroorganisms in freshwater[5,6] and, more recently, in seawater[7–9]. However, aquatic eDNA has yet to be applied for obtaining population genetic information. We studied a recently discovered seasonal aggregation of whale sharks at Al Shaheen oil field offshore of Qatar in the Arabian Gulf[3] using eDNA from 20 seawater samples (Fig. 1a,b; Supplementary Table 1; Supplementary Information). To date over 300 individuals have been identified at this aggregation site (male/female ratio of ~2)[10].

We compared mitochondrial (mtDNA) control region sequences obtained from PCR amplification and Illumina sequencing (metabarcoding) of eDNA samples (two polymorphic regions; DL1: 412 bp and DL2: 476–493 bp) to sequences from tissue samples collected at the same locality (61 individuals; Supplementary Table 2). Considering only known haplotypes, we found similar relative haplotype frequencies in the seawater eDNA compared to the tissue samples (Fig. 1c,d). This suggests that quantitative relationships between haplotypes present at the time of water sampling are reflected in the sequencing data. A mock sample prepared from known haplotypes indicated a positive correlation between DNA template concentration and read output (Supplementary Fig. 4), supporting a quantitative relationship between the two. We retrieved more haplotypes from eDNA (DL1: 7, DL2: 18) than from tissue samples (DL1: 4, DL2: 12) (Fig. 1c,d; Supplementary Figs 1 and 2), indicating that the tissue database did not represent the complete mitochondrial diversity of the aggregation. The four DL1 haplotypes found in the tissue samples were also found within the eDNA and included one haplotype unique to Qatar. Similarly, all twelve DL2 haplotypes found in the tissue samples were found in the eDNA, including one haplotype unique to Qatar. Globally, 19 DL1 haplotypes and 44 DL2 haplotypes were identified (Supplementary Table 3).

Using principal component analysis (PCA) of the relative haplotype frequencies inferred from the eDNA reads and tissue samples, we placed the studied aggregation in the population genetic context of the world's whale sharks[11–13], in which the Atlantic and Indo-Pacific Ocean populations appear to be differentiated[13]. This analysis indicated that the Arabian Gulf aggregation groups with other Indo-Pacific aggregations, but not with the Atlantic (Fig. 1e). Interestingly, the eDNA data clustered closely with local tissue samples and the scores of the first principal component correlated with the distance to the Gulf by the shortest sea route ($P < 0.01$, $R^2 = 0.80$) (Fig. 1e; Supplementary Information). Analysis of the genetic differentiation by an F_{ST} (fixation index) assessment based on the tissue samples (DL2 fragment) confirmed that the Gulf aggregation is not significantly differentiated from other Indo-Pacific populations ($F_{ST} = 0–0.03$, $P > 0.05$), but differs from the Atlantic population ($F_{ST} = 0.30$, $P < 0.001$) (Supplementary Information).

As a proof of concept, we estimated the effective female population size (N_f) on the basis of eDNA from the most polymorphic target region (DL2) (Supplementary Information), using an estimated mutation rate of 0.1% per million years (95% confidence interval (CI): 0.04–0.16%) (Supplementary Fig. 10; Supplementary Information). The resulting N_f estimate was 71,600 (95% CI: 43,618–183,526; nucleotide diversity $\pi = 0.00358$), when scaling haplotype

[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. *e-mail: pfthomsen@snm.ku.dk

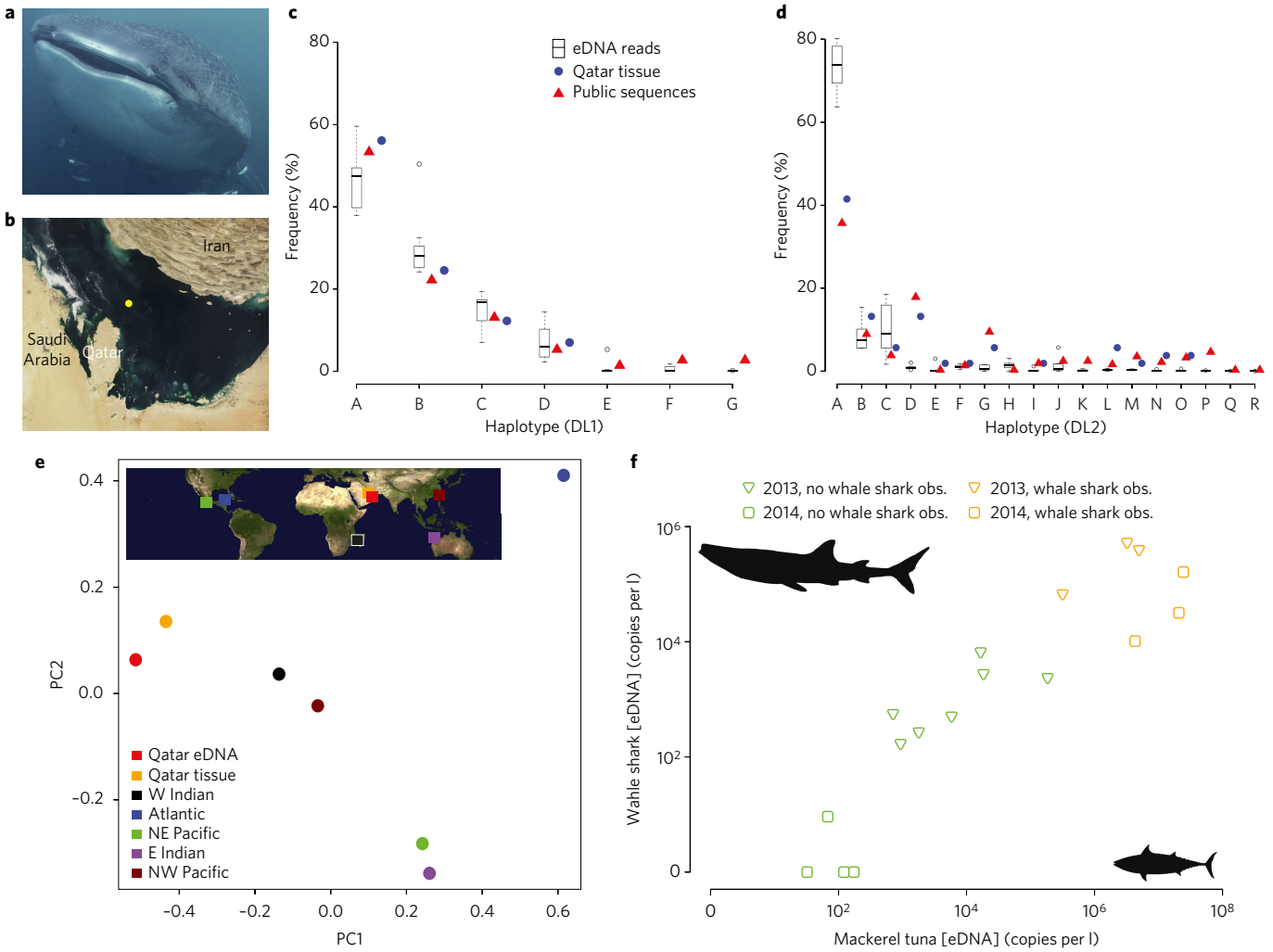

Figure 1 | Results from seawater environmental DNA analyses of a whale shark aggregation in the Arabian Gulf. a, Whale shark from Al Shaheen, Qatar (image courtesy of P. R. Møller). **b**, Sampling area in the Arabian Gulf (map: NASA, Visible Earth). **c**, Frequencies of DL1 haplotypes in eDNA reads (box plots, $n=7\times3$ water samples) and tissue samples from Al Shaheen (blue circles, $n=57$ individuals), and overall haplotype frequencies from NCBI and this study (tissue samples) (red triangles, $n=77$ individuals)[11,25,26]. **d**, Frequencies of DL2 haplotypes in eDNA reads (box plots, $n=5\times3$ water samples) and tissue samples (blue circles, $n=53$ individuals) from Al Shaheen, and overall haplotype frequencies from NCBI, Vignaud *et al.*[13] and this study (tissue samples; red triangles, $n=370$ individuals)[11,13,25,26]. Box plot whiskers: most extreme data point ≤1.5 times the interquartile range from the box. Open black circles: outliers defined as data points >1.5 times the interquartile range from the box. **e**, PCA of DL1 and DL2 sequences from eDNA reads, tissue samples and sequences from NCBI and Vignaud *et al.*[13] (PC1: 57% and PC2: 22% of variance) (map: NASA, Visible Earth). **f**, Correlation between whale shark and mackerel tuna eDNA concentrations, based on qPCR of 17 samples collected in 2013 and 2014 with or without concurrent visual observations of whale sharks. See also Supplementary Information.

frequencies to 100 individuals. The estimated daily number of individuals in the Al Shaheen aggregation[10] is approximately 124, and up to ~200 individuals were present during water sampling based on fin counts (Supplementary Table 1). Scaling haplotype frequencies to 20, 50, 100, 150, 200, 250 and 300 individuals, respectively, resulted in an average estimate of 75,543 females (95% CI: 54,714–96,372). On the basis of tissue samples, N_f was estimated at 138,400 (95% CI: 85,087–351,654; $\pi=0.00692$). These estimates are assumed to reflect the entire Indo-Pacific N_f, as little genetic subdivision has been reported within this region[11–13].

While the N_f estimate from eDNA was approximately half of that estimated from the tissue samples, we find the overlap between CIs promising for eDNA as a proxy for estimating effective population sizes. Our regional estimates are meaningful compared with global estimates of effective population size that are based on complete control region sequences (119,000–238,000 females, no reported CI)[11] and microsatellites (103,572 individuals, standard error range: 27,401–179,794)[12]. Importantly, these estimates are all based on estimated mutation rates, which are difficult to determine accurately[14].

To account for errors generated during amplification and sequencing, which could lead to false positive haplotypes, we cleaned our data using observed (mock sample) and *in silico* estimated error rates before performing the above analyses (Supplementary Figs 3 and 5; Supplementary Information). Interestingly, *in silico* error modelling showed that some haplotypes were more likely to arise as false positives than others (Supplementary Figs 5–7). Cleaning removed one DL1 haplotype and five DL2 haplotypes (<5% of reads in both cases) (Supplementary Information).

As extensive knowledge of haplotypes is not always available, we performed a new analysis on our eDNA reads using only reference sequences from six individuals (mock sample) (Supplementary Figs 8 and 9). All of the original DL1 haplotypes and 10 of the 18 DL2 haplotypes were retained. Additionally, several unknown putative

haplotypes (DL1: 9, DL2: 6) were also found. Nevertheless, results were very similar to those found using a reference database for both N_f (63,400, 95% CI: 38,525–162,899, based on 100 individuals) and PCA, demonstrating that eDNA metabarcoding data can be used independently for population genetic inference with little prior knowledge of the studied population.

Owing to its high detection rates, cost-efficiency and non-invasiveness compared with traditional survey methods[15], eDNA analysis is increasingly recognized as a valuable tool for ecological inference and management of aquatic biodiversity[9,15]. However, to realize its full potential, aquatic eDNA needs to advance from species detection to the analysis of populations. Much remains to be investigated regarding the relationship between eDNA data and abundance or biomass, including the influences of abiotic factors. The unknown number of source individuals for an eDNA sample represents another challenge; choosing a number of individuals for scaling is a major assumption and at present requires additional information. Lastly, conservative or non-conservative data cleaning criteria may lead to under- or overestimation of genetic diversity, respectively. Future advances may facilitate the retrieval of longer eDNA fragments and provide higher read coverage, but it is unclear whether the identification of individuals will be possible. Nevertheless, we have demonstrated that reliable estimates of haplotype frequencies, genetic diversity and population subdivision can be retrieved from eDNA—even in the absence of a reference database. The data derived from eDNA required fewer resources and a smaller sampling effort, compared to that derived from tissue samples (Supplementary Information).

The whale sharks in Qatar are reported to aggregate at Al Shaheen to feed on fish spawn from mackerel tuna (*Euthynnus affinis*)[3]. To investigate aquatic eDNA as a potential proxy for studying trophic interactions, we quantified eDNA from both species using quantitative PCR (qPCR). The concentration of whale shark eDNA correlated strongly with that of mackerel tuna ($P<0.001$, $R^2=0.84$) (Fig. 1f). We argue that this result most probably reflects the predator–prey relationship observed between the two species. Alternatively, the tuna may follow the sharks, as reported from the Azores[16], but this has never been observed in Al Shaheen.

Sea currents can potentially move genetic material over large distances, leading to detections of non-local eDNA. However, samples collected concurrently with visual observations of whale sharks contained higher concentrations of whale shark and mackerel tuna eDNA ($P<0.001$ for both, Wilcoxon test) (Fig. 1f; Supplementary Information), supporting a local origin of the sampled eDNA. This is in line with research indicating the differentiation, at scales of ~60 m, of marine eDNA[7] and degradation within days[8]. We performed an experiment on local seawater, which suggests that whale shark eDNA in the Gulf degrades on a similar timescale (Supplementary Fig. 11; Supplementary Information). Thus, while more work is needed, we find it reasonable to assume that our results reflect local population composition due to limited long-distance movement of eDNA.

Ongoing research on large oceanic species, such as the whale shark, includes tissue sampling, acoustic surveys, satellite tagging, aerial surveys and photo identification[17]. Most are dependent on good weather conditions and visibility, and are restricted to individuals near the surface. Aquatic eDNA sampling overcomes these challenges and offers high sensitivity[8,15]. To our knowledge, this study represents the first to show that aquatic eDNA can be used for population-level inferences and for identifying species co-occurrences that may indicate trophic interactions. This broadens the scope of eDNA research and facilitates the informed management of aquatic biodiversity and resources.

Methods

Seawater samples were collected in May 2013 and May–June 2014 at 15 locations near the Al Shaheen oil field (20 samples in total, Supplementary Table 1). Nineteen samples of 3 × 500 ml (1.5 l total) were collected at the surface and filtered through sterile 0.22 µm Sterivex-GP filters (Merck Millipore) using 60 ml syringes (HSW Soft-Ject). An additional sample of 6 × 30 l was collected to measure eDNA degradation. Prior to water sampling, the number of sharks in the aggregation was estimated by counting fins at the surface.

Whale shark tissue samples were taken with a biopsy spear in 2011–2014 and preserved in 96% ethanol. Sharks were photographed and later identified to the individual level (Supplementary Information).

The Qiagen DNeasy Blood and Tissue Kit was used for DNA extraction from both tissue (manufacturer's protocol) and water (modified protocol) samples. Tissue-extracted DNA was PCR amplified using the primers WSCR1-F and WSCR1-R[11], which target the complete control region. PCR products were Sanger sequenced at Macrogen Europe and sequences were quality checked in Geneious v. 7.1.7 (Biomatters Ltd) and assigned to haplotypes with DnaSP v. 5.10.1[18].

In eight samples, taken where whale sharks were visually observed (Supplementary Table 1), we used PCR to amplify two polymorphic regions of the whale shark mtDNA control region (DL1: 412 bp, DL2: 476–493 bp). The three eDNA extracts from each 3 × 500 ml sample were combined in pools. The DL1 and DL2 regions were PCR amplified using primers tagged with oligonucleotides eight nucleotides in length[19]. A unique combination of tags on the forward and reverse primers was used for each PCR replicate (six replicates per sample). PCR products from the samples that yielded positive amplification (DL1: 7 samples, DL2: 5 samples) were purified using Qiagen MinElute kit. Libraries were prepared using the NEBNext DNA Library Prep Master Mix Set for 454 (New England Biolabs Inc.) and sequenced at Macrogen Europe on the Illumina MiSeq platform (DL1: 250 bp paired-end, DL2: 300 bp paired-end). A PhiX spike-in and a mock sample prepared from tissue extracts of six individuals (relative concentrations of 1 to 1:1000) were included in the sequencing runs. Sequences were analysed in OBITools[20]. As read quality for DL2 dropped after ~200 bp, paired-end reads were joined end-to-end and the low-quality middle sequence was removed using a custom Python script. Only 100% matches to known whale shark haplotypes were considered.

For the PCA analysis, haplotype frequencies from eDNA were first scaled to a total of 100 individuals, corresponding approximately to the number of individuals observed (between ~20 and ~200 sharks) when the water samples included for sequencing were taken. Frequencies below 1% were rounded up, so each haplotype was represented by at least one individual. Sequences from the National Center for Biotechnology Information (NCBI) and Vignaud *et al.*[13] represented 32 individuals from Mozambique, 16 from Taiwan and the Philippines, 146 from Ningaloo Reef in Australia and 38 from the Gulf of California in Mexico (Indo-Pacific populations), as well as 32 individuals from Isla Holbox in Mexico (Atlantic population).

A sequencing error rate of 0.3% was estimated from the PhiX output. A putative combined PCR and sequencing error rate of 1.3% was calculated on the basis of low-frequency spurious haplotypes retrieved from the mock sample. Haplotypes appearing at a frequency below these rates were removed from the data.

When the analysis was redone without the reference database, cleaning was done on the basis of the error rate observed in the mock sample (1.3%) and assuming that the most abundant sequence from a PCR was authentic. In addition, sequences were required to be present in at least two PCRs.

The mutation rate of the DL2 region was estimated in BEAST v. 1.8.2[21] using a fossil-calibrated phylogeny, on the basis of the alignment of forty shark species and a relaxed clock model. Nucleotide diversity was determined in DnaSP. Effective female population size was calculated as $N_f=\pi/2\mu$, with μ being the mutation rate[22–24], assuming a model of constant size and a generation time of 25 years[1]. The concentrations of whale shark and mackerel tuna eDNA in the 3 × 500 ml water samples collected on 27–28 May 2013 and 19–20 May 2014 (17 samples) were estimated by qPCR on a Stratagene Mx3005P, using Taqman primer/probe qPCR assays. Both assays targeted ~100 bp mtDNA fragments of the *CYTB* gene. Standard dilutions were prepared from PCR amplicons of tissue-extracted DNA.

To estimate decay rates, the 6 × 30 l water sample was divided into two 90 l buckets that were placed in sunshine and shade, respectively. A 500 ml sample was collected from each bucket every morning and evening (more often on the first three days) for eight days, giving a total of 22 samples per bucket. (This corresponds to a removal of ~12% of the starting volume by the end of the experiment.) Whale shark eDNA concentrations were estimated by qPCR as above and an exponential decay model was fitted to the data.

Detailed descriptions of all methods can be found in the Supplementary Information.

Data availability. Illumina MiSeq raw sequence data are available from the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.kn206). Control region sequences for individual whale sharks generated from tissue samples have been added to Genbank (NCBI Accession numbers KX944487 to KX944547). Input files for phylogenetic analysis of the DL2 fragment in BEAST are available as Supplementary Data files.

Received 19 April 2016; accepted 11 October 2016; published 21 November 2016

References

1. Pierce, S. J. & Norman, B. *Rhincodon typus.* http://dx.doi.org/10.2305/IUCN.UK.2016-1.RLTS.T19488A2365291.en (IUCN, 2016).
2. Rowat, D. & Brooks, K. S. *J. Fish Biol.* **80,** 1019–1056 (2012).
3. Robinson, D. P. *et al. PLoS ONE* **8,** e58255 (2013).
4. Dicken, M. L., Booth, A. J. & Smale, M. J. *ICES J. Mar. Sci.* **63,** 1640–1648 (2006).
5. Ficetola, G. F., Miaud, C., Pompanon, F. & Taberlet, P. *Biol. Lett.* **4,** 423–425 (2008).
6. Thomsen, P. F. *et al. Mol. Ecol.* **21,** 2565–2573 (2012).
7. Port, J. A. *et al. Mol. Ecol.* **25,** 527–541 (2016).
8. Thomsen, P. F. *et al. PLoS ONE* **7,** e41732 (2012).
9. Thomsen, P. F. & Willerslev, E. *Biol. Conserv.* **183,** 4–18 (2015).
10. Robinson, D. P. *et al. PLoS ONE* **11,** e0158593 (2016).
11. Castro, A. L. F. *et al. Mol. Ecol.* **16,** 5183–5192 (2007).
12. Schmidt, J. V. *et al. PLoS ONE* **4,** e4988 (2009).
13. Vignaud, T. M. *et al. Mol. Ecol.* **23,** 2590–2601 (2014).
14. Kondrashov, F. A. & Kondrashov, A. S. *Phil. Trans. R. Soc. B.* **365,** 1169–1176 (2010).
15. Valentini, A. *et al. Mol. Ecol.* **25,** 929–942 (2016).
16. Fontes, J., McGinty, N., Machete, M. & Afonso, P. in *QScience Proc. (The 4th Int. Whale Shark Conf.)* Vol. 2016, iwsc4.17 (2016).
17. Meekan, M., Speed, C., Planes, S., McLean, C. & Bradshaw, C. *Population Monitoring for Whale Sharks* (*Rhincodon typus*) (Australian Institute of Marine Science, 2008).
18. Librado, P. & Rozas, *J. Bioinform.* **25,** 1451–1452 (2009).
19. Meyer, M., Stenzel, U. & Hofreiter, M. *Nat. Protoc.* **3,** 267–278 (2008).
20. Boyer, F. *et al. Mol. Ecol. Resour.* **16,** 176–182 (2016).
21. Drummond, A. J., Suchard, M. A., Xie, D., Rambaut, A. *Mol. Biol. Evol.* **29,** 1969–1973 (2012).
22. Watterson, G. A. *Theor. Popul. Biol.* **7,** 256–276 (1975).
23. Tajima, F. *Genetics* **105,** 437–460 (1983).
24. Nei, M. & Li, W.-H. *Proc. Natl Acad. Sci. USA* **76,** 5269–5273 (1979).
25. Alam, M. T., Petit R. A. III, Read, T. D. & Dove, A. D. M. *Gene* **539,** 44–49 (2014).
26. Schmidt, J. V. *et al. Endang. Species Res.* **12,** 117–124 (2010).

Acknowledgements

We thank the Qatar Ministry of Environment for their collaboration and invaluable support. In particular, the crew on board *R/V Saqt Al Khaleej* is thanked for help with logistics for the water sampling. We thank the Maersk Oil Research and Technology Centre (MO-RTC) in Doha, Qatar, for being the main sponsor of the project. Special thanks to A. S. Al-Emadi (Head of MO-RTC) and J. Al-Khori (Technical Manager of MO-RTC) for supporting the project. The Danish National Research Foundation and the Natural History Museum of Denmark are thanked for additional funding. We thank T. B. Brand and the rest of the staff at the Centre for GeoGenetics, University of Copenhagen, as well as K. Magnussen and the Danish National Sequencing Centre for laboratory support. M. Krag is thanked for help with tissue samples. L. Olsen and P. Gravlund, National Aquarium Denmark (Den Blå Planet), provided a *Stegostoma fasciatum* tissue sample. J. V. Schmidt, University of Illinois at Chicago; R. W. Jabado, UAE University, Abu Dhabi, United Arab Emirates; and N. S. Blom, Danish Technical University, are thanked for scientific input. E. Vissing is thanked for the custom Python script.

Author contributions

P.F.T., E.E.S., I.B.N., P.R.M. and S.S.B. conceived and designed the experiments. E.E.S., I.B.N., P.F.T., P.R.M. and S.S.B. performed the experiments. E.E.S., I.B.N., P.F.T. and M.W.P. analysed the data. E.E.S., P.F.T., P.R.M., S.S.B., E.D.L., D.P.R., S.W.K., M.W.P., M.A.J., L.O. and E.W. contributed materials or analysis tools. E.E.S., P.F.T., P.R.M., I.B.N., E.D.L. and S.W.K. wrote the paper.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.F.T.

How to cite this article: Sigsgaard, E. E. *et al.* Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nat. Ecol. Evol.* **1,** 0004 (2016).

Competing interests

The authors declare no competing financial interests.

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

Eva Egelyng Sigsgaard, Ida Broman Nielsen, Steffen Sanvig Bach, Eline D. Lorenzen, David Philip Robinson, Steen Wilhelm Knudsen, Mikkel Winther Pedersen, Mohammed Al Jaidah, Ludovic Orlando, Eske Willerslev, Peter Rask Møller and Philip Francis Thomsen

Nature Ecology & Evolution **1,** 0004 (2016); published 21 November 2016; corrected 19 December 2016.

The original version of this Brief Communication contained artefacts in Fig. 1 caused by errors in the production process. The figure has been corrected in all versions of the Brief Communication.