SC/66b/DNA/03

Pilot study to assess the utility of ddRAD sequencing in identifying species-specific and shared SNPs among Blainvilles (Mesoplodon densirostris) and Cuviers (Ziphius cavirostris) beaked whales

E. L Carroll, C. Reyes, O. E. Gaggiotti, M. T. Olsen, D. J. Maaholm, M. Rosso, N. Davison, V. Martin, A. Schiavi, and N. Aguilar de Soto

INTERNATIONAL WHALING COMMISSION

Papers submitted to the IWC are produced to advance discussions within that meeting; they may be preliminary or exploratory.
It is important that if you wish to cite this paper outside the context of an IWC meeting, you notify the author at least six weeks before it is cited to ensure that it has not been superseded or found to contain errors.

Pilot study to assess the utility of ddRAD sequencing in identifying species-specific and shared SNPs among Blainville's (*Mesoplodon densirostris*) and Cuvier's (*Ziphius cavirostris*) beaked whales

E. L Carroll[1], C. Reyes[2,3], O. E. Gaggiotti[1], M. T. Olsen[4], D. J. Maaholm[4], M. Rosso[5], N. Davison[6], V. Martin[7], A. Schiavi[2], and N. Aguilar de Soto[2,8]

1. Scottish Oceans Institute, University of St Andrews, East Sands, St Andrews, Fife, KY16 8LB, Scotland, UK

2. BIOECOMAC, La Laguna University, Tenerife, Canary Islands, Spain

3. Sea Mammal Research Unit, University of St Andrews, East Sands, St Andrews, Fife, KY16 8LB, Scotland, UK.

4. Evolutionary Genomics Section, Natural History Museum of Denmark, University of Copenhagen, DK-1350 Copenhagen K, Denmark

5. CIMA Research Foundation, 17100, Italy

6. Scottish Marine Animal Stranding Scheme, SRUC Veterinary Services, Drummondhill, Inverness, IV2 4JZ, Scotland

7. Sociedad para el studio de Cetaceos en Canarias. SECAC. Lanzarote. Canary Islands, Spain

8. Centre for Research into Ecological & Environmental Modelling (CREEM), The Observatory, Buchanan Gardens, University of St Andrews, Fife, KY16 9LZ, Scotland, UK

ABSTRACT

Restriction site-associated DNA (RAD) sequencing has become a popular approach to genotyping non-model organisms for ecological and evolutionary studies. However, there is difficulty in predicting how many variable loci will be recovered for a given protocol, combination of restriction enzymes and/or size selection criteria. Here we undertake a pilot study of a double digest RAD (ddRAD) protocol in Blainville's and Cuvier's beaked whales. Four samples from each species were run, with all samples from Blainville's coming from El Hierro, Canary Islands, and the Cuvier's samples coming from the Canary Islands, Scotland and the Mediterranean. The pilot study produced 9.2M quality controlled reads for the Blainville's and 16.4M quality controlled reads for the Cuvier's beaked whales. After loci construction and filtering in program STACKS, this produced 8143 variable RAD loci for Blainville's and 14095 variable RAD loci for Cuvier's beaked whales at moderate depths (20x). The higher variability in Cuvier´s beaked whales is probably due to the difference in sequencing success between the species and the broader geographic range of the Cuvier's compared with the Blainville's samples. In addition, we analysed the data using PYRAD to identify loci in common across the two species; this revealed 9666 loci at 20x depth in common between at least one sample per species.

INTRODUCTION

Restriction site-associated DNA sequencing (RADseq) is now commonly used in ecological and evolutionary studies to identify thousands of polymorphic genetic markers in taxa with no prior genomic resources (Andrews *et al.* 2016). One common variant of RADseq is double-digest RAD (ddRAD), where two restriction enzymes are used and loci with cut sites from both enzymes are selected using specific adaptors (Peterson *et al.* 2012). ddRAD is seen as more 'tuneable' than conventional RADseq, as by manipulating the choice of enzymes and size selection window the user can optimise the number of loci generated by the technique (Andrews *et al.* 2016).

However, despite general guidance in the literature (Peterson *et al.* 2012), it can be difficult to *a priori* predict the number of loci a particular combination of restriction enzymes and size selection will generate. This information can be critical in determining how many samples to multiplex on a high-throughput sequencing lane. Low coverage will result in low sequencing depths per locus, limiting the confidence in individual genotype assignments (Nielsen *et al.* 2011). Alternatively, multiplexing fewer samples per lane will result in high

coverage, but potentially fewer samples per study, limiting the power and/or scope of studies. Given limited funding resources, optimising the number of samples to multiplex is a critical step to in any large-scale ddRAD sequencing study.

Here we undertake a pilot study in Blainville's (*Mesoplodon densirostris*) and Cuvier's (*Ziphius cavirostris*) beaked whales to assess the effectiveness of a particular ddRAD protocol that is currently being used in other marine mammal species. Specifically, we aim to estimate the total number and proportion of variable RAD loci generated at different sequencing depths as a function of sequencing effort, both within and between the two species. We use the ddRAD protocol developed by Peterson *et al* (2012) and optimised by K. Andrews (University of Idaho) and E. L. Carroll for marine mammals.

METHODS

Skin samples were collected from either dead stranded whales or from free-ranging whales using a lightweight biopsy dart fired from a modified veterinary capture rifle (Krützen *et al.* 2002). Four samples from Blainville's beaked whales and two samples from Cuvier's beaked whales were collected using biopsy sampling off El Hierro, Canary Islands. Further samples of Cuvier´s beaked whale were collected; one biopsy collected from a whale in the Ligurian Sea, Mediterranean, and one sample collected from a whale stranded in Scotland.

Genomic DNA was extracted using a standard proteinase K digestion and phenol/chloroform methods (Sambrook *et al.* 1989) modified for small tissue samples (Baker *et al.* 1994). DNA samples were quantified with the Qubit dsDNA HS assay kit (Life Technologies) and standardised to a concentration of 20 ng/uL. We then followed the ddRAD protocol developed by Peterson *et al.* (2012), modified as follows. Briefly, a total of 250 ng of DNA per sample was digested with MspI and Hind III overnight at 37°C, followed by a 20 minute heatkill step at 65°C to deactivate the restriction enzymes. We then ligated adaptors with forward barcodes onto the samples using 15 uL of ligation mix (4 uL Buffer , 0.5 uL T4 ligase and 10.5 uL PCR H20) and 15 uL of annealed adaptors. Ligation was conducted in a PCR machine with the following protocol: 22°C for 2 hours; 65°C for 20 min and 8°C hold. At this stage, the four samples from each species had unique forward barcodes annealed and were then pooled to form a library. Following clean up with PureLink PCR microkit columns (Invitrogen), libraries underwent size selection using a Pippin Prep (Sage Science) and a target range of 300-400 bp. To anneal Illumina sequencing primers and reverse indexes, libraries were divided into 8 parts and PCR was undertaken using Phusion high-fidelity PCR kits to manufacturer's recommendation (10 cycles only). Libraries were again pooled after PCR, cleaned using AMPURE-XP beads and eluted in 10 uL of qiagen EB. Library quality and quantity were checked with Bioanalyzer and qPCR. Libraries were then pooled in equimolar amounts for sequencing on an Illumina MiSeq at Edinburgh Genomics.

We undertook two separate analyses of the ddRAD data; a species-specific analysis in program STACKS (Catchen *et al.* 2013) and a combined analysis in program PYRAD (Eaton 2014). We elected to undertake the combined analysis in PYRAD because it has a global alignment method for identifying loci that is supposed to process insertion-deletion mutations more effectively than STACKS, making it more suitable for phylogenetic datasets. For the species-specific analyses, we demultiplexed and truncated reads to 130 bp, and discarded those reads with a raw phred score of <20 using the command process_radtags. We then used STACKS to identify RAD loci from the data in the absence of a reference genome, using the denovo_map command. For this, we set the minimum number of reads needed to identify an allele (-m) as 3 and the number of mismatches allowed between loci when processing a single individual was set at 2 (-M). Other settings were left at the STACKS defaults.

Data in the resulting STACKS catalogue were filtered in several ways to estimate the total number and number of variable loci produced in the pilot study. We examined all loci that were present in at least three of four samples per species and that were present at minimum depths of 10x, 20x and 30x. We then estimated the number of variable loci, which we define as those with between 1 and 4 SNPs, and present in at least 3 of 4 samples, at different read depths. To examine the number of loci per sample, we plotted the number of variable loci at each depth by the total number of quality controlled (QC) reads.

We exported RAD loci present at a depth of 20x with between 1 and 4 SNPs from STACKS in software STRUCTURE format to analyse any signals of population structure in the small species-specific datasets. To do this, we ran three iterations of STRUCTURE on the admixture setting, with K =1 to 4, and 100,000 burn in and 1,000,000 MCMC steps. Optimal K was selected using the Evanno *et al*. (Evanno *et al.* 2005) method in program STRUCTURE HARVESTER (Earl & vonHoldt 2012), and CLUMPAK (Kopelman *et al.* 2015) was used to identify the most consistent admixture proportions for each individual over the three STRUCTURE runs.

For the combined analysis, we used PYRAD to demultiplex samples and then merged paired end reads and trimmed adaptors using PEAR (Zhang *et al.* 2014). Quality control was undertaken with PYRAD, which

discarded reads based on quality scores and the number of ambiguous bases in the sequence. We then clustered reads into loci within and then across samples, with merged paired-end reads and unassembled reads run under different PYRAD settings. We then report the number of shared loci between the two species, defined as those present in at least one sample from each species, at minimum depths of 10x, 20x and 30x.

RESULTS AND DISCUSSION

There was a bias in the number of reads towards the Cuvier's beaked whale library, resulting in 16,364,711 QC paired-end reads for the Cuvier's beaked whales and 9,239,361 for the Blainville's beaked whales. This is likely pipetting error, or because the qPCR results were at the high end of the reference range, meaning the libraries were pooled based on data that were less reliable than if the values had been mid-range. Subsequently, the number of QC reads, unique stacks and polymorphic loci were lower in the Blainville's data compared with the Cuvier's data (Table 1).

Despite the differences in QC reads per library, the total number of rad loci present in the STACKS catalogue in at least 3 samples was similar between the species at the 10x and 20x depths (Table 2). The number of loci (total and polymorphic) that had 30x coverage was substantially lower in the Blainville's than in the Cuvier's, probably attributable to the difference in number of reads per sample. The proportion of loci that were variable was consistent within species, but different between them; it was approximately one third in Blainville's and one half in Cuvier's. This likely reflects the broader geographic range the Cuvier's beaked whale samples came from, as they are likely to harbour more diversity. Plotting the number of variable loci per sample at different read depths as a function of number of reads that passed quality control (Fig 1) indicated that >4M paired end reads is required per sample to obtain 10,000 variable loci at depth of 30x.

We conducted the STRUCTURE analysis with 8143 loci for Blainville's beaked whales; all loci present at 20x depth. For the Cuvier's beaked whales, we subsampled 8000 loci from the 14095 variable loci with ≥20x read depth. Analysis of the STRUCTURE results for the Blainville's samples indicated K=2 was the most likely value. However, inspection of the admixture proportions when K=2 showed that all individuals assigned 50% to cluster 1 and 50% to cluster 2 (Fig 1). This suggests STRUCTURE is assigning individuals randomly to K clusters due to the lack of underlying population structure (Latch *et al.* 2006; Martien *et al.* 2007).

The STRUCTURE results for the Cuvier's samples also indicated K=2, with the Scottish sample clearly assigning to a different cluster than the Ligurian Sea and Canary Islands samples (Fig. 1). The Scottish sample was removed and the STRUCTURE analysis rerun with the two Canary Islands and single Ligurian Sea sample. The choice of K=1 to 3 limits the use of the Evanno method, as delta K cannot be calculated for K=1 or K=3. However, K=2 clearly separated the Ligurian Sea samples from the Canary Islands samples. It is worth noting that K=3 had the lowest mean log likelihood averaged over three replicates; when K=3 each sample was assigned to a separate cluster.

The combined species analysis based on PYRAD had a lower number of reads that passed QC; this is partly because, where possible, paired-end reads were merged for this analysis (Table 1). This was not done for the STACKS analysis and could positively bias the number of loci. There were 9666 loci shared between at least one sample from each species at moderate read depths (20x; Table 2).

The results suggest that ~4M QC paired end reads per sample are sufficient to produced 10,000 reasonable quality SNPs, present in at least 3 of 4 samples per species. Given our knowledge of beaked whale biology and genetic diversity, we believe that 10,000 loci would be sufficient to detect structure, if any exists, as suggested by the preliminary analysis in STRUCTURE for Cuvier's beaked whales. In order to calculate how many samples to multiplex per lane, we need to adjust for (1) sequencing error (perhaps 20%) and (2) phiX spiking on Illumina runs to increase library diversity (perhaps 5%). Given an Illumina HiSeq 2500 v4 chemistry produces around 250 M PE reads, this leaves ~190M reads after error and phiX, meaning 45-50 samples could be multiplexed per lane. Caveats on this analysis include the fact we have not attempted to estimate linkage disequilibrium across the loci and in future will need to conduct additional filtering on RAD loci (e.g. minor allele frequency), which could mean we are over-estimating the number of loci we have recovered from our pilot study.

Table 1: Results of analyses per sample using STACKS (non-merged paired end reads) and PYRAD (merged and unassembled paired end reads), showing the number of reads that passed quality control (QC reads), number of unique RAD loci (U.loci) and number of variable RAD loci (Var. loci) per sample, and number of loci per sample (N. loci) at different sequencing depths (10x, 20x, 30x) per sample.

	STACKS				**PYRAD**			
	Sample site	**QC reads**	**U.loci**	**Var. loci**	**QC reads**		**N.loci**	
M. densirostris						10x	20x	30x
Md01T	Canary Is	2444347	113201	12195	973800	22531	8906	1676
Md03R	Canary Is	2371874	114069	12447	859717	21126	6455	909
Md07R	Canary Is	1528725	99585	8948	1048480	22975	9392	2080
Md04T	Canary Is	2894415	122782	13687	576869	13014	1309	119
mean		2309840	112409	11819	864717	19912	6516	1196
Z. cavirostris								
Zc06EH	Canary Is	4399357	165212	35070	1650312	27138	15507	7441
Zc09EH	Canary Is	5790903	179128	37361	2045950	28899	17238	8911
Zc16ALBA	Scotland	2060362	124901	24980	760047	8530	591	102
Zc35LS	Ligurian Sea	4114089	171705	27124	1499509	20145	10051	4998
mean		4091178	160236	21134	1488955	21178	10847	5363

Table 2: Results of species-level analyses in STACKS and PYRAD, showing the total number (All) and number of variable loci (Var.) found in at least three of four samples per species at different read depths per sample (NL_{10X}, NL_{20X}, NL_{30X}) from the STACKS analysis and the number of loci shared between the species from the PYRAD analysis (Shared).

		NL_{10X}	**NL_{20X}**	**NL_{30X}**
M. densirostris	All	59335	24882	5081
	Var.	19435	8143	1696
Z. cavirostris	All	53648	25267	12145
	Var	29910	14095	6754
Shared	All	19574	9666	2808

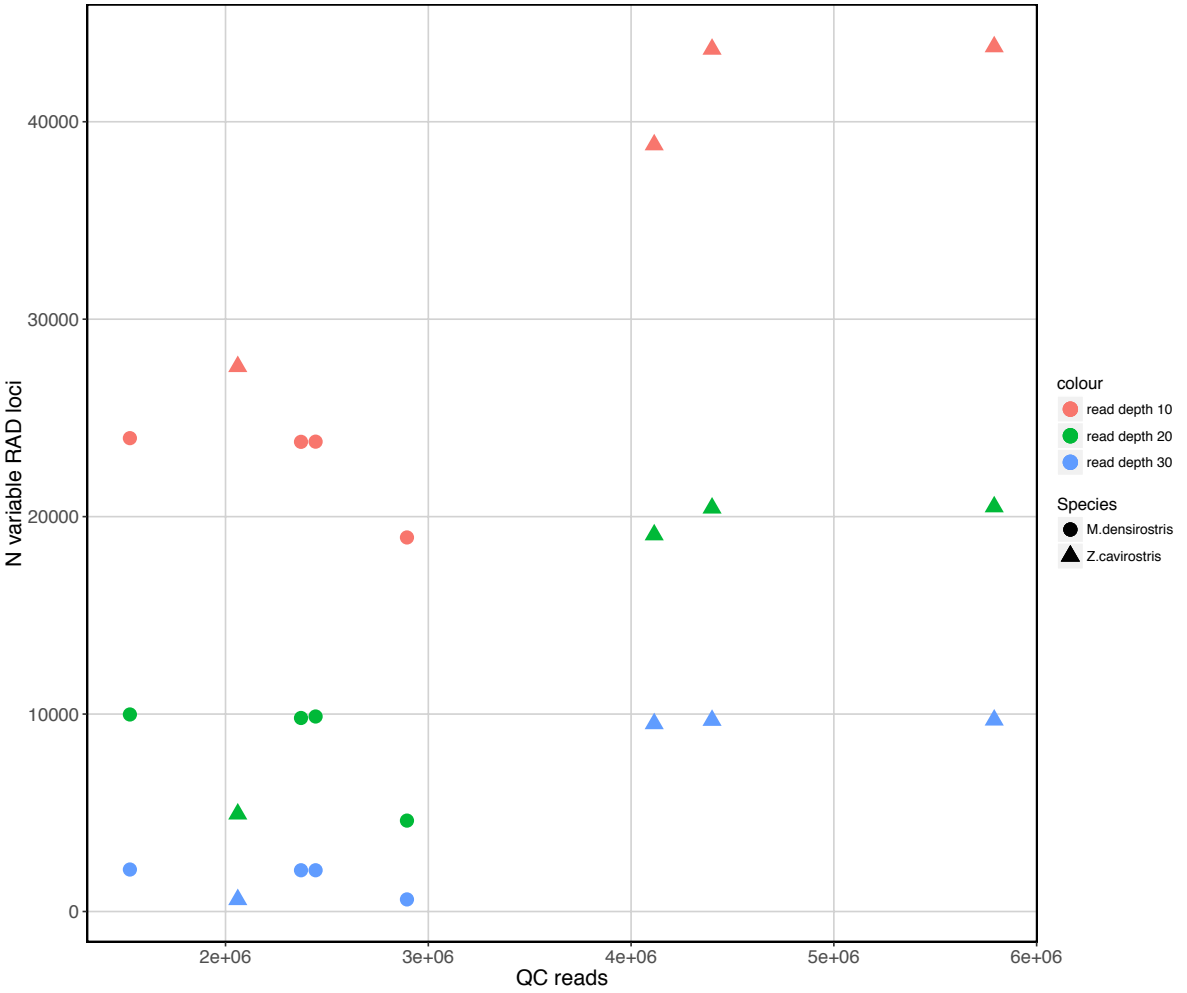

Figure 1: Number of loci present in each sample at various read depths as function of number of reads that passed quality control. The loci are those we define as variable, defined as having between 1 and 4 SNPs and present in at least 3 of 4 individuals, based on a species-specific STACKS analysis.

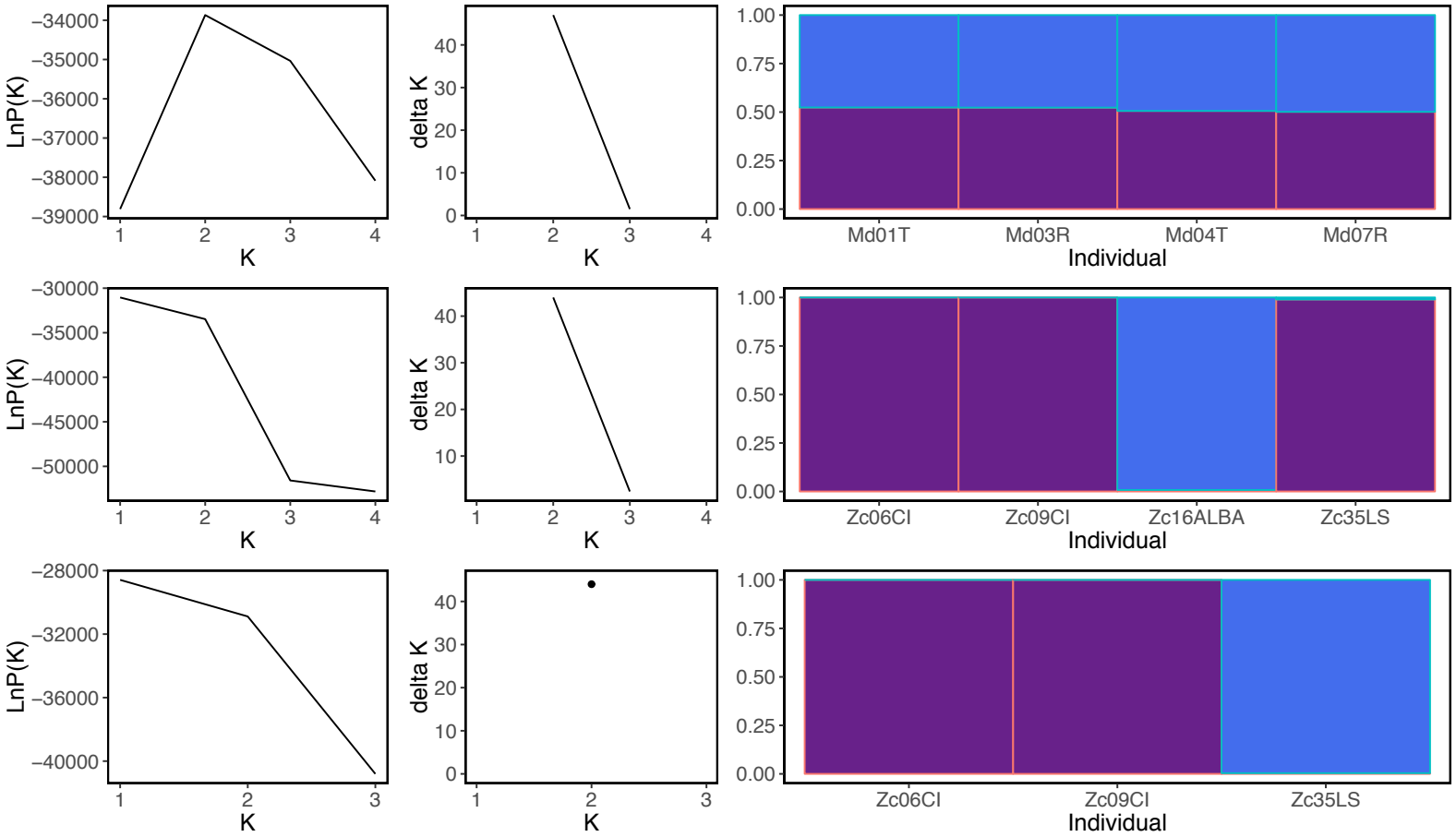

Figure 2: STRUCTURE analysis results. Left hand side shows mean log likelihood averaged over three replicates (LnP(K)) and second order rate of change constant (delta K) plotted against K for Blainville's (top panel) and Cuvier's (middle panel, all samples, bottom panel, Canary Islands and Ligurian Sea samples) beaked whale's. Right hand side shows the proportion of each individuals' genotype that assigns to each cluster.

CITATIONS

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet*, **17**, 81–92.

Baker CS, Slade RW, Bannister JL *et al.* (1994) Hierarchical structure of mitochondrial DNA gene flow among humpback whales, Megaptera novaeangliae, world-wide. *Molecular Ecology*, **3**, 313–327.

Catchen J, Hohenlohe P a, Bassham S, Amores A, Cresko W a (2013) Stacks: an analysis tool set for population genomics. *Molecular ecology*, **22**, 3124–40.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Eaton D (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, 1179–1191.

Krützen M, Barré L, Möller L *et al.* (2002) A biopsy system for small cetaceans; darting success and wound healing in Tursiops spp. *Marine Mammal Science*, **18**, 863–878.

Latch E, Dharmarajan G, Glaubitz J, Rhodes O (2006) Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, **7**, 295–302.

Martien K, Archie EA, Taylor BL (2007) Simulation-based performance testing of the Bayesian clustering program STRUCTURE. *Unpublished report (SC/59/SD3) presented to the Scientific Committee of the International Whaling Commission, Cambridge, UK*.

Nielsen R, Paul J, Alberechtsen A, Song Y (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–452.

Peterson B, Weber J, Kay E, Fisher H, Hoekstra H (2012) Double digest RADseq: An inexpensive method for De Novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual 2nd ed.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)*, **30**, 614–20.

ACKNOWLEDGEMENTS

ELC is supported by a Newton Fellowship from the Royal Society and a Marie Slodowska Curie Fellowship (Behaviour-Connect) funded by EU Horizon2020 program. NAS is supported by a Marie Slodowska Curie Fellowship (ECOSOUND) funded by EU Horizon2020 program. Data from El Hierro was collected within project "Monitoreo de cetáceos en El Hierro" supported by Fundación Biodiversidad-MAGRAMA and data analysis was supported by project "Assessing resilience of beaked whales to human impacts" funded by ONR. Biopsies off El Hierro were gathered with permit from the Spanish ministry MAGRAMA to the University of La Laguna (ULL) and with ethical approval number CEIBA2015-0141 from CEIBA (Committee for Ethics in Research and Animal Experimentation of ULL). The sample from the Ligurian Sea was collected under Permit MATTM 0018799/PNM.