MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 254–268 doi: 10.1111/1755-0998.12059

Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics

BRITTANY L. HANCOCK-HANSER,* AMY FREY,* MATTHEW S. LESLIE,† PETER H. DUTTON,* FREDERICK I. ARCHER* and PHILLIP A. MORIN*

**Protected Resources Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 8901 La Jolla Shores Drive, La Jolla, CA 92037, USA, †Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive, MC 0202, La Jolla, CA 92093, USA*

Abstract

Next-generation sequencing (NGS) is emerging as an efficient and cost-effective tool in population genomic analyses of nonmodel organisms, allowing simultaneous resequencing of many regions of multi-genomic DNA from multiplexed samples. Here, we detail our synthesis of protocols for targeted resequencing of mitochondrial and nuclear loci by generating indexed genomic libraries for multiplexing up to 100 individuals in a single sequencing pool, and then enriching the pooled library using custom DNA capture arrays. Our use of DNA sequence from one species to capture and enrich the sequencing libraries of another species (i.e. cross-species DNA capture) indicates that efficient enrichment occurs when sequences are up to about 12% divergent, allowing us to take advantage of genomic information in one species to sequence orthologous regions in related species. In addition to a complete mitochondrial genome on each array, we have included between 43 and 118 nuclear loci for low-coverage sequencing of between 18 kb and 87 kb of DNA sequence per individual for single nucleotide polymorphisms discovery from 50 to 100 individuals in a single sequencing lane. Using this method, we have generated a total of over 500 whole mitochondrial genomes from seven cetacean species and green sea turtles. The greater variation detected in mitogenomes relative to short mtDNA sequences is helping to resolve genetic structure ranging from geographic to species-level differences. These NGS and analysis techniques have allowed for simultaneous population genomic studies of mtDNA and nDNA with greater genomic coverage and phylogeographic resolution than has previously been possible in marine mammals and turtles.

Keywords: cetacean, conservation, DNA capture array, green sea turtle, next-generation sequencing, population structure

Received 3 February 2012; revision received 30 November 2012; accepted 6 December 2012

Introduction

High-throughput 'next-generation' sequencing techniques are enhancing the way we address phylogenetic and population genetic questions, but applications in nonmodel organisms have been limited by the need for better methods of repeatedly sequencing the same set of loci in tens to hundreds of individuals. Early applications have been based on PCR amplification of loci prior to pooling and sequencing (Gilbert *et al.* 2008; Chan *et al.* 2010; Morin *et al.* 2010; Vilstrup *et al.* 2011). A recent development for broad genome coverage of short anonymous nuclear fragments is restriction-site associated DNA (RAD) sequencing (Baird *et al.* 2008; Elshire *et al.* 2011). These two applications represent the relatively low-throughput and high-throughput approaches to applying Next-generation sequencing (NGS) methods to larger numbers of samples and loci than has been previously practical with Sanger sequencing. For many applications in population genetics and phylogenetics, however, an intermediate approach is needed, where dozens to hundreds of targeted nuclear loci and/or complete mitochondrial genomes are generated from dozens to hundreds of samples. We present here a relatively low-cost multiplexed DNA capture and NGS sequencing approach for simultaneous mitogenome and targeted nuclear locus sequencing applications in nonmodel organisms with limited or no genomic information.

Correspondence: Brittany L. Hancock-Hanser , Fax: +1 (858) 5467003; E-mail: brittany.hancock-hanser@noaa.gov

The use of complete mitogenomes in phylogenetics has repeatedly been shown to substantially increase resolution and provide better topologies and divergence time estimates than shorter mitochondrial DNA (mtDNA) sequences such as the control region (Mueller 2006; Chan *et al.* 2010; Duchene *et al.* 2011; Knaus *et al.* 2011). Recent publications include the application of mitogenomics to determine deep (e.g. Jackson *et al.* 2009; Singh *et al.* 2009; Inoue *et al.* 2010; Lavoue *et al.* 2011; Pacheco *et al.* 2011; Yamanoue *et al.* 2011) as well as more recent radiations (Chan *et al.* 2010; Morin *et al.* 2010; Vilstrup *et al.* 2011; Wielstra & Arntzen 2011), and the use of larger numbers (tens to hundreds) of mitogenomes to understand inter- and intra-specific diversity and evolutionary patterns (e.g. Chan *et al.* 2010; Foote *et al.* 2010; Morin *et al.* 2010; Yamanoue *et al.* 2011). Not only are these techniques being used for contemporary DNA but also ancient DNA studies are increasingly employing mitogenomics for better resolution of phylogeny and phylogeography (e.g. Gilbert *et al.* 2008; Enk *et al.* 2011; Horn *et al.* 2011). The major limitation to using complete mitogenomic data for phylogeographic and phylogenetic studies has been the cost and effort required to generate the data. To date, the majority of these studies have used traditional, labour intensive PCR and Sanger sequencing methods (e.g. Knaus *et al.* 2011; Lavoue *et al.* 2011; Pacheco *et al.* 2011; Wielstra & Arntzen 2011; Yamanoue *et al.* 2011; Yu *et al.* 2011; Shamblin *et al.* 2012).

To avoid the need for individual PCR amplifications of many loci per sample, capture hybridization methods have been developed to enrich genomic DNA samples for preselected genes or DNA fragments (Bashiardes *et al.* 2005; Noonan *et al.* 2006; Hodges *et al.* 2009; Mamanova *et al.* 2010). These methods have typically been applied to enrich genomic samples from single individuals, but recent developments have led to enriching pooled DNA libraries for greater cost-effectiveness (Faircloth *et al.* 2012; Lemmon *et al.* 2012; Peterson *et al.* 2012). When gene sequences from target organisms or related species are known, DNA capture arrays can be commercially synthesized for genomic DNA library enrichment (e.g. SureSelect; Agilent Technologies Inc.), custom-made from PCR products or synthesized RNA probes attached to magnetic beads (Gnirke *et al.* 2009; Maricic *et al.* 2010). We have combined methods for highly multiplexed genomic library preparation (Meyer & Kircher 2010) with methods for capture array enrichment (Hodges *et al.* 2009) that enable efficient and low-cost per sample sequencing of complete mitochondrial genomes and low coverage of up to 118 nuclear loci for up to 100 samples for SNP discovery in a single NGS sequencing lane. To date, we have obtained sequences from the fin whale (*Balaenoptera physalus*), sperm whale (*Physeter macrocephalus*), three species of beaked whales (*Ziphius cavirostris*, *Mesoplodon densirostris*, *M. europaeus*), pantropical spotted dolphin (*Stenella attenuata*), spinner dolphin (*Stenella longirostris*) and the green sea turtle (*Chelonia mydas*). We present a summary of results from over 500 mitogenomes, and discuss optimal sample characteristics, methods and capture array design for obtaining consistent depth of coverage of mitogenome and nDNA loci. Finally, we investigate the use of cross-species capture to expand the number and variety of sequences that can be enriched from genomic DNA of species for which we have little or no genomic information.

Methods

Sample selection and DNA extraction

With few exceptions, genomic DNA (gDNA) extractions were conducted using silica-based filter purification (Qiaxtractor® DX reagents; Qiagen) following manufacturers' instructions, performed on a JANUS® automated work station (Perkin–Elmer). During the initial digestion of the sample, additional proteinase K and overnight digestion at 37 °C was sometimes added to assist with breakdown of cetacean and turtle skin samples. Samples extracted using other methods such as sodium chloride protein precipitation (Miller *et al.* 1988) or phenol/chloroform procedures (Sambrook *et al.* 1989) were put through a secondary silica-based purification (Qiaquick; Qiagen) to remove enzymatic inhibitors (e.g. melanin; Yoshii *et al.* 1993) that are common to cetacean and sea turtle skin sample DNA extracts.

Capture array design

We designed five different capture arrays (Table 1). MtDNA genomes used in each array design were obtained from GenBank® (NCBI/NIH; accession numbers NC_001321, NC_002503, NC_000886, NC_005273, EU_557096). Nuclear loci sequences for the cetacean arrays differed for each species, and were obtained by sequencing conserved mammalian loci (Aitken *et al.* 2004) from the fin whale, sperm whale (Morin *et al.* 2007a; Mesnick *et al.* 2011), spotted and spinner dolphin, and three species of beaked whale samples. For the spotted dolphin, spinner dolphin and beaked whale nuclear loci, we also used the BLASTN tools with default configuration for 'near exact matches' in the Ensembl genome annotation project (Flicek *et al.* 2011) to extract orthologs of 75 sequences that included these sections from the bottlenose dolphin (*Tursiops truncatus*) draft genome (assembly turTru1, Jul 2008; database version 69.1). Orthologs were identified based on high sequence similarity and identification of the gene in the Ensemble genome browser display. This allowed us to expand our original

Table 1 Array design characteristics. Array design data in Dryad data repository (doi:10.5061/dryad.cv35b)

Species	Sequencing technology‡	No. of arrays§	mtDNA replicates¶	nDNA replicates¶	mtDNA probe interval (bp)††
Fin whale	GAII	3	5	45	3
Sperm whale	GAII	2	5	29	3
Green Sea turtle	GAII	1	1	12.5**	7
Spotted and Spinner dolphin*	HiSeq2000	2	1	13	15
Beaked whale†	HiSeq2000	1	1	13	15

*One of the arrays included only spinner dolphins.
†Three species of beaked whales were pooled in one library.
‡Type of Illumina sequencing technology used.
§The physical number of capture arrays that were used.
¶Number of copies of the mitogenome or nuclear loci probe sets that were put on the array.
**Average of the replications times the number of loci in each subset and divided by the total number of loci.. Three subsets of the loci were replicated 16 times (70 loci), 10 times (17 loci) and 6 times (31 loci) on the array respectively.
††The number of base pairs between the beginning of each probe tiled across the mitogenome.

sequences to approximately 1000 bp, typically including portions of at least one exon and one intron (Supplemental Table S1). The green turtle nuclear loci were obtained from sequences generated from sequencing known microsatellite flanking regions and amplified fragment length polymorphism (AFLP) fragments (Roden *et al.* 2009 and unpublished data) (Supplementary Table S2).

The actual array design characteristics for each library followed closely with the design outlined in Hodges *et al.* (2009) with a few minor modifications. All arrays were SureSelect DNA capture arrays with 244 K 60 bp probes on a 1 inch by 3 inch glass slide (Agilent Technologies Inc.). The capture arrays varied in the numbers of individual nuclear loci, numbers of copies of each mtDNA and nDNA probe sequence, and distance between each probe within the mtDNA genome (Table 1). Arrays were designed using the Agilent eArray software (https://earray.chem.agilent.com/earray/; design data files available in the Dryad data repository, doi:10.5061/dryad.cv35b).

Library preparation and array capture

The DNA library preparation and array capture protocols were used as described in Meyer & Kircher (2010) and Hodges *et al.* (2009), respectively, with minor modifications (see supplementary material for complete protocol). Although Meyer & Kircher (2010) state that very low amounts of DNA (as low as 100 pg) can be used in the library preparation with positive results downstream, we used a target amount of 100 ng per sample to ensure good coverage of both the entire mitogenome and a set of nuclear loci. We used the Quant-iT ™ PicoGreen® dsDNA Assay Kit (Invitrogen) in conjunction with a fluorospectrometer to ensure accurate measurement of double stranded DNA in extracts from various methods that could produce potentially spurious spectrophotometer readings caused by RNA or chemical and tissue contaminants.

The first part of the library preparation (Meyer & Kircher 2010) involved shearing the DNA into fragments that are 200–500 bp long with a mean of 250 bp. In contrast to Meyer & Kircher (2010), 80 uL (rather than 50 μL) of total volume was used in the shearing (sonication) procedure, using a Bioruptor UCD-200 (Diagenode). Twenty microlitres of the sheared DNA was electrophoresed in a 2% agarose gel to confirm the presence of fragments of the desired size range. If the fragments were longer than approximately 400–500 bp, one to two more sonication cycles were performed. Blunt-ends were repaired with 20 μL of DNA (instead of 50 μL) and 20 μL of reagent mix, with volumes adjusted to maintain reagent and enzyme concentrations. Adaptor fill-in incubation was extended to 60 min. instead of 30 min. Purifications via Solid Phase Reversible Immobilization (SPRI) were performed as described in Meyer & Kircher (2010); however, a Vortemp® (Labnet) at 55 °C for 10–15 min (no shaking) was used to dry the tubes after the second ethanol wash step.

Prior to starting the indexing PCR, quantification of the adaptor-ligated fragments was performed to estimate the approximate quantity of the target DNA in the library preparation solution. An indexing PCR was performed on the positive control to create standards for the qPCR step as described in 21.ii of the Meyer & Kircher (2010) protocol. Ten-fold serial dilutions of the quantified product were made using AE buffer (Qiagen). Quantitative PCR was performed as described in Meyer & Kircher (2010). The quantitative PCR results were used to determine, based on DNA quantity, how many PCR cycles should be used in the indexing PCR (see below).

The final steps prior to pooling involved indexing each of the samples by PCR amplification using indexed primers, purification, quantifying the indexed product and pooling them into a single tube for library capture on the array. Each sample was indexed with a different reverse primer in the indexing PCR, using 10 μL of the adaptor-ligated library. For most of the libraries, with the exception of the fin whales and one sperm whale library, we used more PCR cycles than recommended by Meyer & Kircher (2010): if the template contained ≥10 ng/μL, 25 cycles were used in the PCR; if the concentration was <10 ng/μL, 30 cycles were used. Products were electrophoresed on a 2% agarose gel; if any of the lanes containing sample appeared blank, then a subsequent indexing PCR was completed for that sample using 30 cycles (products were combined if the second PCR appeared to still be weak or blank in the gel). While more cycles than recommended were used in four library preparations, we suggest starting with 15 cycles and increasing the number of cycles if no band is visible on the agarose gel to avoid possible overcycling, leading to biased replication of fragments at this PCR step (see discussion). After the final SPRI purification, we quantified the amplified products with a Nanodrop® (Thermoscientific) as recommended by Hodges *et al.* (2009) and then pooled the DNA in equimolar quantities for a total combined quantity of 20 μg of DNA (target = 20 μg of DNA total, minimum amount acquired = 15 μg) in a final volume of 138 μL. If the pooled volume exceeded 138 μL, the pooled library was concentrated in a Savant SpeedVac® (Thermoscientific).

The capture array hybridization method described by Hodges *et al.* (2009) involves mixing the pooled product with a buffer solution and oligo-blockers (including a blocker of repetitive elements, Human Cot-1 DNA), then washing this entire mixture over the surface of the array at 65 °C for 65 h in a rotating chamber. We use a combination of oligos from both protocols in the hybridization mixture to accommodate the Meyer & Kircher (2010) index primers (Table 2). We also advise rotating the chamber at 20 r.p.m rather than 12 r.p.m for maximum coverage of the array slide, following the recommendations of the array manufacturer (Agilent Technologies, Inc.). After hybridization, the array was washed and the enriched library eluted as described by Hodges *et al.* (2009) and according to manufacturer's instructions (Agilent Technologies, Inc.) (Fig. 1).

Following concentration of the enriched library in a Savant SpeedVac® as described in Hodges *et al.* (2009), the DNA was amplified in five replicate reactions. The master mix recipe for this amplification, thermocycler conditions and number of cycles followed Hodges *et al.* (2009). The five products were pooled and 10 μL of the product was electrophoresed on a 2% agarose gel to confirm amplification. The remaining product was purified using a Qiaquick cleaning kit as detailed in Hodges *et al.* (2009), and eluted in a final volume of 60 μL. Twenty microlitres of the product was gel purified using a 2% low-melt agarose and the Zymoclean™ Gel DNA recovery kit (Zymo Research Corporation) to isolate fragments in the range of approximately 250–300 bp prior to single-end sequencing (each library in a single lane) on an Illumina Genome Analyzer II (read length 80–100 bp) with a cBot single read cluster generation kit or a HiSeq2000 Analyzer (read length 100 bp) with the TruSeqSR cluster kit (Illumina, Inc,) (Table 1). All libraries were sequenced by The DNA Array Core Facility (The Scripps Research Institute, La Jolla, CA).

Table 2 Blocking oligos used in the hybridization mixture

Primer name	Stock concentration (μM)	Source citation
BO 1	200	Hodges *et al.* 2009
BO 3	200	Hodges et al. 2009
B03.P7.part1.F	200	Meyer & Kircher 2010
B04.P7.part1.R	200	Meyer & Kircher 2010
B05.P7.part2.F	200	Meyer & Kircher 2010
B06.P7.part2.R	200	Meyer & Kircher 2010

Assembly of mitogenomes and nuclear sequences

Additional filtering of read quality was generally not performed prior to mitogenome assembly as the high coverage resulted in high-quality consensus sequences. For low-coverage nuclear loci, filtering to remove short reads (e.g. <20 bp) and reads that did not have at least 95% of nucleotides that have quality scores >15 provided a slight reduction in the number of putative single nucleotide polymorphisms (SNPs) that showed up in only a single individual and were likely to be false positives. Results differed by species and library, so filtering and assembly were done iteratively with different parameters until sequence quality stabilized.

Assembly of mitogenomes and nuclear sequences was performed using one of the three methods. Individual assemblies to a reference sequence were conducted in CLC Genomics Workbench v4.1 (CLCbio) or in Geneious Pro (v. 5.5.3) (Biomatters Ltd.). For batches of samples we used custom scripts (Dryad data repository doi:10.5061/dryad.cv35b) in the R computing environment (R Development Core Team 2006) to iteratively run publicly available analysis packages for quality filtering (fastx toolkit; http://hannonlab.cshl.edu/fastx_toolkit/), assembly (bwa; Li & Durbin 2009), multiple alignment (mafft; Katoh *et al.* 2005) and SNP detection (gatk; DePristo *et al.* 2011; Nielsen *et al.* 2011).

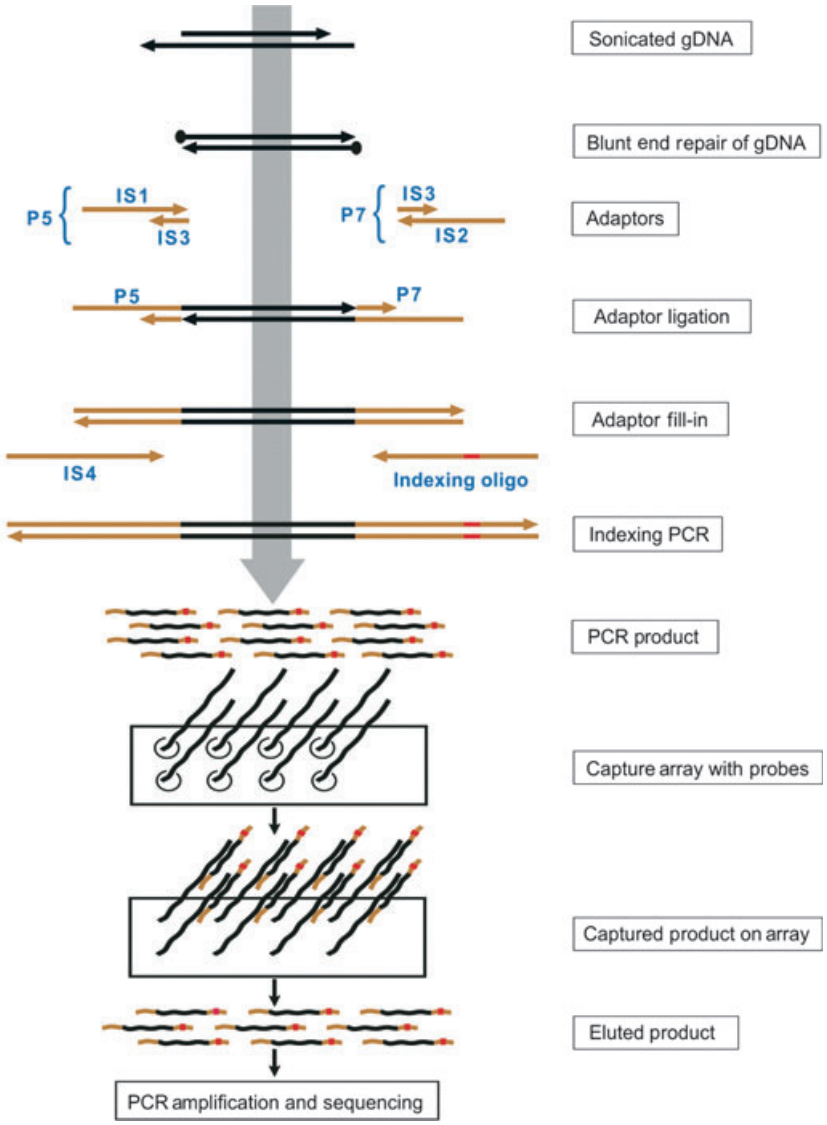

Fig. 1 Schematic of the library preparation and capture array hybridization method

To evaluate the efficiency of cross-species capture and assembly (the leatherback turtle reads were generated by hybridization to an array design based on green turtle sequence), the leatherback turtle reads were first assembled to a reference leatherback turtle mitogenome sequence (Duchene *et al.* 2012) to determine the capture efficiency when green turtle baits were used on the capture array. Assembly parameters in CLC Genomics Workbench were: Similarity = 0.8, Length fraction = 0.8, Insertion cost = 3, Deletion cost = 3, Mismatch cost = 3. To test for the effects of assembly to a divergent reference sequence, the previously assembled reads were then also assembled to the green turtle mitogenome reference sequence (Accession number AB012104) in Geneious Pro using custom settings: Allow gaps: Max per read = 10, word length = 24; ignore words repeated more than one times; Max mismatches per read = 25%; Max gap size = 3, Max ambiguity = 4. Evaluation of the divergence between genomes relative to the depth of coverage in the assembly was done by setting parameters for highlighting low-coverage regions to show regions with depth of coverage <10. All of these regions include gaps in coverage, plus flanking regions where depth of coverage is low so that the consensus sequence is more likely to contain errors. Raw sequence read FASTQ files for marine turtle sequences (Duchene *et al.* 2012) are available on request for readers interested in evaluating distribution and assembly characteristics of the raw reads (supplementary Table S3).

Calculations of percent reads on target were based on the total number of mapped reads for all samples within a species, divided by the total number of reads for all samples. The expected number of reads from an

unenriched library was calculated as the size of the target sequence (16 500 for mtDNA, variable for nuclear loci; Table 5), multiplied by the expected number of copies per cell (one for nuclear, 300 for mtDNA; Morin *et al.* 2007b), and divided by the genome size. Since we do not yet know the genome size for a cetacean or marine turtle, we used the size of the human genome, 3.3 billion, as the estimated size. The enrichment factor was calculated as the total number of mapped reads divided by the expected number of reads from an unenriched library.

Results

A total of 550 individuals across eight species were sequenced and analysed using these NGS techniques, including 153 fin whales, 100 sperm whales, 52 beaked whales, 116 spinner dolphins, 45 spotted dolphins and 84 green turtles (plus 2–4 individuals of each of the six other sea turtle species). An average of 21 μg of indexed DNA (pooled samples) was used for capture enrichment hybridization (N = 3 arrays) for the fin whale, 16.3 μg for the sperm whale (N = 2), 22 μg for the beaked whale (N = 1), 28 μg for the spinner dolphin (N = 2), 11.4 μg for the spotted dolphin (N = 1) and 43 μg were used for the sea turtles (N = 1). For a summary of the mtDNA and nDNA statistics based on species, see Tables 3, 4 and 5.

The percentage of all sequence reads that contained index sequences ranged from 91% to 97% per library, with an average of 94% (HiSeq2000 data, N = 4 libraries). Capture array enrichment for mitochondrial and nuclear loci varied substantially between array designs and sample sets. The portion of sequence reads that mapped to the targeted reference sequences (mtDNA and nDNA) ranged from approximately 3% to 55%, of which 0.2% to 1.7% was for nuclear loci, and all remaining mapped reads were for mitochondrial DNA (percent reads on target, Tables 4 and 5). However, the proportion of unique mapped reads to total mapped reads (identical reads included) ranged from 0.7% to 39%, indicating wide variation in the level of clonality among reads due to PCR amplification. We calculated the number of reads that would be expected to map to our mtDNA and nuclear references if there were no enrichment from the capture array hybridization, and found that mtDNA was enriched on average by 125-fold (range 12–367), and nuclear DNA was enriched on average 565-fold (range 362–920). The lower enrichment factor for mtDNA is expected given that mtDNA is already typically represented by about 300 copies per cell in preserved cetacean biopsy samples (Morin *et al.* 2007b) and could represent a much higher portion of the cellular DNA from degraded samples.

Mean depth of coverage of mitogenomes was generally high, ranging from 22 to 195 reads per nucleotide

Table 3 Summary statistics for libraries and raw sequence data

Raw data

Species	PCR product amt (ng)*	Number of unique samples across all arrays	Number of runs†	Total number of reads (millions)	Mean/Median of total number of reads per individual	Min/Max of total number of reads per individual
Fin whale	378.3	153‡	166	74.80	450 000	6 041
					273 855	4 630 000
Sperm whale	324.5	100‡	103	70.75	686 852	3 704
					539 042	2 630 762
Green turtle	500	84§	84	26.8	318 634	6 995
					304 271	938 778
Spinner dolphin	244.5	116‡	116	112.8	972 670	289 447
					855 815	2 756 921
Spotted dolphin	253	45‡	45	58.00	1 300 000	383 286
					1 140 000	2 750 000
Beaked whales	376.9	52‡	53	89.1	1 680 947	556 289
					1 552 152	4 000 000

*Amount of total PCR product that was pooled on average per individual.

†In both the fin and sperm whale arrays some samples were duplicated for replication purposes. One sample in the beaked whale array had two sets of data.

‡Total number of samples from all hybridization arrays of that species, sequenced separately in one to three different lanes on an Illumina GA-II or HiSeq2000. 50–86 samples were pooled on each capture array.

§Eighty four green turtle samples plus 16 samples of other turtle species (total = 100 samples). Only green turtle data are considered in the text and this table.

Table 4 Summary statistics of mtDNA assemblies

Species (No. arrays)	Number of unique samples across all arrays	Total number of mapped reads (millions)	Number of unique reads*	Percent reads on target†	Mean/Median of reads per individual‡	Min/Max of reads per individual‡	SD of reads per individual§	Mean coverage¶	Number of individuals with mean depth of coverage >20
Fin whale (3)	153	41.30	3429 800	55.1	22 416 23 641	2055 32 548	7534	124.1	151
Sperm whale (2)	100	19.00	2446 301	26.8	24 463 27 750	685 32 007	8584	147.8	94
Green turtle (1)	84	4.6	736 538	17.1	8768 6631	43 31 305	7397	52.7	65
Spinner dolphin (2)	116	2.4	959 723	2.1	8274 7903	884 31 554	5000	50.42	95
Spotted dolphin (1)	45	1.02	159 165	1.8	3548 2411	371 22 632	3936	21.6	12
Beaked whales (1)	52	8.1	481 135	9.1	9252 5981	1445 29 107	7848	56.5	44

*Identical reads have been filtered out (from total number of mapped reads) to leave just unique reads mapped to the reference sequence.
†the ratio of total mapped reads to total number of reads, as a percent.
‡'reads per individual' refers to the total number of reads divided by the number of individuals.
§Standard deviation of reads per individual across all arrays.
¶Mean depth of coverage of unique reads per nucleotide site in the assembled sequence.

Table 5 Summary statistics of 790 nDNA assemblies

Species (No. arrays)	Number of unique samples across all arrays	Total number of mapped reads	Number of unique reads	Total sequence length*	Total number of loci†	Percent reads on target‡	Mean/Median reads per individual§	Min/Max reads per individual§	SD of reads per individual¶	Mean coverage**
Fin whale (3)	153	17 2252	74 601	17 942	43	0.2	488 455	21 1425	331	2.1
Sperm whale (2)	100	25 2131	86 604	26 297	49	0.4	866 771	1 3810	722	3.3
Green turtle (1)	84	36 8989	97 740	60 434	118	1.4	1163 1191	5 3334	708	2
Spinner dolphin (2)	116	1 545 590	590 573	87 269	85	1.4	5091 5318	360 15 034	3718	5.8
Spotted dolphin (1)	45	595 201	31 570	87 269	85	1	701 611	179 2682	439	1
Beaked whales (1)	52	1 481 587	179 532	59 667	75	1.7	3452 3276	98 10 339	2360	5.8

*The sum of individual sequence lengths.
†The total number of loci that were used in the array design.
‡See Table 4.
§See Table 4.
¶See Table 4.
**Mean depth of coverage (unique reads/nucleotide site).

position across arrays, and resulted in full coverage of the complete mitogenome sequence in all but a few individuals from most libraries (except for spotted dolphins and cross-species capture; see below) (Table 4). We found that samples with mean depth of coverage greater than 20 have an average of 13 zero-coverage nucleotides in the mitogenome assemblies, resulting in the majority of samples having complete or nearly complete mitogenome sequences (Table 4). Even samples with average depth of coverage between 10 and 20 typically had fewer than 25 zero-coverage nucleotides, and average depth of coverage as low as seven resulted in fewer than 100 zero-coverage bases in several samples (data not shown). We attributed low coverage of some samples to a combination of DNA degradation, over estimates of the DNA concentration prior to library preparation, and use of <100 ng of DNA for library preparation when sufficient DNA was not available.

Mean nuclear locus depth of coverage was consistently low, ranging from 1 to 5.8 reads per nucleotide position (Table 5). This low level of coverage was not uniformly distributed within and among loci (Supplementary Table S6), so that there were some loci or parts of the targeted sequences that had consistently higher depth of coverage while others had low or no coverage across all individuals.

We used the GATK software (McKenna *et al.* 2010; Nielsen *et al.* 2011) to identify SNPs based on multi-sample analysis followed by genotype validation in individual samples. SNP validation varied slightly by species, but in general we considered variable sites identified by GATK in nuclear loci to be potential SNPs if there were heterozygotes and/or homozygotes of the alternate allele, depth of coverage was at least 5–7 reads per individual for called genotypes, and the minor allele frequency was ≥ 0.05 in the sample set. Individual SNPs were validated based on genotypes across all samples with adequate coverage. We excluded apparent SNPs as possible paralogs if they were heterozygous in most or all samples or were found on reads with several additional variants relative to the reference sequence. A total of 88 318 bp (86 loci) were screened for SNP discovery in spinner dolphins (*S. longirostris*), and 59 667 bp (75 loci) of nuclear sequence were screened to discover SNPs for two species of beaked whales (Table 5). We identified 132 putative SNPs from the spinner dolphin sample set, representing approximately one SNP per 670 bp of sequence, 117 potential SNPs for Cuvier's beaked whale (*Z. cavirostris*) and 188 for Blainsville's beaked whale (*M. densirostris*), representing approximately one SNP per approximately 300–500 bp of sequence.

We compared the results for 13 loci that were previously sequenced from the Cuvier's beaked whale (*Z. cavirostris*) using Sanger sequencing of PCR products from 20 individuals (high-quality sequence from 8 to 18 individuals per locus) with the same subset of 13 loci obtained from our capture enrichment and NGS results from 22 individuals (high-quality sequence from 15 to 21 individuals). Using the conservative criteria described above to exclude low depth of coverage or rare SNPs, we identified 18 putative SNPs common to both data sets. Fourteen additional SNPs were from Sanger data only, and 34 from NGS only. When individual NGS sequences were inspected for SNPs found by Sanger sequencing, additional SNPs were confirmed that were not initially identified, typically because they were only found in one individual and had low depth of coverage.

Of 71 *Z. cavirostris* SNPs selected for SNP assay design for Amplifuor genotyping (Morin & McCarthy 2007), 59 assays were successfully designed and 53 were optimized for genotyping in the lab, which resulted in resolution of variable genotypes for 50 of the SNPs. Thus, 72% of our selected SNPs from the NGS analysis resulted in SNP assays that could be genotyped, and 96% of the genotyped SNPs were polymorphic (unpublished data).

A portion of our samples had low quality DNA (e.g. beach-stranded cetaceans). Although results were variable for these samples, they tended to favour mtDNA, resulting in an overabundance of fragments mapped to the mtDNA genome relative to the nDNA target fragments. In a single pool of fin whale samples that were hybridized to a capture array prior to sequencing, four of the nine samples that were from beach-stranded whales had the highest mean depth of coverage of the mitogenome across all samples in the pool, and the nine stranded animal samples averaged about five times the mean depth of coverage of 40 skin biopsy samples from live animals (mean depth of coverage = 3856 and 775 respectively) in the entire pool.

We examined cross-species capture for the most divergent species of marine turtles, the green turtle (used as the capture array bait) and leatherback turtle (Dutton *et al.* 1996). For example, the leatherback sample with the highest number of reads (216 893), had 52 963 reads that assembled to the leatherback reference mitogenome, resulting in full coverage of all parts of the mitogenome except for the control region. All leatherback turtle reads assembling to the reference were then aligned to the green turtle mitogenome reference. This resulted in 26 regions with gaps and flanking regions of <10 reads, excluding the repeat region in the control region. Mean similarity between the aligned leatherback and green turtle reference mitogenomes was 83.8% (SD = 6.7) in the gaps/low-coverage regions, but 89.4% for the remaining mitogenome regions with ≥ 10 reads. The similarity in gap regions ranged from 65.6% to 92.0%, and 75.0% of the gaps occur at similarity <88.0%. The G/C content

was slightly lower in the gaps/low-coverage regions relative to the higher depth of coverage regions (39.0% vs. 39.4%). Assembly and phylogenetic analysis of the mitogenomes from other marine turtle species is described elsewhere (Duchene *et al.* 2012).

Discussion

We have demonstrated that the capture array and next-generation sequencing techniques adopted here were successful in generating whole mitochondrial genomes and numerous nuclear sequences from large numbers of samples across seven cetacean species and green turtles. Over 500 mitogenomes have been generated and with few exceptions, 100% of the mitochondrial reference sequence was covered for each of the sequenced individuals. While the capture array enrichment varied substantially among arrays, up to 55% of the sequence reads mapped to reference sequences, resulting in mean depth of coverage of 22–148 reads for mitogenomes. We also demonstrated library enrichment for targeted regions from divergent sequences of nontarget species. As described in Duchene *et al.* (2012), using the green turtle as capture array bait, we were able to sequence complete mitogenomes (except the control region) for all the other sea turtle species, resulting in reference mitochondrial genomes for several species that were previously unavailable. Finally, we were able to generate sequences for a larger number of nuclear loci than with traditional sequencing, for the purpose of SNP discovery. With further refinement (see below), there is the potential to use these methods for direct genotyping of SNPs from the NGS data.

DNA samples of low quality can be used to generate mitogenomes using this method, whereas methods based on long range PCR are not successful with these types of samples. While we believe it is possible to use historical or ancient samples, these samples tend to be of low DNA concentration. Therefore, larger gaps may be present in the coverage of either the mitochondrial genome or the target nuclear fragments. As mentioned in the methods section, it is possible to increase the number of indexing PCR cycles for samples that have low DNA quantity, but that may primarily increase coverage depth of the common fragments and not fill in gaps. In fact, we found that using 25 cycles as the low number of cycles in the indexing PCR may have led to the increased clonality (presence of identical sequences, presumably replicated from the same starting molecule) of common fragments for many samples, not just the samples of low quality (data not shown). We recommend starting with a lower number of cycles in the indexing PCR as indicated by Meyer & Kircher (2010) and Mamanova *et al.* (2010). A qPCR assay to estimate the genome copy number can also be used to optimize the number of cycles (see Meyer & Kircher 2010), but we found that it was more efficient to use the same number of cycles on all samples, then follow up with more cycles only for those showing poor amplification. In addition, use of different PCR enzyme and buffer systems than used here may improve the representation of target DNA fragments in sequencing libraries, and are recommended to reduce biases in fragment size and GC-content (Dabney & Meyer 2012). Additional measures, such as using T-A ligation (instead of blunt-end ligation) and double indexing, can be used to improve ligation efficiency and sequencing accuracy, which is particularly important for ancient DNA and other poor quality samples (Mamanova *et al.* 2010; Kircher *et al.* 2012). In previous studies, using blunt-end instead of T-A ligation has been found to cause chimeras (self-ligation of both target DNA and adapters) to appear during the library preparation. Although in practice we see little evidence of this happening in our libraries, it is likely that some chimeras form and could be sequenced if the DNA is over-sonicated, resulting in fragments shorter than the NGS read length (Lodes 2012).

As we have continued to develop and enhance our protocols for creating next-generation sequencing data, three notable improvements have emerged based on our results. First, although the manufacturers of the capture arrays (Agilent Technologies Inc.) and Hodges *et al.* (2009) recommend using a total of 20 μg of DNA product when hybridizing to the array, we have used additional DNA product to expand the number of samples and loci sequenced. For the green turtle array, we used more than double the amount of recommended index PCR product and increased sample size from 50 to 100. For this library (relative to fin and sperm whales with N = 50 samples per array, also sequenced on the Illumina GAII platform), the mean depth of coverage for mtDNA decreased (possibly due to multiple factors; see below), but remained above 50, and nuclear locus coverage was not noticeably different (Tables 4 and 5). Second, both the fin whale and sperm whale arrays were designed with five copies of the mitochondrial genome probes attached to them, while the beaked whale, spinner dolphin, spotted dolphin and turtle arrays had only one copy. The mtDNA mean depth of coverage declined when we reduced the number of copies on the array to 1, but averaged over 50 for all arrays that had 1 copy of the mtDNA probe set (except spotted dolphins; see below). Finally, the mtDNA probes were spaced every 3, 7 or 15 bp apart depending on the species (Table 1). Spacing the mtDNA probes up to 15 bp apart (4X coverage of the mitogenome sequence) did not reduce mtDNA depth of coverage (e.g. the green turtle array had a probe interval of 7 bp, and a lower mean depth of coverage than the beaked whale array with a probe interval of 15 bp; Table 4).

We saw an increase in the proportion of reads mapped to nuclear loci from the fin whale and sperm whale libraries (0.2% and 0.4% respectively) to the later libraries ($\geq$1%). This improvement in the proportion of mapped nDNA reads is possibly due to two factors: 1) all the arrays subsequent to the fin whale and sperm whale arrays had fewer copies of the mtDNA sequence probes on them, and a larger mtDNA probe interval, resulting in less competition between mtDNA and nDNA for capture and sequencing capacity. Second, the arrays described above were sequenced using more advanced technology (from GAII to HiSeq 2000), which produced longer reads on average and an increased number of reads per array (Table 1).

The greatest limitation of combined mtDNA and nuclear capture on a single array appears to be the relatively low depth of coverage of the nuclear loci. This depth of coverage limits the use of nuclear sequence data to SNP discovery (the identification of variable nucleotide sites in the sample set) in most cases, rather than genotyping of all individuals directly from the sequence data (e.g. Elshire *et al.* 2011). On some future capture arrays, we will experiment with separating the mitochondrial genome from the nuclear loci when greater nDNA depth of coverage is required. With these arrays, each one will have either mtDNA reference sequence capture baits or nuclear locus capture baits only. Sequencing could be done either in two separate lanes to maximize the number of nuclear reads, or after hybridization and elution, pooling unequal portions of the enriched mtDNA and nuclear libraries to be sequenced on one lane to reduce cost, although the ratio would need to be empirically tested. Either approach should have the advantage of providing both high depth of coverage for mitogenomes and nDNA loci for the majority of samples (cf. Mamanova *et al.* 2010), and could allow significant expansion of the number of nuclear loci sequenced per individual.

Our goals for individual species projects have varied, resulting in different array designs aimed at generating primarily mitochondrial sequences, or a combination of mitogenomes and nuclear sequences for SNP discovery. We have also attempted to adjust the portions of reads that map to mtDNA vs. nDNA so that we can get closer to the optimal depth of coverage for both from a single capture array. As shown in Table 5, we were able to increase the number and average length of nuclear loci, and also increase the number of pooled samples, without losing coverage density of nuclear loci. Our ability to assemble complete mitogenomes from most samples was also not impaired by the increase in pooled individuals (the spinner and spotted dolphin and green turtle arrays). With the exception of spotted dolphins, mean depth of coverage of mitogenomes remained above 50 reads/nucleotide site as we increased the mitogenome probe interval from 3 bp to 15 bp, and the number of copies of the mitogenome probe sets were reduced from five to one. Presumably, these changes would have resulted in concomitant increases in the mean depth of coverage of the nuclear loci if we had held the nuclear locus number and mean length constant, but we chose to use the potential extra read capacity to increase the number of loci, mean length and number of samples pooled on the array. Despite the above increases, we still saw the mean depth of coverage increase from 2.1 to 3.3 for fin and sperm whale arrays, where the total nuclear sequence length was 18 kbp and 26 kbp, respectively, to 5.8 for beaked whales and spinner dolphins, where the total nuclear sequence length was approximately 60 kbp and 87 kbp respectively. The spotted dolphin and turtle arrays, however, each had lower nuclear depth of coverage. We believe that the spotted dolphin array suffered from poor capture overall, most likely due to some problem in the library preparation or hybridization, and the turtle array had many very short nuclear sequences, some of which included tandem repeats, which caused highly variable and overall low depth of coverage across the nuclear loci. In the time since we conducted our initial experiments, the number of reads per sequencing lane has increased almost six-fold, so even without additional changes to our array designs, we anticipate significantly better depth of coverage for both nuclear and mitochondrial loci from a single capture array.

Any capture or genomic reduction (e.g. RAD-TAG; Baird *et al.* 2008) method has the potential to also generate sequences of paralogous loci or nuclear pseudogenes of mitochondrial loci (NuMts; Bensasson *et al.* 2001; Lopez *et al.* 1994). We have completed phylogenetic analysis of marine turtle and beaked whale data sets, including investigation of divergent reads in mitochondrial and nuclear locus assemblies (Duchene *et al.* 2012; Morin *et al.* 2012b), and have also completed SNP discovery for three species (see below). It is clear that when assembly stringency is low, there can be multiple divergent sequences in some regions of the mitogenome assembly, presumably representing NuMts that were captured with the true mitochondrial sequences. As the mitochondrial DNA is typically in 100 to 1000-fold excess relative to nuclear copies in the genome (Morin *et al.* 2007b) and differential amplification of NuMt and true mitochondrial fragments is not expected in NGS libraries due to similar base composition (Mamanova *et al.* 2010), we would expect that NuMts would be a small fraction of reads and might only cause problems in assembly when stringency is low and/or depth of coverage is low. To date, we have not found evidence of divergent reads at high frequencies in mitogenome assemblies, resulting in consistent and repeatable consensus sequences even when presumptive NuMt

sequence reads have been observed in the assemblies. Especially in areas of low coverage, however, it is important to visually inspect assemblies to confirm that divergent NuMts are not present at high enough frequency to alter the consensus sequence.

We have only performed analysis of nuclear loci for SNP discovery (detecting the location of an individual SNP) from three species to date, primarily because software to do SNP discovery for large numbers of sequences and samples is not readily available, and the process remains labour intensive from this type of data (see, e.g. Nielsen *et al*. 2011; Nekrutenko & Taylor 2012). We have implemented a custom SNP discovery pipeline that uses the GATK software for SNP discovery, followed by generation of a database of SNP genotypes and read counts per genotype based on a set of rules that take into account the depth of coverage and minimum and maximum percent of alternate alleles for calling homozygotes and heterozygotes. The putative SNPs were inspected to identify those that appear to segregate in approximately Hardy–Weinberg proportions and have little or no evidence of paralogous loci (e.g. homozygotes consisted of only one allele in the assembled reads). Results for beaked whales indicate that with approximately 20 samples per species we are able to detect over 100 SNPs in the 75 nuclear loci.

Both Sanger and NGS SNP discovery methods were limited by high variability in the samples sequenced for each locus or position in the locus. This problem is partially overcome by sequencing larger numbers of samples, so that the methods meant to detect SNPs from larger numbers of low-coverage samples can detect SNPs from the combined data set (e.g. GATK, reviewed in Nielsen *et al*. 2011). The overall low coverage and variation in coverage depth within and among loci means that we will potentially miss detecting SNPs with both methods, but the effect of that on our ability to detect a sufficient number of SNPs will depend on the goals of the project. If discovery of large numbers of SNPs (>100) is the primary goal, then we suggest using methods recently described for genome-wide nuclear locus enrichment and sequencing (Faircloth *et al*. 2012; Lemmon *et al*. 2012), or genome-wide SNP discovery methods such as RAD sequencing (Miller *et al*. 2007; Baird *et al*. 2008; Rowe *et al*. 2011; Peterson *et al*. 2012). However, the method described here is time and resource efficient for the combined sequencing of mitogenomes and discovery of a number of nuclear SNPs adequate for studies of population structure and phylogeography (e.g. Morin *et al*. 2009, 2012a; Finger *et al*. 2011; Kogura *et al*. 2011; Mesnick *et al*. 2011) or for selection of informative SNPs for mixed-stock analysis or species identification (e.g. Smith & Seeb 2008; Bowden *et al*. 2012). Regardless of the goal, the labour involved in SNP discovery is significantly less using NGS methods than traditional PCR and Sanger sequencing (both in the laboratory and in data analysis) and has the potential to become significantly more automated for NGS methods (e.g. Nielsen *et al*. 2011 and references therin).

In considering whether to use NGS technology to sequence entire mitogenomes, a cost estimate analysis between Sanger sequencing technology and NGS technology is warranted. To sequence the entire mitochondrial genome of a cetacean, ≥ 25 partially overlapping fragments at nearly 700 bp each would have to be PCR amplified and sequenced (cetacean mitochondrial genomes are typically about 16 400 bp long). For 50 samples sequenced using 25 primer pairs, we estimate that the cost would be approximately $9900 (supplies only for PCR, product cleaning and sequencing). Sequencing the same number of samples using the NGS technology, we have outlined here costs approximately $3400 ($960 for library preparation, $630 for capture array hybridization, plus an estimated $1750 per lane for sequencing on the Illumina platform). The initial cost of materials (e.g. 100 index primers) and equipment for preparing NGS libraries is significant, but can be spread over many subsequent projects. It should be noted that the methods for enriching large numbers of nuclear sequences for phylogenetic analysis or genome-wide SNP discovery, while more efficient for generating large numbers of loci per sample, require a more substantial initial investment in the more costly in-solution enrichment libraries (e.g. approximately $1700 to $15 000 for MYcroarray and Agilent SureSelect kits, respectively; Faircloth *et al*. 2012; Lemmon *et al*. 2012). These methods are likely to be more cost-effective for larger numbers of samples or nuclear loci within a project, or spread across projects when the same capture sequences can be used (e.g. cross-species capture), and have the added benefit of requiring less specialized laboratory equipment than hybridization to solid capture arrays.

If traditional Sanger sequencing was used to generate mitogenomes as well as 80 nuclear loci for 50–100 samples, between 10 000 and 20 000 sequences would need to be generated to match what is obtained in one NGS lane. Furthermore, Sanger sequencing would require running two 96-well plates every day for a year to generate these data considering that it takes approximately 3–4 h to do a PCR or sequencing run. Although next-generation sequencing entails considerable laboratory work, it takes less than a month to generate the libraries and they can be sequenced in a matter of days. With further advances in efficiency and automation, this time and cost will undoubtedly decrease (e.g. Fisher *et al*. 2011; Rohland & Reich 2012).

The impact of these technologies on phylogenetics, phylogeography and population genetics is already

becoming apparent. Whole mitogenomes were previously relegated to use in analysis of deep evolutionary relationships, or where needed to resolve tree topologies that could not be resolved with shorter sequences. With NGS technologies, they are increasingly being applied to large numbers of samples both above and below the species level, resulting in highly supported phylogenies and more precise and accurate estimates of divergence times, and even allowing the identification of genes under selection. Generating whole mitogenome sequences or large numbers of nuclear loci was previously limited by the need to PCR amplify and sequence many shorter fragments (<1000 bp). With capture arrays, cross-species capture and highly multiplexed library enrichment and sequencing, we can efficiently and cost-effectively apply mitogenomics to hundreds of samples of nonmodel species and perform SNP discovery on dozens to over 100 loci, all from a single capture array design and one or a few Illumina sequencing lanes. Continued improvements in sequencing technology will allow scaling up of the number and length of reads, resulting in higher depth of coverage without any substantial changes in protocols or resources. The potential for doing nuclear SNP genotyping directly from NGS data is already being realized for a small number of model organisms, and with the methods presented here, should be possible in the near future for many more nonmodel organisms.

Acknowledgements

We are grateful to Matthias Meyer for pre-publication access to manuscripts and methods, and helpful discussion in adapting methods for cetacean and sea turtle tissues. Steve Head and the TSRI DNA Array Core Facility were extremely helpful not only in sequencing our libraries but in giving advice about the library preparation procedure. Funding came from the NMFS Southwest Region and PRD/SWFSC. Thank you to Aimee Lang, Barb Taylor, Bill Perrin and three anonymous reviewers for helpful comments.

References

Aitken N, Smith S, Schwarz C, Morin PA (2004) Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. *Molecular Ecology*, **13**, 1423–1431.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Bashiardes S, Veile R, Helms C *et al.* (2005) Direct genomic selection. *Nature Methods*, **2**, 63–69.

Bensasson D, Zhang D, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution*, **16**, 314–321.

Bowden R, MacFie TS, Myers S *et al.* (2012) Genomic tools for evolution and conservation in the chimpanzee: *Pan troglodytes ellioti* is a genetically distinct population. *PLoS Genetics*, **8**, e1002504.

Chan YC, Roos C, Inoue-Murayama M *et al.* (2010) Mitochondrial genome sequences effectively reveal the phylogeny of *Hylobates* gibbons. *PLoS ONE*, **5**, e14419.

Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, **52**, 87–94.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Duchene S, Archer FI, Vilstrup J, Caballero S, Morin PA (2011) Mitogenome phylogenetics: The impact of using single regions and partitioning schemes on topology, substitution rate and divergence time estimation. *PLoS ONE*, **6**, e27138.

Duchene S, Frey A, Alfaro-Núñez A *et al.* (2012) Marine turtle mitogenome phylogenetics and evolution. *Molecular Phylogenetics and Evolution*, **65**, 241–250.

Dutton PH, Davis SK, Guerra T, Owens D (1996) Molecular phylogeny for marine turtles based on sequences of the ND4-leucine tRNA and control regions of mitochondrial DNA. *Molecular Phylogenetics and Evolution*, **5**, 511–521.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Enk J, Devault A, Debruyne R *et al.* (2011) Complete Columbian mammoth mitogenome suggests interbreeding with woolly mammoths. *Genome Biology*, **12**, R51.

Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Finger AJ, Anderson EC, Stephens MR, May BP (2011) Application of a method for estimating effective population size and admixture using diagnostic single nucleotide polymorphisms (SNPs): implications for conservation of threatened Paiute cutthroat trout (Oncorhynchus clarkii seleniris) in Silver King Creek, California. *Canadian Journal of Fisheries and Aquatic Sciences*, **68**, 1369–1386.

Fisher S, Barry A, Abreu J *et al.* (2011) A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology*, **12**, R1.

Flicek P, Amode MR, Barrell D *et al.* (2011) Ensembl 2011. *Nucleic Acids Research*, **39**, D800–D806.

Foote AD, Morin PA, Durban JW *et al.* (2010) Positive selection on the killer whale mitogenome. *Biology Letters*, **7**, 116–118.

Gilbert MTP, Drautz DI, Lesk AM *et al.* (2008) Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. *Proceedings of the National Academy of Sciences*, **105**, 8327.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Hodges E, Rooks M, Xuan Z *et al.* (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols*, **4**, 960–974.

Horn S, Durka W, Wolf R *et al.* (2011) Mitochondrial genomes reveal slow rates of molecular evolution and the timing of speciation in beavers (*Castor*), one of the largest rodent species. *PLoS ONE*, **6**, e14622.

Inoue JG, Miya M, Lam K *et al.* (2010) Evolutionary origin and phylogeny of the modern holocephalans (Chondrichthyes: Chimaeriformes): a mitogenomic perspective. *Molecular Biology and Evolution*, **27**, 2576–2586.

Jackson J, Baker CS, Vant M *et al.* (2009) Big and slow: phylogenetic estimates of molecular evolution in baleen whales (suborder mysticeti). *Molecular Biology and Evolution*, **26**, 13.

Katoh K, Kuma K, Miyata T, Toh H (2005) Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Informatics*, **16**, 22–33.

Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, **40**, e3.

Knaus BJ, Cronn R, Liston A, Pilgrim K, Schwartz MK (2011) Mitochondrial genome sequences illuminate maternal lineages of conservation concern in a rare carnivore. *BMC Ecology*, **11**, 10.

Kogura Y, Seeb JE, Azuma N *et al.* (2011) The genetic population structure of lacustrine sockeye salmon, *Oncorhynchus nerka*, in Japan as the endangered species. *Environmental Biology of Fishes*, **92**, 539–550.

Lavoue S, Miya M, Arnegard ME *et al.* (2011) Remarkable morphological stasis in an extant vertebrate despite tens of millions of years of divergence. *Proceeding of the Royal Society of London, B*, **278**, 1003–1008.

Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.

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

Lodes M (2012) Chimera-free library prep for NGS platforms. *Genetic Engineering & Biotechnology News*, **32**, Available from http://www.genengnews.com/keywordsandtools/print/1/25605/ (accessed 10 April 2012)

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ (1994) Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution*, **39**, 174–190.

Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.

Maricic T, Whitten M, Paabo S (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE*, **5**, e14004.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Mesnick S, Taylor B, Archer EI *et al.* (2011) Sperm whale population structure in the eastern North Pacific inferred by the use of single nucleotide polymorphisms (SNPs), microsatellites and mitochondrial DNA. *Molecular Ecology Resources*, **11**(suppl. 1), 278–298.

Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, doi:10.1101/pdb.prot5448.

Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research*, **16**, 1215.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Morin PA, McCarthy M (2007) Highly accurate SNP genotyping from historical and low-quality samples. *Molecular Ecology Notes*, **7**, 937–946.

Morin PA, Aitken NC, Rubio-Cisneros N, Dizon AE, Mesnick SL (2007a) Characterization of 18 SNP markers for sperm whale (*Physeter macrocephalus*). *Molecular Ecology Notes*, **7**, 626–630.

Morin PA, Hedrick NM, Robertson KM, LeDuc CA (2007b) Comparative mitochondrial and nuclear quantitative PCR of historical marine mammal tissue, bone, baleen, and tooth samples. *Molecular Ecology Notes*, **7**, 404–411.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, **9**, 66–73.

Morin PA, Archer FI, Foote AD *et al.* (2010) Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. *Genome Research*, **20**, 908–916.

Morin PA, Archer FI, Pease VL *et al.* (2012a) An empirical comparison of SNPs and microsatellites for population structure, assignment, and demographic analyses of bowhead whale populations. *Endangered Species Research*, **19**, 129–147.

Morin PA, Duchene S, Lee N, Durban J, Claridge D (2012b) *Preliminary analysis of mitochondrial genome phylogeography of Blainville's, Cuvier's and Gervais' beaked whales*. p. 17, SC/64/SM14, International Whaling Commission, Scientific Meeting 64, Panama City, Panama.

Mueller RL (2006) Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. *Systematic biology*, **55**, 289–300.

Nekrutenko A, Taylor J (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, **13**, 667–672.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Noonan JP, Coop G, Kudaravalli S *et al.* (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science*, **314**, 1113–1118.

Pacheco MA, Battistuzzi FU, Lentino M *et al.* (2011) Evolution of modern birds revealed by mitogenomics: timing the radiation and origin of major orders. *Molecular Biology and Evolution*, **28**, 1927–1942.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

R Development Core Team (2006) *R: a language and environment for statistical computing*. R foundation for Statistical Computing, www.r-project.org

Roden SE, Dutton PH, Morin PA (2009) Characterization of SNP markers for the green sea turtle (*Chelonia mydas*). *Molecular Ecology Resources*, **9**, 1055–1060.

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.

Rowe HC, Renaut S, Guggisberg A (2011) RAD in the realm of next-generation sequencing technologies. *Molecular Ecology*, **20**, 3499–3502.

Sambrook J, Fritsch EF, Maniatis T (1989) *In Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Springs Harbor Press, Cold Springs Harbor, NY.

Shamblin BM, Bjorndal KA, Bolten AB *et al.* (2012) Mitogenomic sequences better resolve stock structure of southern Greater Caribbean green turtle rookeries. *Molecular Ecology*, **21**, 2330–2340.

Singh T, Tsagkogeorga G, Delsuc F *et al.* (2009) Tunicate mitogenomics and phylogenetics: peculiarities of the *Herdmania momus* mitochondrial genome and support for the new chordate phylogeny. *BMC Genomics*, **10**, 534.

Smith CT, Seeb LW (2008) Number of alleles as a predictor of the relative assignment accuracy of STR and SNP baselines for chum salmon. *Transactions of the American Fisheries Society*, **137**, 751–762.

Vilstrup JT, Ho SYW, Foote AD *et al.* (2011) Mitogenomic phylogenetic analyses of the Delphinidae with an emphasis on the Globicephalinae. *BMC Evolutionary Biology*, **11**, 65.

Wielstra B, Arntzen JW (2011) Unraveling the rapid radiation of crested newts (*Triturus cristatus* superspecies) using complete mitogenomic sequences. *BMC Evolutionary Biology*, **11**, 162.

Yamanoue Y, Miya M, Doi H *et al.* (2011) Multiple invasions into freshwater by pufferfishes (teleostei: tetraodontidae): a mitogenomic perspective. *PLoS ONE*, **6**, e17410.

Yoshii T, Tamura K, Taniguchi T, Akiyama K, Ishiyama I (1993) [Water-soluble eumelanin as a PCR-inhibitor and a simple method for its removal]. *Nihon Hoigaku Zasshi*, **47**, 323–329.

Yu L, Wang X, Ting N, Zhang Y (2011) Mitogenomic analysis of Chinese snub-nosed monkeys: evidence of positive selection in NADH dehydrogenase genes in high-altitude adaptation. *Mitochondrion*, **11**, 497–503.

B.L.H. performed the research, analysed data and wrote the manuscript; A.F., and M.S.L., performed the research and analysed data; F.I.A. analysed data and contributed to analysis by developing R-scripts to automate multiple analysis steps using publicly available programs; P.H.D. designed the research; P.A.M. designed the research, analysed data and wrote the manuscript

Data Accessibility

Array Sequences: Supplementary Material.

Array Designs: Dryad data repository (doi:10.5061/dryad.cv35b)Selected DNA consensus sequences: GenBank accessions JX45971, JX454972, JX454974, JX454976, JX454978, JX454985.

Analysis pipeline R scripts: Dryad data repository (doi:10.5061/dryad.cv35b).

Nuclear consensus sequences for beaked whales from NGS SNP discovery: Supplementary Material.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Indexed library preparation and Agilent capture array hybridization for Illumina Genome Analyzer sequencing

Table S1. 75 Tursiops truncatus CATS orthologs (Aitken *et al.* 2004) for cetacean cross-species nuclear capture array. Orange highlighting represents exons based on annotation from the Ensembl Genome Browser (http://uswest.ensembl.org/index.html)

Table S2. Turtle sequence fragments generated from sequencing known microsatellite flanking regions and amplified fragment length polymorphism (AFLP) fragments (Roden *et al.* 2009 and unpublished data)

Table S3. Supplementary Table 3.Accession numbers and file sizes for raw sequence data (Fastq) files and BAM assembly files for mitogenomes and nuclear loci from published green turtle (C. mydas) mitogenomes. More data about the samples are in Duchene *et al.* 2012.

Table S4. Consensus sequences (relative to the capture array reference sequence) for Ziphius cavirostris nuclear loci, with identified SNPs. Loci are replicated for design of multiple SNPs in the same locus (SNP for assay design indicated by square brackets)

Table S5. Consensus sequences (relative to the capture array reference sequence) for Mesoplodon densirostris nuclear loci, with identified SNPs. Loci are replicated for design of multiple SNPs in the same locus (SNP for assay design indicated by square brackets)

Table S6. Loci coverage per species