Supplemental Information for:

Biogeography in the deep: Hierarchical population genomic structure of two beaked whale species

Aubrie B. Onoufriou, Oscar E. Gaggiotti, Natacha Aguilar de Soto, Morgan L. McCarthy, Phillip A. Morin, Massimiliano Rosso, Merel Dalebout, Nicholas Davison, Robin W. Baird, C. Scott Baker, Simon Berrow, Andrew Brownlow, Daniel Burns, Florence Caurant, Diane Claridge, Rochelle Constantine, Fabien Demaret, Sascha Dreyer, Martina Đuras, John Durban, Alexandros Frantzis, Luis Freitas, Gabrielle Genty, Ana Galov, Sabine S. Hansen, Andrew C. Kitchener, Vidal Martin, Antonio A. Mignucci-Giannoni, Valeria Montano, Aurelie Moulins, Carlos Olavarría, M. Michael Poole, Cristel Reyes Suárez, Emer Rogan, Conor Ryan, Agustina Schiavi, Paola Tepsich, Jorge Urban, Kristi West, Morten T. Olsen & Emma L. Carroll

Electronic Supplementary Materials (ESM) Table of Contents:

Table of Contents

Supplementary Spreadsheet Tables (SST) List:

Location: "Onoufriou_BW_Global_DNA_SupplementaryTables_Final.xlsx"

Worksheet 1 "SST1. Full Sample List": List of all samples included in both the ddRAD and mitogenome analyses. Sample names are in the column headed "ITABW-ID", indicating the identifier used in the newly established ITAWB database. The following are also provided per

sample when available: species, sampling date, sampling location, sampling region, sampling ocean, whether or not they were included in the ddRAD or mitogenome analysis, the population defined using either method, and a detailed description of the sample's origin (who contributed the sample, the origin of the sample, sample type, any permits or considerations for collection of the sample and any impact minimisation or assessment steps taken during sample collection).

Worksheet 2 "SST2. BEAST Partitions": List of partitions inferred with PARTITIONFINDER and analysed in a Bayesian phylogenetic framework implemented in BEAST.

Worksheet 3 "SST3. NCBI mtDNA List": List of the already published mitogenome sequences used in the mitogenome analysis by NCBI accession number, sample ID, species and sampling locality.

Worksheet 4 "SST4. mtDNA Seq. Stats": Summary of shotgun sequencing statistics for the samples used in the mitogenome analyses. For each species, Copenhagen sample ID, corresponding ITABW ID, sampling locality, number of total reads, reads mapped to the reference mitogenome and coverage are given. Summaries of each read statistic are given overall for each species.

Worksheet 5 "SST5. mtDNA Div-Diff": Diversity and differentiation statistics for the mtDNA data including whole mitogenomes and extracted control regions. Ocean-basin-level statistics are provided including sample size (N), segregating sites (S), haplotypes (h), haplotype diversity (Hd), nucleotide diversity (Nu) and fixed differences (FixDiff). Total and pair-wise differentiation statistics are also provided: Fst and dA.

Worksheet 6 "ST6. ddRAD Fst": Differentiation statistics (Fst) for the ddRAD data. Ocean-basin and population-level statistics (Fst point estimate and 95% confidence interval) are provided including sample size (n).

Worksheet 7 "SST7. Haplotypes": List of samples and their mitogenome haplotypes.

Electronic Supplementary Material (ESM) 1: Tissue Archive, DNA Extraction and Sample Selection

Tissue samples were either collected specifically for this project or provided on loan from archives maintained by contributors. A large set of samples was provided by Dr. Merel Dalebout, who collated samples for investigations into beaked whale systematics and genetic diversity (Dalebout, 2002; Dalebout, Baker, Mead, Cockcroft, & Yamada, 2004; Dalebout et al., 2007, 2014; Dalebout, Mead, Baker, Baker, & van Helden, 2002; Dalebout et al., 2005, 2003; Dalebout, Ruzzante, Whitehead, & Øien, 2006; Dalebout, Steel, & Baker, 2008; Gomerčić et al., 2006; Van Helden et al., 2002). All samples were either skin biopsies sampled directly from free-swimming animals (Krützen et al., 2002; Lambertsen, 1987), tissue collected from dead beach-cast or ship-strike individuals, or already extracted DNA provided by NOAA's Southwest Fisheries Science Centre Marine Mammal and Turtle Molecular Research Sample Collection as already extracted DNA (https://www.fisheries.noaa.gov/west-coast/science-data/marine-mammal-and-sea-turtle-research-tissue-collection). Freshly collected tissue samples were typically stored in either DMSO or 70-99% ethanol and stored at -20°C.

The purpose of the current study was to develop an understanding of the global baseline genetic diversity and structure of Blainville's and Cuvier's beaked whales (*Mesoplodon densirostris* and *Ziphius cavirostris,* respectively and henceforth 'Blainville's' and 'Cuvier's'), providing context for investigating the impacts of anthropogenic activities on resident populations in future studies. Balancing the coverage of samples across their respective distributions and budgetary constraints, it was decided that five lanes of sequencing on a HiSeq 2500 (Illumina) would be used for the current study. The pilot study by Carroll et al. (2016) concluded that sequencing up to 50 individuals per HiSeq2500 lane would generate ~10k variable SNPs per individual, a number likely to detect structure in these two beaked whale species.

Of the 89 Blainville's individuals in the sample archive, n=67 were available as tissue, n=21 were already extracted DNA, and one individual had both a tissue and DNA sample. Of the 340 Cuvier's individuals, n=289 were available as tissue, n=29 were DNA, n=22 had both tissue and already extracted DNA. DNA for the ddRAD and mitogenome analyses was extracted from approximately 30-50mg of tissue, using the Phenol:Chloroform:Isoamyl Alcohol method described by Sambrook, Fritsch, & Maniatis (1989) and modified for use in small tissue samples by Baker et al. (1994). In addition, a subset of the Cuvier's samples used in the mitogenome analysis were extracted using a KingFisher Duo™ (Thermo Scientific™) automated extraction and purification instrument, following the manufacturer's instructions. Extracted DNA was checked for quality using a NanoDrop™ (Thermo Scientific™) spectrophotometer and gel electrophoresis, and quantified using a Qubit (Invitrogen) fluorometer.

As the protocol for double-digest restriction site-associated DNA sequencing (ddRAD) requires both high molecular weight and high concentration DNA (>20ng/ul), a scoring system was developed to rank samples (Table S1.1) prior to preparing libraries for pooling and sequencing. DNA was run on 1.2% agarose gels to assess the overall quality of the sample and the

concentration was measured using spectrophotometry (NanoDrop) to obtain an approximate value. Some samples that yielded poor quality DNA were extracted up to two more times (n=4 Blainville's, n=68 Cuvier's). More precise measurements of DNA quantity were made using fluorometry (Qubit) for n=88 Blainville's and n=302 Cuvier's that either had visible amounts of DNA on the agarose gel or quantifiable amounts of DNA on the spectrophotometer. Samples selected based on their DNA score were pooled into libraries with individuals that shared the same score whenever possible.

Samples were also selected to ensure every geographical location possible was covered, and to fill in the rest of the sequencing lanes, samples from well-studied resident populations were prioritised (Hawai'i, Canary Islands, Bahamas, and Ligurian Sea). Although of poorer quality, DNA samples with scores lower than 'Good' were included in the libraries as many of them came from poorly sampled areas. Table S1.2 shows the geographical origin for the 170 Cuvier's and 55 Blainville's samples that were sequenced. Following sequencing and the bioinformatic steps outlined in Supplementary 3 and 4, a number of individuals was removed from the analysis that failed to pass quality control (QC). Table S1.3 shows the number of individuals from both Cuvier's and Blainville's that either passed or failed QC according to the score assigned to them based on DNA quality/quantity before sequencing.

The final list of indivdiuals used in this study is found in Supplementary Table 1 (SST1). Once duplicate individuals were removed, the number of unique individuals for ddRAD sequencing remaining was *n*=161 Cuvier's and *n*=55 Blainville's.

Table S1.1. The scoring system developed to rank DNA samples before ddRAD library preparation based on the amount of DNA in the sample (measured using a Qubit fluorometer) and the molecular weight according to 1.2% agarose gels (HMW =High Molecular Weight, Smear=degraded DNA of varying sizes, LMW= Low Molecular Weight).

Score	DNA Concentration (Qubit)	DNA Gel Result
Great	>20ng/ul	HMW
Good	>20ng/ul	HMW + smear
Good	>20ng/ul	Smear
Good	>20ng/ul	No visible DNA
Good	15-20ng/ul	HMW
OK	>20ng/ul	LMW
OK	15-20ng/ul	Smear
OK	<15ng/ul	HMW
Some	<15ng/ul	Faint HMW
Some	<15ng/ul	Smear
Some	<15ng/ul	No visible DNA

Table S1.2. Broad geographic origin of the samples selected for ddRAD library preparation and sequencing for *n*=170 Cuvier's (*Ziphius cavirostris*) and *n*=55 Blainville's beaked whales (*Mesoplodon densirostris*).

	Cuvier's (*n*=170)	Blainville's (*n*=55)
North Atlantic	89	34
North Pacific	28	8
South Pacific	15	8
South Africa	2	5
Mediterranean	36	Not present

Table S1.3. The number of Cuvier's (*Ziphius cavirostris*) and Blainville's beaked whales (*Mesoplodon densirostris*) that passed or failed the ddRAD quality control steps based on the DNA quality/quantity score described in Table S1.1.

	Cuvier's (*n*=170)		Blainville's (*n*=55)	
Quality Score	No. Passed QC	No. Failed QC	No. Passed QC	No. Failed QC
Great	102	29	39	9
Good	19	15	2	0
OK	2	3	1	3
Some	0	0	1	0

ESM2: ddRAD and Mitogenome Library Preparation and Sequencing

ddRAD builds upon the earlier RADseq method (Baird et al., 2008) by adding a second restriction enzyme (RE) to the digest and an explicit size-selection step, allowing researchers to have more control over the fraction of the genome that is sequenced (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). In this protocol, samples were digested with one RE that targeted a commonly occurring motif (MspI, 4bp) and a rarely occurring motif (HindIII, 6bp). Unique P1 adaptors containing individual barcodes and PCR primers, and universal P2 adaptors with PCR primers, were ligated to both ends of the digested DNA. The samples were then cleaned, pooled and size selected using a Pippin Prep (Sage Science). A PCR step added a secondary identifier (reverse index) to the P2 end and Illumina flow cell annealing sequences on both ends. After this step, samples were cleaned, pooled and sequenced. By adding a second RE digest, and eliminating random mechanical shearing and broad size selection, individual studies are much more reproducible and precise. Limiting the DNA window that is sequenced and precisely selecting for size, means that the sequenced fragments from different individuals are more likely to be recovered from the same region of the genome (Peterson et al., 2012).

The following ddRAD protocol was optimised for beaked whale tissue by Carroll et al., (2021, 2016). Samples selected for sequencing were grouped based on their score, normalised to 20ng/µl and a total of 250ng of DNA per individual underwent an overnight digestion at 37°C with MspI and HindIII. After a 20-minute heat kill step at 65°C, adaptors were ligated with one of 10 forward barcodes per sample, using the temperature profile of 22°C for 2 hours followed by 65°C for 20 minutes. With unique barcodes now ligated, up to 10 samples, grouped according to quality classification, were pooled to form a library and cleaned using three PureLink PCR Micro Kit columns (Invitrogen) per library. Following the final elution step, 30µl of cleaned ligate underwent size selection to a 300-400 bp range using a Pippin Prep (Sage Science). The resulting size-selected ligate was divided into 8 wells and library-specific reverse indices were annealed during low-cycle number PCR using a Phusion High-Fidelity PCR kit (ThermoFisher). PCR products were pooled and cleaned using AMPURE-XP (Beckman-Coulter) beads and eluted to a final volume of 15µl in EB buffer (Qiagen). The final libraries were sent to the National High-Throughput DNA Sequencing Centre at the University of Copenhagen where the quantities and quality were determined with qPCR and Bioanalyzer (Agilent Genomics). Finally, the libraries were normalised and up to five libraries (~50 samples) were pooled into sequencing lanes, and sequenced on a HiSeq2500 V4 chemistry (Illumina®).

For the mitogenome sequencing, we used the Carøe et al. (2018) protocol. Briefly, DNA was fragmented to approximately 350 bp, using the M220 Focused-Ultrasonicator™ (Covaris), according to the manufacturer's protocol. After fragmentation, samples were quantified on the Agilent 2200 TapeStation according to the protocol for genomic DNA, to verify fragmentation success. DNA libraries were built, using the blunt-end single-tube protocol described by Carøe et al. (2018) with a few modifications. To each library, 2 µl of 10 µM Illumina® adapters were added to the fragmented DNA, followed by a MiniElute (Qiagen) clean-up step before indexing with P5 and P7 indices. Libraries were sequenced using paired-end (PE) 150 bp chemistry on two lanes of Illumina® HiSeq 4000 at the National High-Throughput Sequencing Centre at

University of Copenhagen, Denmark. In addition, 16 libraries characterized by average fragment lengths <300bp were sequenced on a single lane of HiSeq4000 using single-end 80 bp chemistry.

ESM3: ddRAD **STACKS** Parameter Optimisation

The STACKS SNP discovery pipeline (Rochette, Rivera-Colón, & Catchen, 2019) implements the Bayesian genotype caller (BGC) algorithm of Maruki & Lynch (2015, 2017) in the command 'gstacks'. This algorithm uses a Bayesian genotype-frequency prior that takes into account population-level allele frequencies, does not assume Hardy-Weinberg equilibrium and estimates error rates directly from the sequence data (not from read quality scores) when calling genotypes (Maruki & Lynch, 2015, 2017). The program first estimates significant polymorphic loci from the read alignments with confidence set by the parameter "var-alpha" in 'gstacks' (Maruki & Lynch, 2015; Rochette et al., 2019). Genotypes for each individual are called at these loci, using a method that takes into account allele balance and read depth (Maruki & Lynch, 2017). Confidence in the genotype calling is done using a likelihood ratio test which compares the likelihood of the two most likely genotypes (Maruki & Lynch, 2015).

A subset of demultiplexed and quality-controlled (QC) reads from Cuvier's (*n*=40) and Blainville's *(n*=55) individuals was selected to optimize the parameters used in the STACKS SNP discovery pipeline. The subset of high-quality samples (>98% retained reads and >1,000,000 retained reads) were selected to cover the widest geographical range of Cuvier's and reduce computing time, while all Blainville's individuals were selected. In summary, the following parameters were modified to optimise the 'gstacks' command for each dataset: "min-mapq" (minimum mapping quality score to consider a read; 10,20), "max-clipped" (maximum soft-clipping level as a fraction of the read length; 0.1, 0.2), "var-alpha" (SNP discovery threshold; 0.05, 0.01) and "gt-alpha" (genotype calling threshold; 0.05, 0.01). Samples were selected for optimisation based on the number and proportion of retained reads and to ensure an even distribution from all geographical regions. The best combination of parameters was selected based on resulting datasets with the highest number of SNP loci and the lowest amount of missing data.

At the end of all STACKS and filtering steps, the optimal parameters were selected based on maximizing the total number of final SNP loci and reducing the amount of missing data per species. The final optimised parameters were the same for both Cuvier's and Blainville's samples: mapq=10, sclip=0.2, var_alpha=0.05, gt_alpha=0.05.

ESM4: ddRAD STACKS Protocol, Loci Quality Control and Filtering Steps

Following the discovery and genotyping of SNPs in 'gstacks', individuals and their genotyped loci can be analysed in a framework incorporating some sort of group assignment (such as geographical origin or sex) and then filtered according to minor allele frequency or locus frequency within the entire population in the STACKS 'populations' module. To reduce bias from potentially arbitrary population designations, no such population framework was provided in the current study. Massively parallel sequencing can lead to high error rates and genotypic uncertainties that can be introduced at any step throughout the analysis (O'Leary, Puritz, Willis, Hollenbeck, & Portnoy, 2018). Fortunately, many errors can be overcome by employing rigorous filtering to identify and reduce errors before analyzing the final dataset (O'Leary et al., 2018). In this study, we took a tiered approach to filtering, starting with low cut-off values for missing data (applied separately per locus and individual) and finalizing the dataset with higher thresholds. This alternative and iterative filtering method, whereby you increase the cut-off threshold, has been shown to retain more loci and individuals as poor-quality individuals can deflate genotype call rates in otherwise acceptable loci, while poor-quality loci can increase the amount of missing data in otherwise acceptable individuals (O'Leary et al., 2018). Below we describe each of the steps that were implemented using R v. 3.6.0 (R Core Team, 2019) and VCFTOOLS v. 0.1.12a (Danecek et al., 2011) to filter individuals and loci based on the amount of missing data, read depth, and quality score (Table S4.1). In Table S4.2, each bioinformatic step is listed, with the resulting number of loci and individuals remaining throughout the process for both Cuvier's and Blainville's datasets.

Table S4.1. List of filtering commands and steps used in the program VCFTOOLS to filter loci and individuals based on locus depth, genotype quality, minor allele frequency (MAF) and missingness. Low stringency indicates that lower cut-off values are used to filter out missing data before iteratively increasing cut-off values, a strategy shown by O'Leary et al. (2018) to increase the proportion of retained loci and individuals.

Stringency	VCFTOOLS Command	Description
Low	--minDP 5 –minGQ20	Recode genotypes with quality <20 and depth <5 to zero
	--maf 0.001	Remove the sites made monomorphic by previous step.
	--max-missing 0.5	Remove sites with >50% missing data
	--missing-indv	Calculate missingness per individual, write a list of individuals with >50% missing data
	--remove	Remove individuals on list with >50% missing data
High	--site-depth	Calculate site depth, write a list of loci with mean site depth >3x the overall mean
	--exclude-positions	Remove sites with site depth >3x the overall mean
	--max-missing 0.75	Remove sites with more than 75% missing data
	--missing-indv	Calculate missingness per individual, write a list of individuals with >25% missing data
	--remove	Remove individuals on list with >25% missing data

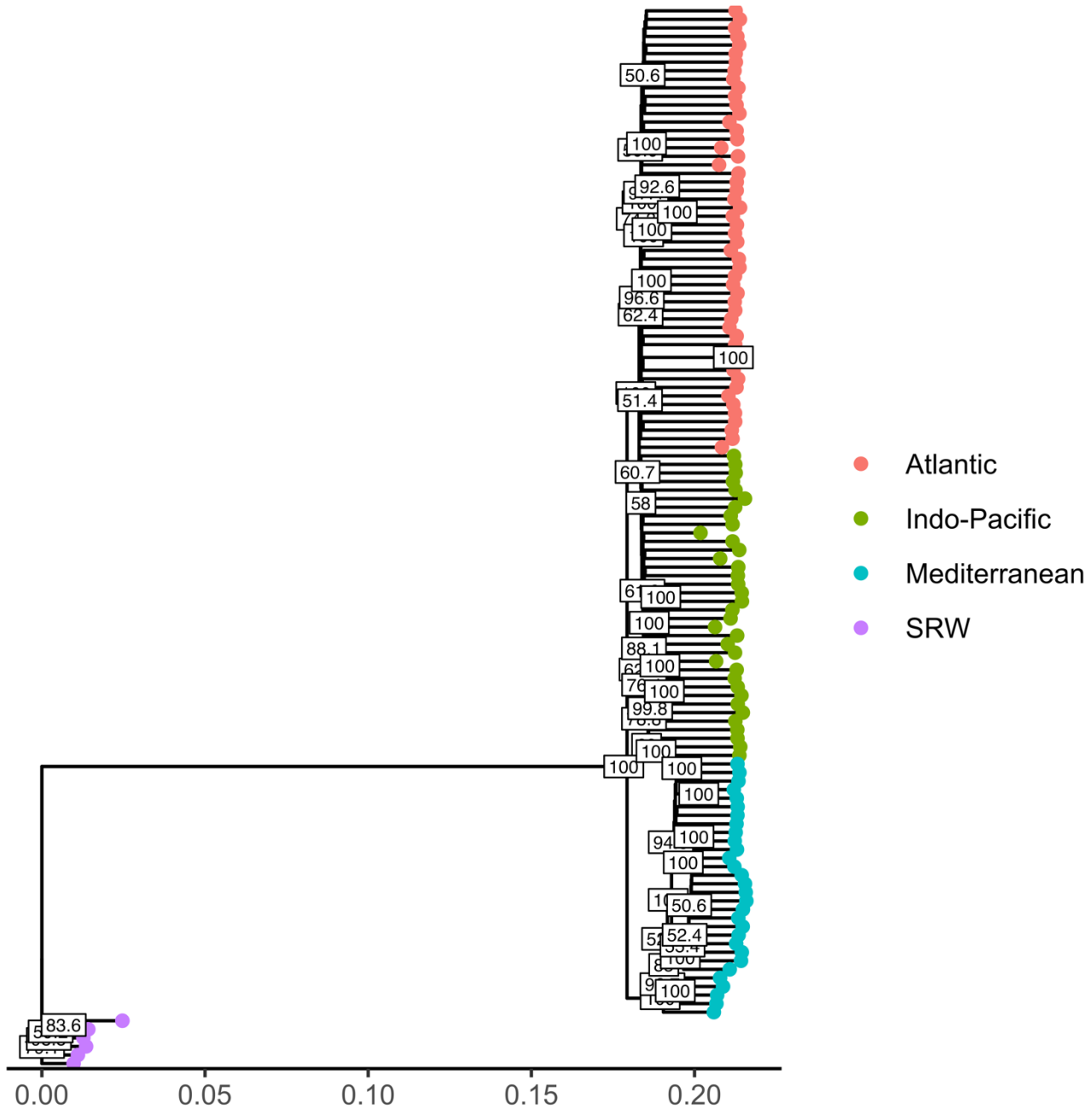
Table S4.2. Summary of each bioinformatic step to discover, genotype and filter loci based on the steps described in Table 1. Data are presented for the Cuvier's and Blainville's beaked whale (*Ziphius cavirostris* and *Mesoplodon densirostris,* respectively) datasets with a summary of each step and the program that was used.

	Species	Cuvier's	Blainville's
	starting sample size	170	56
optimised gstacks parameters	mapq	10	10
	sclip	0.2	0.2
	var_alpha	0.05	0.05
	gt_alpha	0.05	0.05
Process RadTags (STACKS)	Total PE reads	1112692388	339961638
	Retained PE reads	1063993201	322949310
	Mean PE retained reads/sample	6258784	5766952
	% Retained (across all samples)	95.6%	95.0%
gstacks (STACKS)	Individuals Remaining	154	54
	Alignments Read	1085038314	334888850
	Alignments Kept	778173558	249685545
	% Alignments Kept	71.7%	74.6%
	Loci Built and Genotyped	1801998	979828
Populations (STACKS)	Loci Kept	1795750	977060
	No. Sites	477045540	184528823
	No. Polymorphic Sites	2087305	1054577
Filtering (VCFTOOLS)	Sites with >5x depth, >20 genotype quality, >0.001 MAF	1326391	735883
	Sites with <50% missing data	327270	340095
	Sites with depth <3x overall mean depth	326459	339441
	Sites with <75% missing data	262482	296250
Populations (STACKS)	Loci passed filtering (whitelist)	34264	37617
	Loci present in >80% individuals with >0.01 MAF	31734	32610
	No. Sites	9994609	9527357
	No. Polymorphic Sites	30479	271983
	No. Genomic Sites	9610872	9504054
	No. Individuals remaining	123	49
glPlot/Removing Duplicates ('Adegenet')	Final no. Individuals	123	43
	Final no. Loci	30479	13988

ESM5: Phylogenetic Trees with Southern Right Whale Outgroups

Phylogenetic trees using the ddRAD SNP data were generated for both Cuvier's and Blainville's incorporating data from Southern right whales (SRW, *Eubalaena australis*) as the outgroup. The full SNP discovery pipeline as described in SM4 was repeated for the final n=123 Cuvier's and n=43 Blainville's, each time including sequence data from six SRWs. The Cuvier's + SRW dataset was aligned to the same Cuvier's genome (NCBI Genbank database accession: PRJNA399469) and the Blainville's + SRW sequences were aligned to the same Sowerby's genome (*Mesoplodon bidens*: PRJNA399476). All bioinformatic steps were followed as before, with one exception.

The final VCF file was uploaded into R and converted to a "genlight" file as before and BIONJ trees with bootstrap support (in %, based on 100 bootstraps) were produced in R ('poppr' v2.8.5; Kamvar, Brooks, & Grünwald, 2015 and 'ggtree'v.2.0.2; Yu, Smith, Zhu, Guan, & Lam, 2017). The trees were rooted using one of the SRW sequences and re-plotted. The SRW individuals were then dropped from the tree, to better visualize the ocean-level phylogenetic patterns of Cuvier's and Blainville's. The original BIONJ trees (without dropping the SRW outgroup are found in Figures S5.1 and S5.2. The final datasets (including 6 SRW samples) were n=118 Cuvier's (n=33137 SNPs) and n=42 Blainville's (n=29904 SNPs).



297 Figure S5.1. BIONJ phylogenetic tree of 118 Cuvier's beaked whales (*Ziphius cavirostris*) and 6
298 Southern right whales (*Eubalaena australis*) as outgroups, generated using n=33137 ddRAD
299 SNPs.

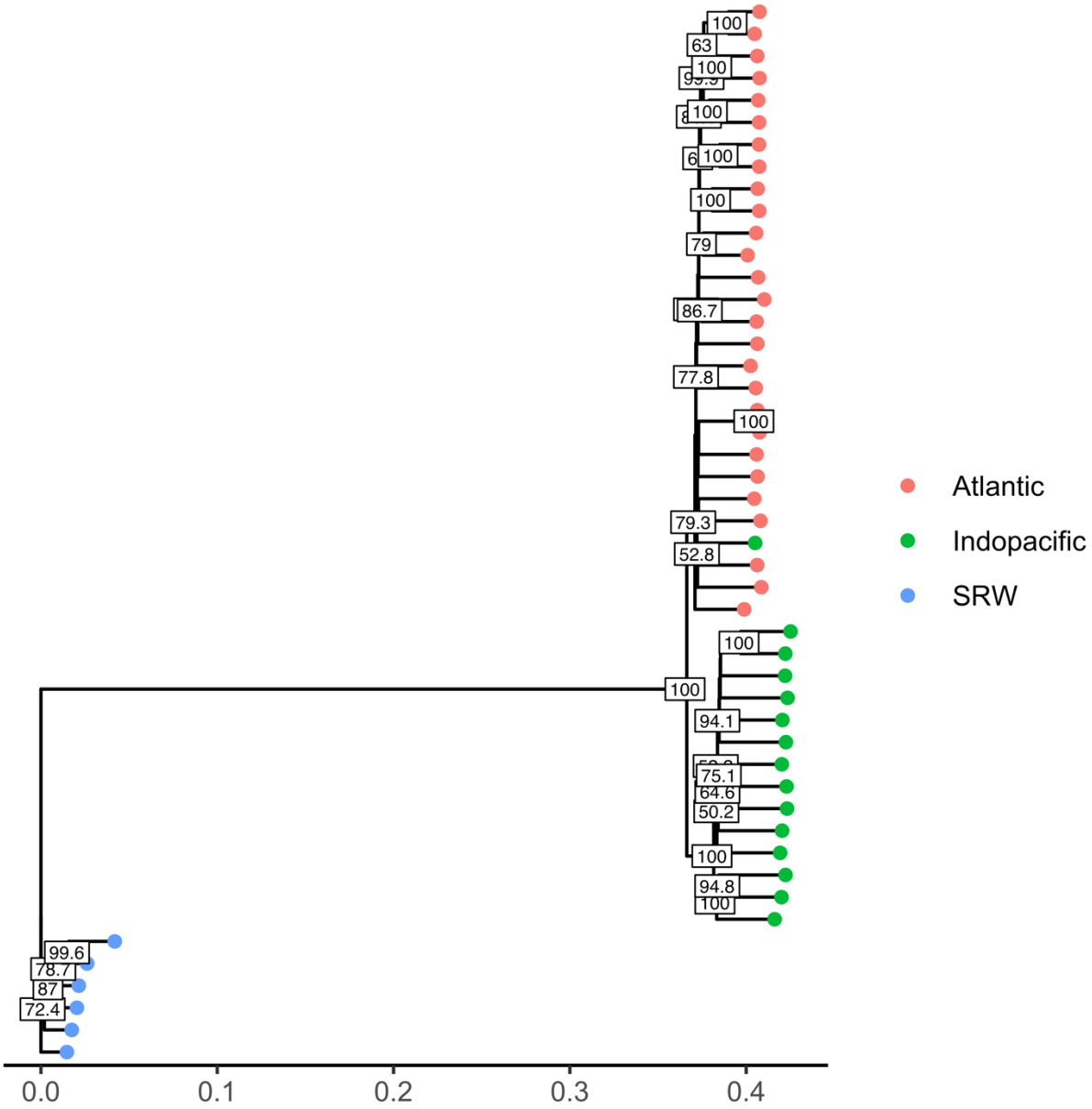


Figure S5.2. BIONJ phylogenetic tree of 42 Blainville's beaked whales (*Mesoplodon densirostris*) and 6 Southern right whales (*Eubalaena australis*) as outgroups, generated using n=329904 ddRAD SNPs.

ESM6: Isolation-by-Distance

Isolation by distance (IBD) was calculated per species, and within ocean basin per species, using a Mantel test in 'ade4' v1.7-16 in R (Dray & Dufour, 2007) and geographical distances calculated via the least cost (LC) path distance over seawater in 'marmap' v1.0.5 (Pante & Simon-Bouhet, 2013). To calculate the LC distance over seawater, sampling locations were plotted on a global bathymetry map (with 4 minute resolution) using the "getNOAA.bathy" function in 'marmap'. The resolution of the world bathymetry map resulted in some sampling locations of stranded individuals to be on land, and therefore incur a great coast in the LC path. Sample coordinates were therefore adjusted to the nearest -200m isobath using the "dist2isobath" function in 'marmap'. The updated sample coordinates and bathymetry map were used to calculate a transition matrix using "trans.mat" in 'marmap', requiring the LC path to have a minimum depth of 200m. Finally, the LC path distance between each individual was calculated using "lc.dist" in 'marmap'. The resulting pairwise matrix of geographic distances was used in combination with pairwise genetic distance (Euclidean) to run a Mantel test using "mantel.randtest" with 999 permutations. Mantel tests were conducted based on 999 replicates for all Cuvier's or Blainville's combined, and for individual ocean basins (Atlantic, Indo-Pacific, and Mediterranean- Cuvier's only) (Table S6.1). All correlation values were positive and all but the Indo-Pacific Blainville's were significant ($p<0.05$). The genetic and geographic distance matrices were plotted with a 2 dimensional kernel density estimation to visualize whether the apparent IBD was the result of a continuous cline or population clustering (Figures S6.1 and S6.2).

Table S6.1. The observation correlation and associated p-value of Mantel tests for Isolation by distance.

		n	Observation correlation (r)	P-value
Cuvier's	All	123	0.308	0.001
	Atlantic	54	0.154	0.001
	Indo-Pacific	36	0.162	0.028
	Mediterranean	33	0.218	0.002
Blainville's	All	43	0.665	0.001
	Atlantic	28	0.110	0.03
	Indo-Pacific	15	0.014	0.427

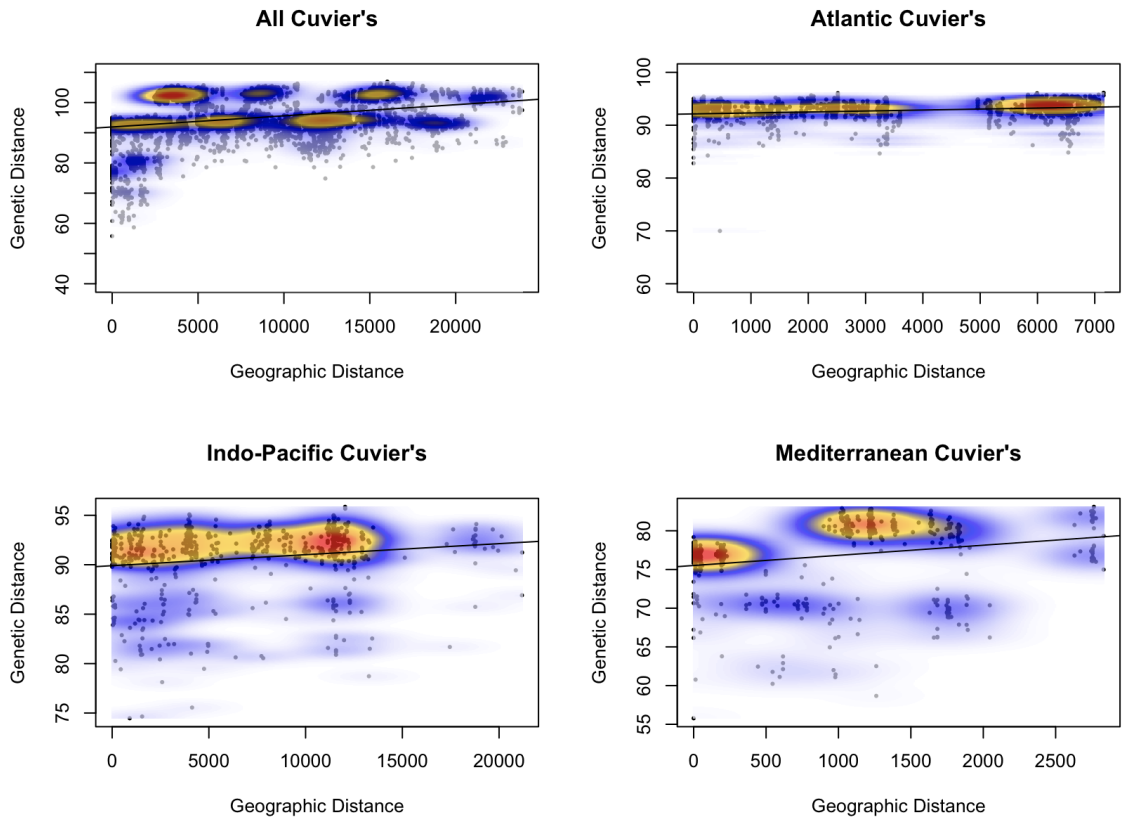


Figure S6.1. Scatterplots of genetic distance (Euclidian) and geographic distance (least cost (LC) path distance over seawater) overlaid with 2-dimensional kernel density estimation from n=123 Cuvier's beaked whales (*Ziphius cavirostris,* n=30479 SNPs).

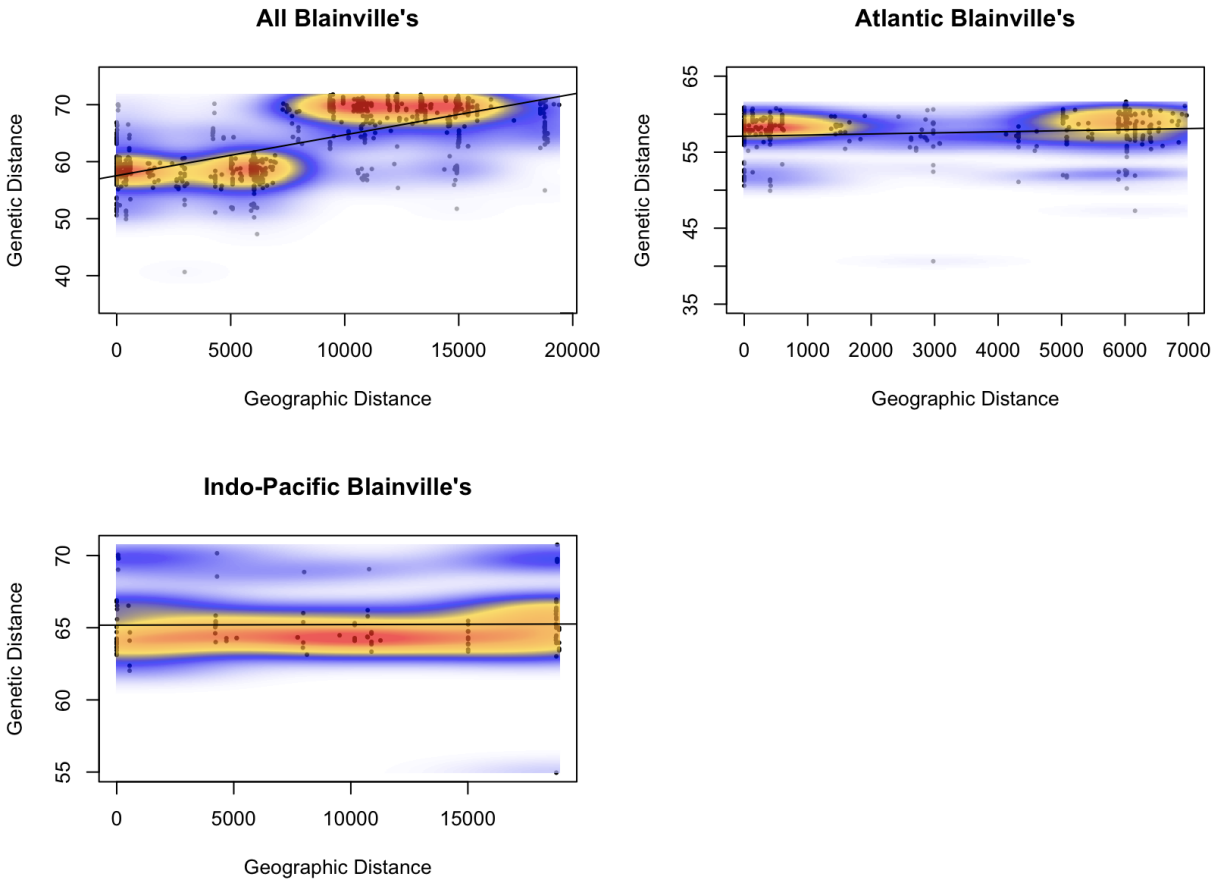
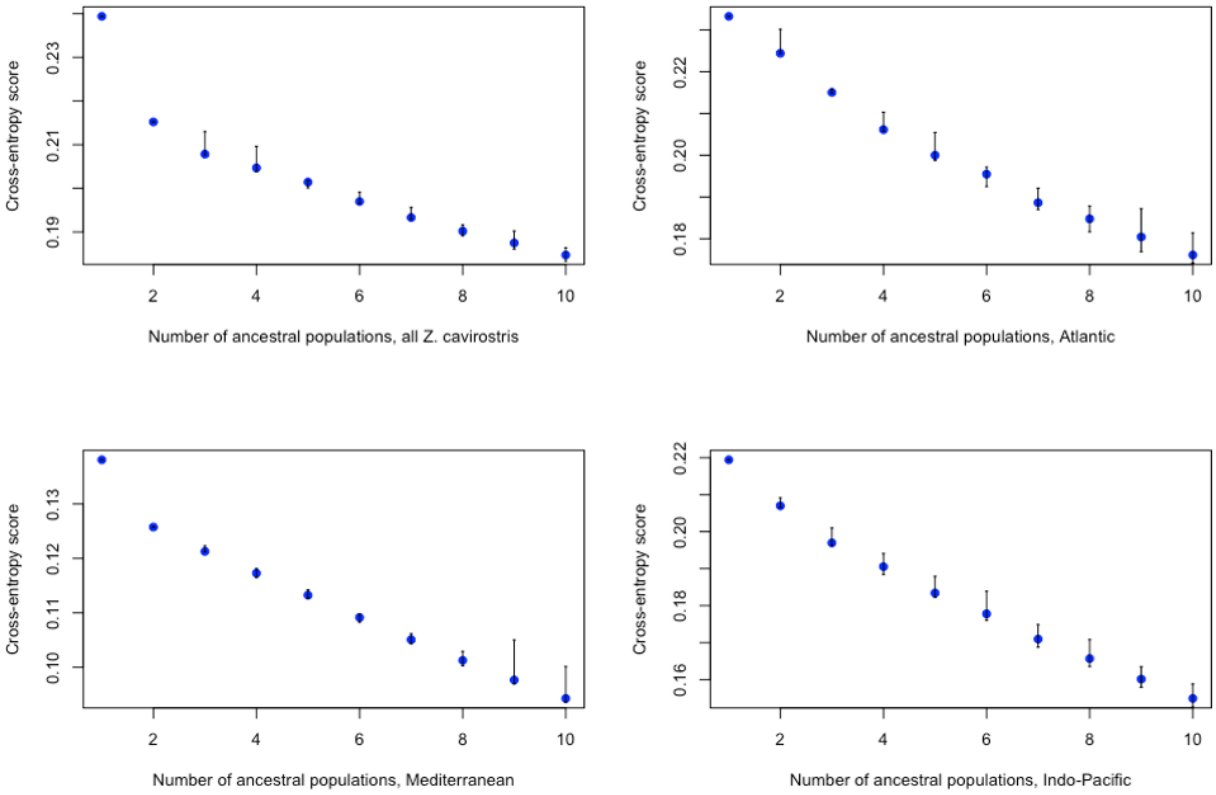


Figure S6.2 Scatterplots of genetic distance (Euclidian) and geographic distance (least cost (LC) path distance over seawater) overlaid with 2-dimensional kernel density estimation from n=43 Blainville's beaked whales (*Mesoplodon densirostris,* n=13988 SNPs).

ESM7: ‘tess3r’ Cross-Entropy Scores

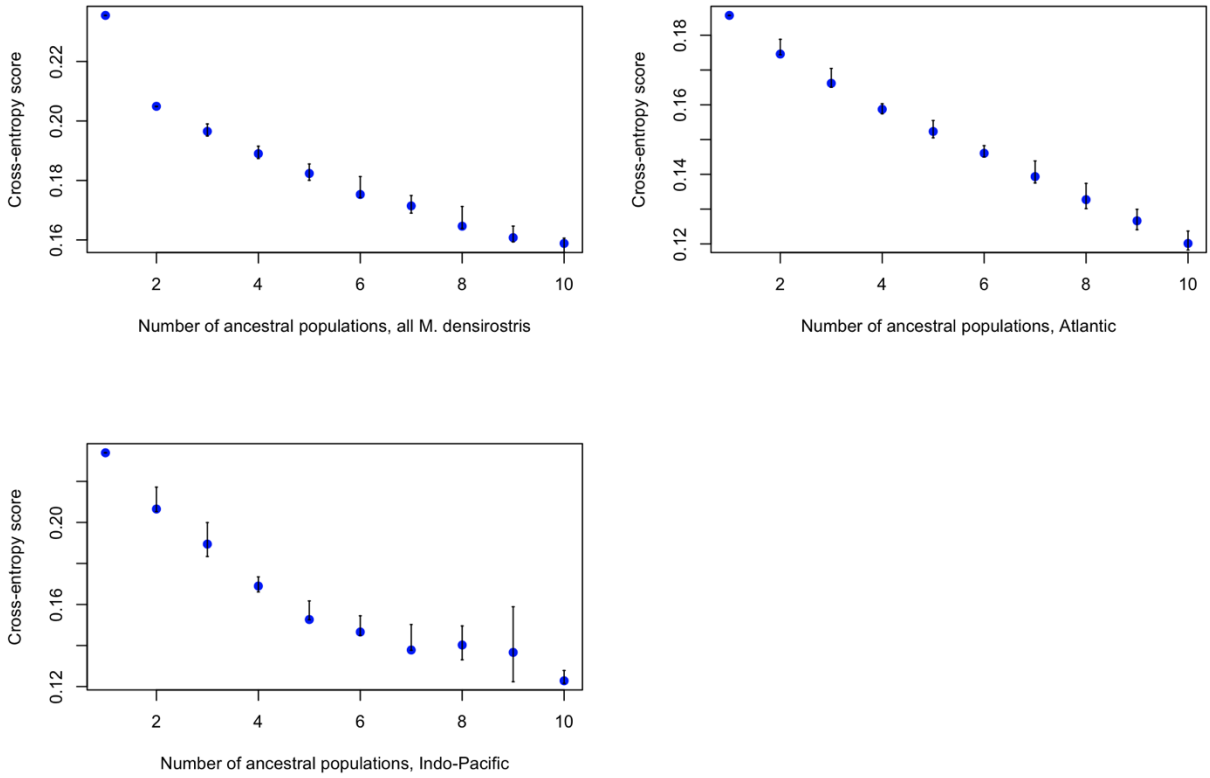
The R package ‘tess3r’ (Caye, Jay, Michel, & Francois, 2018) incorporates genotypic and geographical information (latitude and longitude coordinates for each sample) in a spatially explicit, least-squares optimization approach to estimate ancestry. The user defines and compares results from a range of ancestral population clusters (*K*) with the resulting bar plots displaying ancestry coefficients reflecting the probability of population membership and estimates of admixture. As opposed to the initial version of ‘tess’ (Chen, Durand, Forbes, & François, 2007; Durand, Jay, Gaggiotti, & François, 2009), there is no biological model underlying this version however, the model does expect that individuals sampled in close geographical proximity are more likely to share ancestry than those sampled from further away.

For both Cuvier’s and Blainville’s, tess3r was run for *K*=2-10 and cross-entropy scores were plotted against *K* values to infer the most likely number of genetic clusters. In cross-entropy plots, smaller values indicate better fit with the best estimate of *K* corresponding to the value at which the curve reaches a plateau or starts to increase. In cases where a clear minimum or plateau is not observed, the *K* value that leads to the most parsimonious assignment of individuals (least amount of admixture) to populations can be considered as selection criteria. The figures below display the cross-entropy scores for: Cuvier’s (Figure S7.1) and Blainville’s (Figure S7.2).



401

402 Figure S7.1. Cross-entropy scores of *K*=2-10 genetic clusters generated using 'tess3r' for
403 *n*=123 Cuvier's beaked whales (*Ziphius cavirostris*) sampled from across their global range
404 (top left), the Atlantic (top right), Mediterranean (bottom left) and Indo-Pacific (bottom right).



405

406 Figure S7.2. Cross-entropy scores of *K*=2-10 genetic clusters generated using 'tess3r' for *n*=43
407 Blainville's beaked whales (*Mesoplodon densirostris*) sampled from across their global range
408 (top left), the Atlantic (top right), and Indo-Pacific (bottom left).

409

ESM8: Discriminant Analysis of Principle Components

Discriminant Analysis of Principle Components (DAPC) was conducted in the R package, 'adegenet' (Jombart, Devillard, & Balloux, 2010). DAPC is a useful tool to summarise the amount of genetic differentiation between groups (either determined *a priori* or *de novo* using K-means clustering) while ignoring the amount of variation within groups. To optimise the number of principle components (PCs) to retain from the analysis, 'adegenet' offers a cross-validation tool with "xvalDAPC". This command subsets the data to use as a training set, runs the analysis over a pre-determined number of repeats (n=30), and determines the best number of PCs to retain based on whichever yields the highest predictive success of the training data with the lowest root mean squared error (RMSE). The resulting DAPC can be plotted to observe the spatial structure of SNP genotypes across clusters and the function "assignplot" can be used to visualize the proportion of successful reassignment to the prior groups.

DAPC with cross-validation was conducted for n=118 Cuvier's (excluding Atl_Spain and Indo_Mix). The highest mean success and lowest MSE was achieved when 20 PCs were retained. The resulting scatterplot and assignment plot are found in figure S8.1. DAPC with cross-validation was conducted for n=43 Blainville's and the optimal number of PCs to retain was 10. The resulting scatter and assignment plots are in figure S8.2.

The presence of hierarchical structure requires investigation of more than the first and second axes to resolve finer scales. Scatter plots of the 2^{nd} vs 3^{rd} (Figure S8.3) and 3^{rd} vs 4^{th} (Figure S8.4) axes help to discriminate between the genetic clusters found in Cuvier's within the Atlantic and Indo-Pacific.

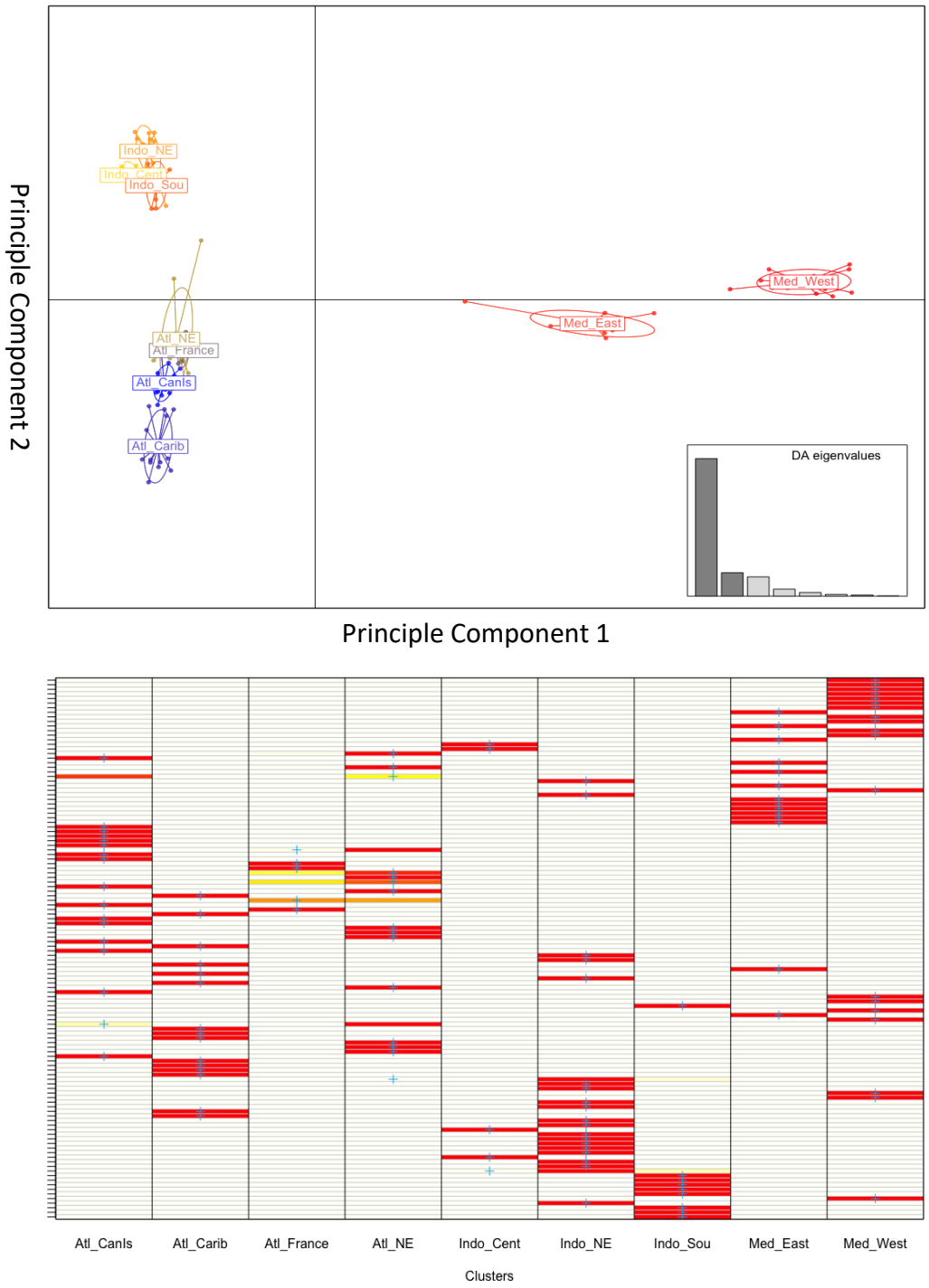


Figure S8.1. DAPC scatter (top) and assignment (bottom) plot for n=118 Cuvier's beaked whales (*Ziphius cavirostris*) generated using cross-validation and retaining 20 PCs. In the assignment plot, each row represents an individual, the blue cross indicates the prior cluster assignment and the colours represent membership probability (red=1, white=0).

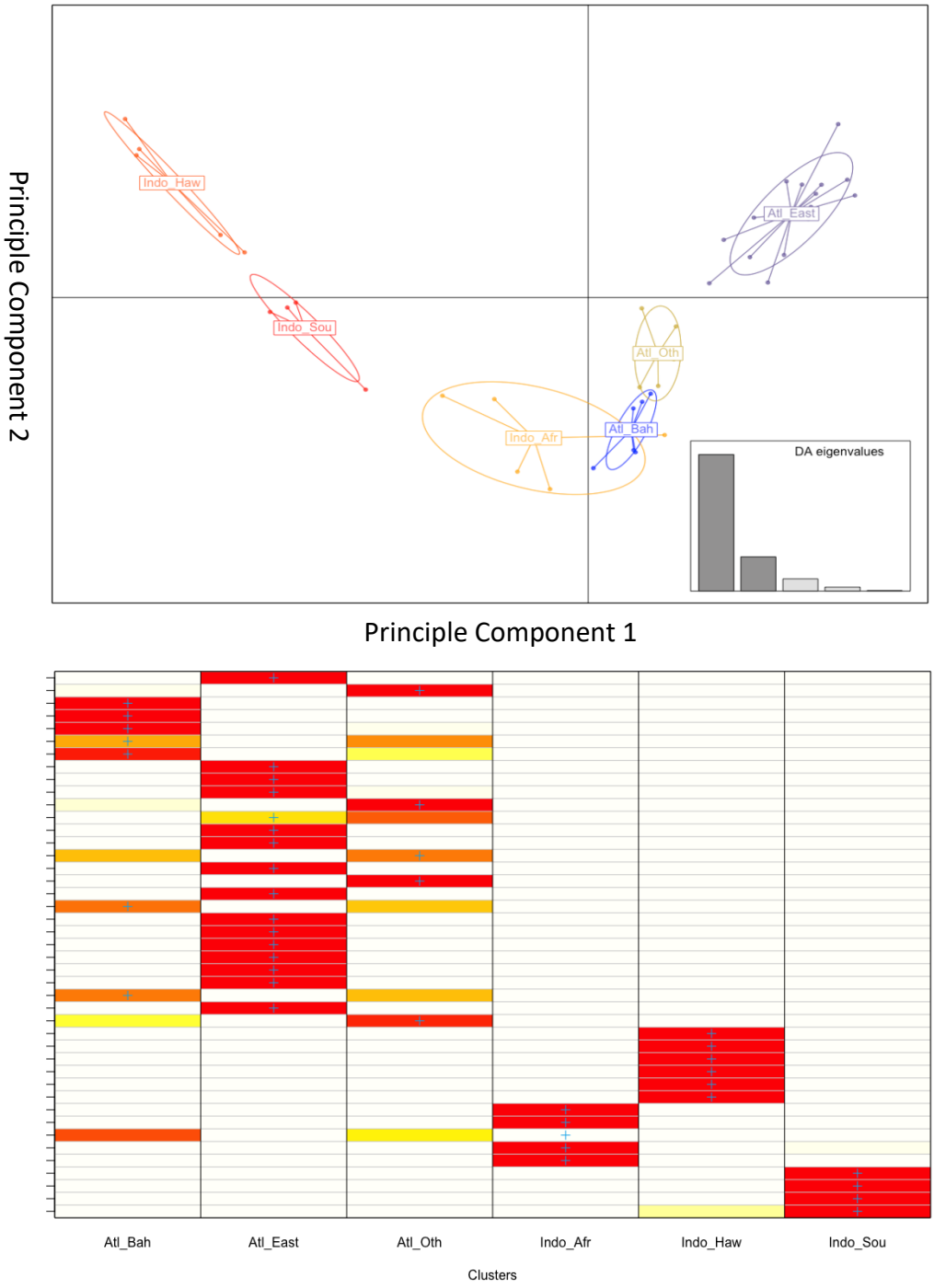
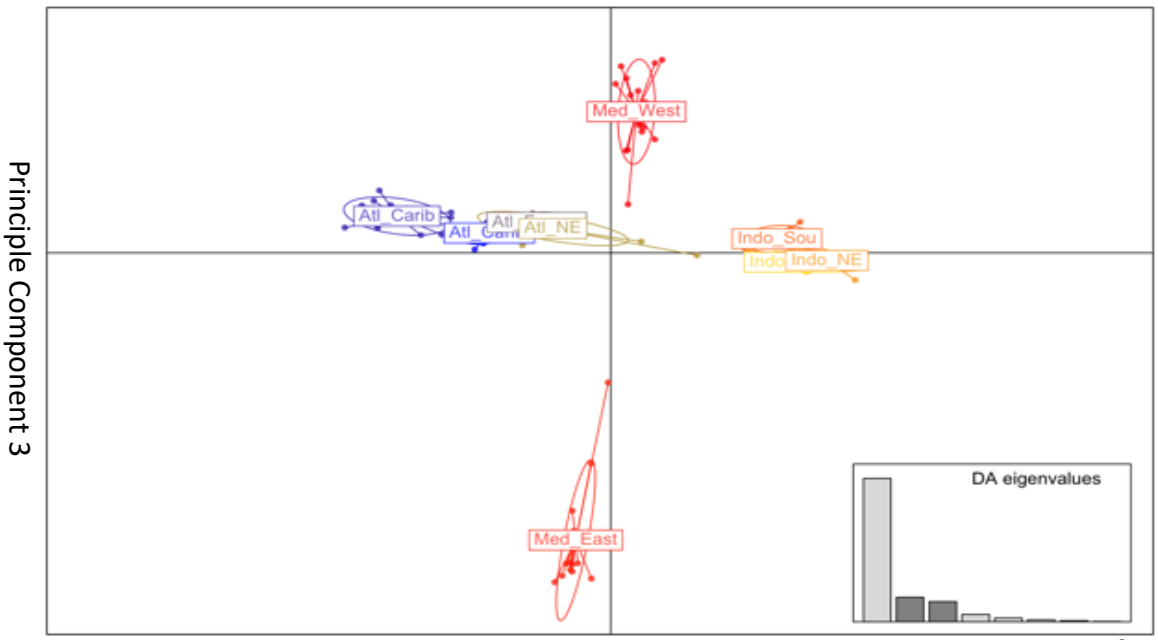


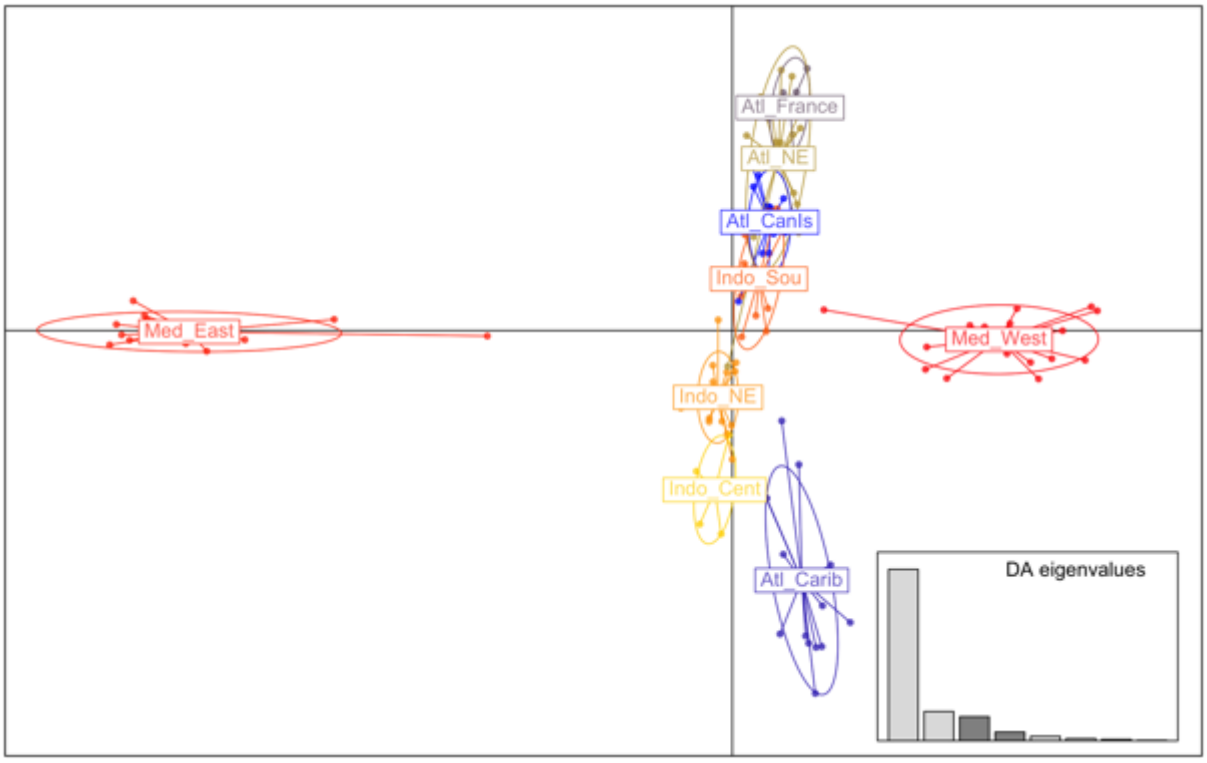
Figure S8.2. DAPC scatter (top) and assignment (bottom) plot for n=43 Blainville's beaked whales (*Mesoplodon densirostris*) generated using cross-validation and retaining 10 PCs. In the assignment plot, each row represents an individual, the blue cross indicates the prior cluster assignment, and the colours represent membership probability (red=1, white=0).



540

541

542 Figure S8.3. DAPC scatter plot (2nd and 3rd axes) for n=118 Cuvier's beaked whales (*Ziphius*
543 *cavirostris*) generated using cross-validation and retaining 20 PCs.



544

545

546 Figure S8.4. DAPC scatter plot (3rd and 4th axes) for n=118 Cuvier's beaked whales (*Ziphius*
547 *cavirostris*) generated using cross-validation and retaining 20 PCs.

ESM9: ddRAD Sequencing Results

From the ITABW collection, 225 individuals were selected for ddRAD analysis (Cuvier's *n*=170, Blainville's *n*=55), balancing DNA quality and quantity, and covering as much of each species' broad geographical ranges as possible (See ESM1 for sample selection process). The samples were run across five HiSeq 2500 lanes, generating a total of 340 million and 1.113 billion PE reads across libraries of Blainville's and Cuvier's samples, respectively. Following demultiplexing and initial QC, 323 million and 1.064 billion PE reads were retained for Blainville's and Cuvier's samples, respectively. The number of reads and proportion of those retained were consistent across libraries. In the Blainville's libraries (*n*=7), the mean number of retained reads was 5.77 million per sample (97% retained, standard error of the mean (SE)=1.4%) and in the Cuvier's libraries (*n*=18), the mean number of retained reads was 6.26 million per sample (94.6% retained, SE=1.0%).

The final ddRAD QC dataset included 123 Cuvier's individuals and 30,479 (72.4% individuals retained) and 43 Blainville's individuals and 13,988 SNPs (76.8% individuals retained) (Figure 1). Each Cuvier's individual was genotyped at an average of 29,697 (SE=90.5) SNPs with a mean per locus read depth of 59x (SE=3.6). Each Blainville's individual was genotyped at an average of 13,760 (SE=63.59) SNPs with a mean per locus read depth of 53x (SE=4.4). Overall, both datasets had low levels of missing data (2.6% in Cuvier's and 1.6% in Blainville's). Only SNPs with a genotype quality greater than 20 were kept (99% base call accuracy).

Supplementary References

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), 1–7. https://doi.org/10.1371/journal.pone.0003376

Baker, C. S., Slade, R. W., Bannister, J. L., Abernethy, R. B., Weinrich, M. T., Lien, J., … Palumbi, S. R. (1994). Hierarchical structure of mitochondrial DNA gene flow among humpback whales Megaptera novaeangliae, world-wide. *Molecular Ecology*, *3*(4), 313–327. https://doi.org/10.1111/j.1365-294X.1994.tb00071.x

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M.-H. S., Samaniego, J. A., … Gilbert, M. T. P. (2018). Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution*, *9*(2), 410–419. https://doi.org/10.1111/2041-210X.12871

Carroll, E. L., McGowen, M. R., McCarthy, M. L., Marx, F. G., Aguilar de Soto, N., Dalebout, M. L., … Olsen, M. T. (2021). Speciation in the deep: Genomics and morphology reveal a new species of beaked whale Mesoplodon eueu. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1961). https://doi.org/10.1098/rspb.2021.1213

Carroll, E. L., Reyes, C., Gaggiotti, O. E., Olsen, M. T., Maaholm, D. J., Rosso, M., … Aguilar de Soto, N. (2016). 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. *International Whaling Commission*, (July). https://doi.org/10.13140/RG.2.1.2286.5527

Caye, K., Jay, F., Michel, O., & Francois, O. (2018). Fast inference of individual admixture coefficients using geographic data. *Annals of Applied Statistics*, *12*(1), 586–608. https://doi.org/https://doi.org/10.1214/17-AOAS1106

Chen, C., Durand, E., Forbes, F., & François, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study. *Molecular Ecology Notes*, *7*(5), 747–756. https://doi.org/10.1111/j.1471-8286.2007.01769.x

Dalebout, M. L. (2002). *Species identity, genetic diversity and molecular systematic relationships among the Ziphiidae (beaked whales)*. University of Auckland.

Dalebout, M. L., Baker, C. S., Mead, J. G., Cockcroft, V. G., & Yamada, T. K. (2004). A comprehensive and validated molecular taxonomy of beaked whales, family Ziphiidae. *Journal of Heredity*, *95*(6), 459–473. https://doi.org/10.1093/jhered/esh054

Dalebout, M. L., Baker, C. S., Steel, D. J., Robertson, K. M., Chivers, S. J., Perrin, W. F., … Schofield Jr., T. D. (2007). A divergent mtDNA lineage among Mesoplodon beaked whales: Molecular evidence for a new species in the tropical pacific? *Marine Mammal Science*, *23*(4), 954–966. https://doi.org/10.1111/j.1748-7692.2007.00143.x

Dalebout, M. L., Baker, C. S., Steel, D. J., Thompson, K. F., Robertson, K. M., Chivers, S. J., … Yamada, T. K. (2014). Resurrection of Mesoplodon hotaula Deraniyagala 1963: A new species of beaked whale in the tropical Indo-Pacific. *Marine Mammal Science*, *30*(3), 1081–1108. https://doi.org/10.1111/mms.12113

Dalebout, M. L., Mead, J. G., Baker, C. S., Baker, A. N., & van Helden, A. L. (2002). A new species of beaked whale Mesoplodon perrini sp. n. (Cetacea: Ziphiidae) discovered through phylogenetic analyses of mitochondrial DNA sequences. *Marine Mammal Science*, *18*(3),

577–608. https://doi.org/10.1111/j.1748-7692.2002.tb01061.x
Dalebout, M. L., Robertson, K. M., Frantzis, A., Engelhaupt, D., Mignucci-Giannoni, A. A., Rosario-Delestre, R. J., & Baker, C. S. (2005). Worldwide structure of mtDNA diversity among Cuvier's beaked whales (*Ziphius cavirostris*): Implications for threatened populations. *Molecular Ecology*, *14*(11), 3353–3371. https://doi.org/10.1111/j.1365-294X.2005.02676.x
Dalebout, M. L., Ross, G. J. B., Baker, C. S., Anderson, R. C., Best, P. B., Cockcroft, V. G., … Pitman, R. L. (2003). Appearance, Distribution, and Genetic Distinctiveness of Longman's Beaked Whale, *Indopacetus pacificus*. *Marine Mammal Science*, *19*(July), 421–461.
Dalebout, M. L., Ruzzante, D. E., Whitehead, H., & Øien, N. I. (2006). Nuclear and mitochondrial markers reveal distinctiveness of a small population of bottlenose whales (*Hyperoodon ampullatus*) in the western North Atlantic. *Molecular Ecology*, *15*(11), 3115–3129. https://doi.org/10.1111/j.1365-294X.2006.03004.x
Dalebout, M. L., Steel, D. J., & Baker, C. S. (2008). Phylogeny of the beaked whale genus *Mesoplodon* (Ziphiidae: Cetacea) revealed by nuclear introns: Implications for the evolution of male tusks. *Systematic Biology*, *57*(6), 857–875. https://doi.org/10.1080/10635150802559257
Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330
Dray, S., & Dufour, A. B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, *22*(4), 1–20. https://doi.org/10.18637/jss.v022.i04
Durand, E., Jay, F., Gaggiotti, O. E., & François, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution*, *26*(9), 1963–1973. https://doi.org/10.1093/molbev/msp106
Gomerčić, H., Duras Gomerčić, M., Gomerčić, T., Lucić, H., Dalebout, M. L., Galov, A., … Huber, D. (2006). Biological aspects of Cuvier's beaked whale (*Ziphius cavirostris*) recorded in the Croatian part of the Adriatic Sea. *European Journal of Wildlife Research*, *52*(3), 182–187. https://doi.org/10.1007/s10344-006-0032-8
Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(94), 1–15. https://doi.org/10.1186/1471-2156-11-94
Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, *6*(JUN), 1–10. https://doi.org/10.3389/fgene.2015.00208
Krützen, M., Barré, L. M., Möller, L. M., Heithaus, M. R., Simms, C., & Sherwin, W. B. (2002). A biopsy system for small cetaceans: Darting success and wound healing in *Tursiops* spp. *Marine Mammal Science*, *18*(4), 863–878. https://doi.org/10.1111/j.1748-7692.2002.tb01078.x
Lambertsen, R. H. (1987). A biopsy system for large whales and it's use for cytogenetics. *Journal of Mammalogy*, *68*(2), 443–445.
Maruki, T., & Lynch, M. (2015). Genotype-frequency estimation from high-throughput

sequencing data. *Genetics*, *201*(2), 473–486. https://doi.org/10.1534/genetics.115.179077
Maruki, T., & Lynch, M. (2017). Genotype calling from population-genomic sequencing data. *G3: Genes, Genomes, Genetics*, *7*(5), 1393–1404. https://doi.org/10.1534/g3.117.039008
O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, (June), 3193–3206. https://doi.org/10.1111/mec.14792
Pante, E., & Simon-Bouhet, B. (2013). marmap: A Package for Importing, Plotting and Analyzing Bathymetric and Topographic Data in R. *PLoS ONE*, *8*(9), 6–9. https://doi.org/10.1371/journal.pone.0073051
Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0037135
R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754. https://doi.org/10.1111/mec.15253
Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). *Molecular Cloning, A Laboratory Manual. Second Edition*. Cold Spring Harbror Laboratory Press, Cold Spring Harbor, New York.
Van Helden, A. L., Baker, A. N., Dalebout, M. L., Reyes, J. C., Van Waerebeek, K., & Baker, C. S. (2002). Resurrection of Mesoplodon traversii (Gray, 1874), senior synonym of M. bahamondi Reyes, Van Waerebeek, Cardenas and Yanez, 1995 (Cetacea: Ziphiidae). *Marine Mammal Science*, *18*(July), 609–621.
Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*, 28–36. https://doi.org/10.1111/2041-210X.12628