MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) doi: 10.1111/1755-0998.12448

SPECIAL ISSUE: SEQUENCE CAPTURE

In-solution hybridization for mammalian mitogenome enrichment: pros, cons and challenges associated with multiplexing degraded DNA

MELISSA T. R. HAWKINS,*,†,‡ COURTNEY A. HOFMAN,*,§,¶ TAYLOR CALLICRATE,*,** MOLLY M. MCDONOUGH,*,† MIRIAN T. N. TSUCHIYA,*,†,‡ ELIÉCER E. GUTIÉRREZ,*,† KRISTOFER M. HELGEN† and JESUS E. MALDONADO*,†

*Center for Conservation and Evolutionary Genetics, Smithsonian Conservation Biology Institute, National Zoological Park, Washington, DC 20008, USA, †Division of Mammals, National Museum of Natural History, MRC 108, Smithsonian Institution, PO Box 37012, Washington, DC 20013-7012, USA, ‡Department of Environmental Science & Policy, George Mason University, Fairfax, VA 22030, USA, §Program in Human Ecology and Archaeobiology, Department of Anthropology, National Museum of Natural History, Smithsonian Institution, PO Box 37012, Washington, DC 20013-7012, USA, ¶Department of Anthropology, University of Maryland, College Park, MD 20742, USA, **Department of Animal & Avian Sciences, University of Maryland, College Park, MD 20742, USA

Abstract

Here, we present a set of RNA-based probes for whole mitochondrial genome in-solution enrichment, targeting a diversity of mammalian mitogenomes. This probes set was designed from seven mammalian orders and tested to determine the utility for enriching degraded DNA. We generated 63 mitogenomes representing five orders and 22 genera of mammals that yielded varying coverage ranging from 0 to >5400X. Based on a threshold of 70% mitogenome recovery and at least 10× average coverage, 32 individuals or 51% of samples were considered successful. The estimated sequence divergence of samples from the probe sequences used to construct the array ranged up to nearly 20%. Sample type was more predictive of mitogenome recovery than sample age. The proportion of reads from each individual in multiplexed enrichments was highly skewed, with each pool having one sample that yielded a majority of the reads. Recovery across each mitochondrial gene varied with most samples exhibiting regions with gaps or ambiguous sites. We estimated the ability of the probes to capture mitogenomes from a diversity of mammalian taxa not included here by performing a clustering analysis of published sequences for 100 taxa representing most mammalian orders. Our study demonstrates that a general array can be cost and time effective when there is a need to screen a modest number of individuals from a variety of taxa. We also address the practical concerns for using such a tool, with regard to pooling samples, generating high quality mitogenomes and detail a pipeline to remove chimeric molecules.

Keywords: high throughput sequencing, mitogenome, museomics, targeted sequence capture

Received 16 February 2015; revision received 16 July 2015; accepted 20 July 2015

Introduction

Recent advances in high throughput DNA sequencing technology (HTS) have made large-scale genomic studies more cost-effective, especially in nonmodel organisms (Glenn 2011). However, sample quality continues to burden those who wish to study rare, elusive or even extinct species. For such species, low quality DNA samples may be the only resource available, and include such samples as faecal, road-killed or museum specimens. However, DNA isolated from these types of samples is often fragmented and in low concentration, subject to hydrolytic damage (i.e. cytosine deamination) oxidation (Pääbo *et al.* 1989; Shapiro & Hofreiter 2012), and contamination from exogenous sources and inhibitors, making it extremely challenging to use in genetic analysis (Taberlet *et al.* 1997; Taberlet & Luikart 1999). With recently developed HTS, only small quantities of short DNA fragments are required thereby avoiding several traditional limitations of degraded samples. However, another remaining challenge from these valuable types of samples is to

Correspondence: Melissa T. R. Hawkins, Fax: (202)633-0182; E-mail: robertsmt@si.edu

enrich the endogenous DNA against the (often) large amounts of exogenous DNA. The ability to select and target the appropriate markers would represent a powerful new tool for molecular ecology, phylogenetics, archaeology and biomedical studies.

To accommodate a diverse range of phylogenetic/population genetic questions in our research group, we developed a set of RNA based probes designed to capture and enrich mitochondrial genomes (henceforth mitogenomes) representing seven mammalian orders from degraded DNA samples. It is widely recognized that multiple, independently inherited markers are necessary to reliably infer phylogenetic relationships (Maddison 1997; Brito & Edwards 2009; Toews & Brelsford 2012). Yet studies based on mitochondrial markers continue to be useful due to their high copy number in each cell, an abundance of data publicly available from databases like GenBank, and relatively fast rates of nucleotide substitution for resolving shallow phylogenetic, and population-level questions, and utility for molecular divergence dating (Clark & Hartl 1997; Gutiérrez *et al.* 2010; Duchêne *et al.* 2011; Larsen *et al.* 2012; Siles *et al.* 2013; Voss *et al.* 2013; Gutiérrez *et al.* 2014; Petrova *et al.* 2014; Hofman *et al.* 2015). Mitochondrial genomes have become valuable markers for studies largely based on degraded DNA, i.e. DNA sourced from historical museum specimens (Miller *et al.* 2009; Mason *et al.* 2011; Guschanski *et al.* 2013), archaeological contexts (Krause *et al.* 2010; Adler *et al.* 2013), paleontological sites (Rogaev & Moliaka 2006; Prüfer *et al.* 2014), noninvasive samples (Taberlet *et al.* 1997; Taberlet & Luikart 1999; Bozarth *et al.* 2011a,b; Ahlering *et al.* 2012), or just poorly preserved samples (henceforth all will be termed aDNA).

Museum specimens are increasingly valuable resources for genomic studies. Museum collections often house the only representatives of particular populations or taxa, including endangered and extinct species. As many of these samples are irreplaceable, and as it is often difficult to obtain permission to destructively sample specimens, effective usage of the limited DNA extracts acquired is essential. Traditional Sanger sequencing of museum samples has limitations due to poor PCR amplification success, issues with contamination and often few sequences generated after investing large amounts of time and effort. Historically, shotgun sequencing was the only method used to generate large data sets from aDNA (Hofreiter *et al.* 2001; Pääbo *et al.* 2004; Krause *et al.* 2010) and required deep sequencing (in terms of coverage) to ensure all of the desired fragments were sequenced. Furthermore, shotgun sequencing produces a large proportion of sequences that are off-target, from exogenous sources and a high percentage of the reads are discarded (Green *et al.* 2006). Therefore, many fewer individuals can be pooled on the same sequencing run, with a majority of output consumed by off-target sequences.

Targeted in-solution enrichment, also know as in-solution capture, has made ancient and degraded DNA research more feasible for a large number of samples or taxa (Bi *et al.* 2013). Typically, closely related taxa are used to enrich a degraded DNA extract for targeted loci using sequence probes (McCormack & Faircloth 2013). However, in some cases, the closest relatives of the taxa of interest (to be targeted) are unknown or not available. Historically, studies based on sequences obtained from aDNA via in-solution enrichment often included only a small number of samples (due to technological constraints) and the cost per sample was relatively high because only a single sample was used per reaction (hereafter singleplex). In addition, probes generated in-house via PCR amplification of tissue samples (to yield DNA based probes, as done in Mason *et al.* 2011) are limited by the availability of these tissues, which can be difficult to obtain for endangered taxa or those of international origin. Furthermore, when a diverse group of taxa need to be targeted, separate probe design and synthesis for each taxon becomes expensive.

Here, we present a cost effective method specifically designed to capture mitogenomes from degraded aDNA but that can also be used to capture high-quality samples. The goal of this study was to test this diverse probe set on degraded museum samples and fresh tissue samples, to determine the best practices for this type of tool. This single probe set was designed to work on species selected across seven mammalian orders that represents 2500 species of mammals (approximately 50% of extant mammalian diversity) within these orders. We developed an in-solution set of RNA probes from mitogenomes using publically available and recently sequenced taxa to address phylogenetic and population genetic question for a diverse group of mammals. We demonstrate the utility of multiplexing samples and diluting RNA probes for recovering complete mitogenome sequences of samples representing five mammalian orders with a moderate amount of degradation (obtained from museum specimens) while minimizing cost and time and maximizing effectiveness. We also provide a complete analysis pipeline designed to yield high quality sequence data when multiplexing samples. Additional quality control steps to evaluate sequences for the presence of chimeric molecules, and a tool to test novel taxa for utility with this probe set, are provided in the Supporting Information. In addition, 100 published mtDNA sequences spanning the diversity of mammalian orders were tested for genetic similarity to the included probes, to evaluate if the probe set

designed here could be applied to a wider diversity of mammalian species.

Methods

Capture probe design

Complete mitochondrial genomes from a broad diversity of mammalian taxa were used to design RNA baits for in-solution enrichment. Taxa spanned seven mammalian orders (Monotremata, Rodentia, Scandentia, Carnivora, Chiroptera, Lipotyphla (Soricomorpha), and Artiodactyla) representing 22 genera (Table S1, Supporting Information, Wilson & Reeder 2005; Asher & Helgen 2010). Sequences were downloaded from GenBank (www.ncbi.nlm.nih.gov/genbank) and aligned using MAFFT v1.3.3 (Katoh & Standley 2013) in GENEIOUS v 7.1.4. Capture of targets up to 10–13% divergent from the bait sequence has been reported (Mason *et al.* 2011; Hancock-Hanser *et al.* 2013), and tolerances of up to 20% locally divergent has been observed in some cases (MYcroarray, personal communication). We used the CD-HIT-EST clustering algorithm (www.weizhong-lab.ucsd.edu/cd-hit_suite; Li & Godzik 2006; Huang *et al.* 2010) to test the similarity among the included whole mitogenomes (WMG hereafter) to remove those that would not add additional mitochondrial diversity to the probe set. We set a minimum threshold of 10% divergence between the generated clusters (from CD-HIT-EST) and included only a single representative sequence from each of the clusters (i.e. the genetic distance within a cluster was maximally 10%). This reduced the original 22 WMG to 16 representing all seven mammalian orders included in the design. In addition to these published WMG, we sequenced 15 additional novel WMG's [derived from fresh tissue Long Range (LR) PCR of the mitochondrion, described in the Supporting Information] on a Roche 454 GS Junior or Illumina MiSeq platform for several rodent taxa and clustered these through CD-HIT-EST to remove highly similar sequences. This reduced the 15 newly sequenced WMG to 10, for a final data set of 26 WMG (16 from GenBank and 10 novel WMG).

In addition to the use of WMG's for probe design, we expanded the array to better capture hyper-variable genes, defined here as particular gene regions that undergo faster mutation rates compared to the rest of the mitogenome (Pesole *et al.* 1999). Details of hyper-variable genes included in array design are listed in the Supporting Information.

MYcroarray (Ann Arbor, MI, USA; www.mycroarray.com) performed quality checks of the sequences, and then split them into 120 base pair (bp) probes with 2× tiling. This level of tiling allows for overlapping probes, and in this case the probes were tiled every 60 bp (i.e., one probe would start at position 1 in the mitogenome and continue to position 120; the second probe would start at position 60 and end at position 180 so that each mitogenome position was covered by two probes). This design resulted in 6577 unique probes (see Supporting Information), with 80% representing sequences from WMG alignments and 20% from the hyper-variable regions. These were synthesized in a 20 000 probe MYbaits kit which generates 500 ng per undiluted capture. We used approximately 100 ng of probe per capture pool in this study. Probe dilution and multiplex calculations are further described in the Supporting Information.

Sample selection and extraction

The mammal mitogenome array (hereafter referred to as MMA) was tested using a diverse set of 63 samples, spanning 37 species from five of the seven included mammalian orders (Monotremata, Rodentia, Carnivora, Chiroptera and Artiodactyla; Table 1). Two orders that were included in the probe set were not tested in this experiment (Sorciomorpha and Scandentia). These 63 samples were split amongst 10 multiplexed enrichments. Our goal was to test whether distantly related mammalian taxa can be enriched together with a single array (such that costs are minimized). Most samples (n = 53) were derived from museum specimens collected as early as 1899, and as recently as 2011. One additional enrichment contained a single liver sample that was ground up in a buffer solution in a microtube for allozyme work and kept frozen at −80 °C for >30 years plus an additional nine frozen tissue samples were included in this enrichment test to evaluate the differential enrichment success between degraded and nondegraded samples (all extracted frozen tissues were stored at −20 °C). Additional details regarding sample type, and DNA extraction protocols are detailed in the Supporting Information.

Library preparation and enrichment

We prepared samples for Illumina sequencing using commercially available library preparation kits (Kapa Biosystems Illumina Library Preparation Kit #KK8232; Wilmington, MA, USA). Single indexed TruSeq-style adapters were used (Faircloth & Glenn 2012). Because the majority of samples were derived from museum specimens, endogenous DNA concentration was unknown, with much of the extracted DNA including substantial amounts of bacteria, fungi and other exogenous DNA. To compensate for the unknown

Table 1 Summary of mammal mitogenome array results, arranged by genus. These results are averaged across the number of individuals in each genus

Order	Family	Genus (number of species)	Average no. of raw reads	Average no. of mapped reads	Mean average coverage	Average % of mapped reads
Artiodactyla	Cervidae	*Blastocerus* (1)	8 991 338	4 547 552	34537.5	50.58
		Mazama (6)	3 225 461	146 345	1474.2	12.11
		Odocoileus (4)	6 726 469	1 581 925	17520.9	26.52
		Ozotoceros (1)	533 134	80 275	879.7	15.06
		Pudu (1)	122 282	3740	30.4	3.06
Carnivora	Procyonidae	*Bassaricyon* (1)	1 130 326	642 156	5158.9	56.81
		Bassariscus (1)	1 569 690	795 536	6566.0	50.68
		Nasua (1)	148 564	70 848	577.6	47.69
		Nasuella (1)	175 200	16 401	124.5	9.36
		Potos (1)	2 111 048	1 076 959	7979.6	51.02
		Procyon (1)	661 514	433 356	3556.3	65.51
Chiroptera	Phyllostomidae	*Platyrrhinus* (1)	6 158 018	52 098	462.5	0.85
Monotremata	Tachyglossidae	*Zaglossus (16)*	104 936	53 520	393.0	35.37
Rodentia	Muridae	*Dipodillus* (2)	2 115 117	122 289	1626.6	6.09
		Gerbillus (2)	688 138	31 696	400.3	8.57
	Sciuridae	*Glyphotes* (1)	1 018 043	1405	9.4	0.14
		Callosciurus (5)	519 895	1832	16.8	0.56
		Exilisciurus (1)	43 293	106	0.6	0.25
		Hyosciurus (1)	900 481	5305	46.9	0.59
		Lariscus (5)	718 472	1032	5.5	0.20
		Prosciurillus (6)	2 877 505	39 971	348.7	0.53
		Rhinosciurus (2)	1 010 464	953	5.5	0.34
		Sundasciurus (2)	1 740 101	693	3.9	0.14

concentration of target DNA, a large volume of DNA extract (50 μL) was used for library preparation. Minor modifications were made to the manufacturer's protocol (see Supporting Information) including additional PCR cycles on degraded samples (18 cycles for degraded DNA from museum samples, and 10–14 for frozen tissues). The success of library preparation was determined by visualization on an agarose gel. Additional details regarding postlibrary preparation sample manipulation and multiplex information are detailed in the Supporting Information.

Each pool of libraries was incubated with the RNA probes and buffers as described in the MYcroarray protocol for 24 h at 65 °C. Following incubation, DNA was separated from the probes via magnetic beads and purified with QiaQuick PCR Purification Kits (Qiagen) following MYcroarray's enrichment protocol (version 1.3.8) Detailed protocols for MYbaits kits have been published online (http://ultraconserved.org/#protocols; http://www.mycroarray.com/pdf/MYbaits-manual.pdf).

A total of 10 enrichments were performed (with multiplexes of 4–10 samples per pool), with nine pools containing degraded museum samples, and a single enrichment pool containing fresh tissue samples. Postenrichment pools were amplified for 25 cycles to produce a high enough concentration for gel extraction. QiaQuick Gel Extraction Kits (Qiagen) were used to size select the enriched pools for ~200–500 bp fragments and to remove residual adapter and primer dimer.

Sequencing

Quantitative PCR was performed on enriched pools using an Illumina Library Quantification Kit (Kapa Biosystems) with two replicates of 1:1000, 1:2000 and 1:4000 dilutions for each pool. Pools were combined in equimolar ratios based on the number of samples in each pool. These 63 samples were sequenced with paired-end chemistry and with read length of 143 bp on a single lane of an Illumina HiSeq2500 at the Semel Institute UCLA Neurosciences Genomics Core, and reads were demultiplexed at the core facility.

Test for enrichment success

To determine the efficiency of the hybridization, a relative qPCR was performed on a subset of enrichment pools (specifically Enr. 2 and Enr. 3) to calculate the fold enrichment of pools pre- and postenrichment. Universal mammalian mitochondrial cytochrome *b*

primers were used in the qPCR (detailed in Supporting Information).

Assembly of mitogenomes

To determine the efficiency of the MMA for a diverse set of taxa that vary in level of divergence from the probe set, we tested both de novo assembly and read mapping to reconstruct the mitogenomes. The degraded nature of the aDNA extracted from museum samples yielded sequences of lower quality and shorter length, so to compensate for this, we merged the forward and reverse paired reads with the program PEAR v0.9.4. (Zhang *et al.* 2014). Merging joins the forward and reverse reads when they have a 10 bp or greater overlap (PEAR default setting), which happens when short library inserts are sequenced. This resulted in both longer fragments for mapping and higher quality scores where the forward and reverse sequences overlap. The read merging was not necessary for the 10 frozen tissue samples as the reads were too long to have adequate nucleotide overlap as required by PEAR v0.9.4. All sequences (including those generated from frozen tissue samples) were screened for the presence of adapter sequences, which were removed with CUTADAPT v.1.4.2 (Martin 2011). Next, PRINSEQ-LITE v.0.20.4 (Schmieder & Edwards 2011) was used for quality filtering, trimming reads with average quality scores below 20 and exact PCR replicates (more than three identical copies). The filtered reads were then mapped to a reference sequence of the most closely related species using BWA v.7.10 (Li & Durbin 2009). The 'bwa aln' and 'samse' as well as the 'bwa mem' algorithms were tested on the degraded samples, with 'bwa aln' conducted as specified in Kircher (2012). The reads corresponding to the 10 frozen tissue samples were mapped using the 'bwa mem' algorithm. Additional read mapping programs were tested on a subset of individuals to evaluate the performance of the various mapping algorithms (see Supporting Information).

De novo assembly for distantly related taxa

Another objective of this study was to test the ability of the MMA to produce complete mitogenomes from taxa for which only distantly related reference sequences were publicly available. Specifically, we used the MMA to enrich libraries of four pygmy gerbils (*Gerbillus* spp.). Baits derived from a Mongolian gerbil (*Meriones unguiculatus*, KF425526), the closest relative of the genus *Gerbillus* available on GenBank, was included in the MMA design (approximately 20% divergent) plus additional *Gerbillus* sp. cytochrome *b*, ND5, and control region sequences generated in house via Sanger sequencing using protocols detailed in McDonough *et al.* (2013). A novel reference genome for pygmy gerbils was constructed from de novo assembly of LR PCR products (detailed in Supporting Information) using MIRA v.4.0.2 (Chevreux *et al.* 1999). Contigs derived from MIRA v.4.0.2 were then mapped back to the *M. unguiculatus* mitogenome using GENEIOUS v.7.1.4 to generate the resulting consensus sequences.

Clustering test for off target taxa

A set of 100 GenBank sequences (of various genes depending on availability) from mammalian taxa not included in this array was tested with CD-HIT-EST (Huang *et al.* 2010) to estimate the possibility that the MMA could enrich additional mammalian taxa. All tested sequence accession nos are included in Table S4 (Supporting Information). A mammalian tree generated by Meredith *et al.* (2011) was used to select the species tested, and representatives from most tips were included in our analysis. We evaluated the likelihood that the test sequence would enrich by observing which sequences met a 90% similarity threshold (i.e. 10% sequence divergence when comparing the GenBank sequence against all sequences used in probe synthesis). This clustering method calculated the percent similarity across the entire length of the sequences obtained from GenBank, so a test sequence was required to match with at least 90% of the entire length of the probe set sequence with which it clustered. Similarity thresholds of 85% and 80% were also evaluated to determine sequences that would likely hybridize with this probe set after optimization at a lower (more relaxed) hybridization temperature or different bait tiling strategy (complete list of sequences is provided in Table S3, Supporting Information).

Results

Size distribution of pre- and postenrichment

The libraries were run on agarose gels before enrichment to check for uniform library preparation, and subsequently all pools were visualized after enrichment (following amplification). The size range distribution of the pre-enrichment pools varied significantly between samples, but the majority showed a bright band of DNA library between 150 and 250 bp (DNA insert size ranging ~10–110 bp). Some samples had a large smear with size spanning approximately 150–1000 bp, which may indicate exogenous DNA contamination (potentially fungal or bacterial contaminants). Following enrichment, the 10 pools had a narrow size distribution of approximately 150–250 bp. The tenth pool of fresh tissue samples was slightly larger. This further suggests that the

very large fragments visualized after library preparation were either too long to hybridize to the probes, or more likely, of exogenous origin.

Test for enrichment success

We performed a relative quantification RT-PCR to ensure that the in-solution hybridization was efficiently targeting mitochondrial molecules. The two pools tested (Enr. 2 and Enr. 3) had a large increase in the amount of detected mitochondrial DNA (15 and 2700 fold increase for Enr. 2 and 3, respectively) compared to pre-enrichment. In addition to the two pools, several unpooled samples were also quantified to determine the amplification cycle take-off-point, and the results are presented in the Supporting Information. Specifically, two samples were included from Enr. 2 and 3, sample 2*b* did not hit the take-off-point until cycle 35, much later than the combined samples for Enr. 2 (cycle 28). Sample 3*a* took off at cycle 36, which was slightly earlier than the un-enriched Enr. 3 pool (cycle 38), likely because 3*a* constituted the majority of reads (78%) for Enr. 3.

Assembly of mitogenomes

Read mapping was tested with BWA v7.1.0 Li & Durbin (2009), BOWTIE v2.2.4 Langmead *et al.* (2009) and STAMPY v1.0.26 Lunter & Goodson (2010). STAMPY mapped the most number of reads, followed by BOWTIE, and BWA mapped the fewest reads. Although the number of mapped reads may indicate successful mapping, we compared the three methods for a subset of samples, and recovered a much higher number of heterozygous sites from STAMPY and BOWTIE. Here, we remained conservative and assumed mitochondrial haploidy and therefore utilized BWA following the widely used ancient DNA specific parameters detailed in Kircher (2012). Additional detail regarding the detection of heterozygous sites and coverage bias can be found in the Supporting Information.

Variation was observed in both the number of raw reads per sample and the extent of mitogenome coverage. This variation likely reflects the quality of the DNA samples used as starting material as well as hybridization bias resulting from pooling the captures. For example, even though the samples were pooled in equimolar ratios, all enrichments included one sample that had a higher percentage of the raw reads when compared to the others (at least 75% more than the second sample in the enrichments of degraded material; Fig. 1). On average, the dominant samples (defined here as the sample which recovered the majority of reads postcapture) in each pool had 63.7% of the total number of raw reads per enrichment, varying from 31.0% to 98.4%. The enrichment with the most balanced reads (dominant sample with 31.0%) was that resulting from pooling

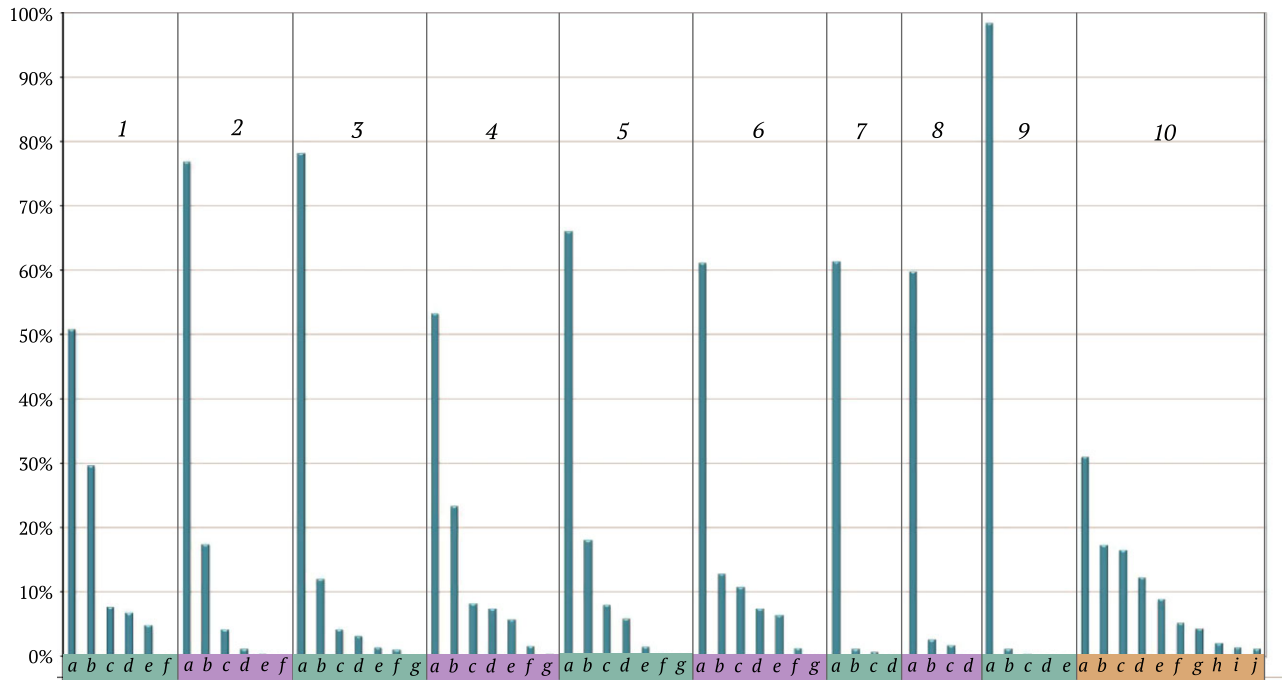

Fig. 1 The percentage of reads attributed to each sample across 10 multiplexed enrichment pools. Samples names have been replaced with abbreviations, with the enrichment pool listed in the figure, and the sample denoted as a letter along the *x*-axis (*a*–*e*, etc.) and each enrichment is labelled in the whitespace along the *y*-axis, all samples are defined in Table 2.

libraries corresponding to the 10 freshly preserved tissue samples. The distribution of reads across all samples, separated by enrichment is shown in Fig. 1.

Overall, the percentage of raw reads per sample varied from 0.02% of the total reads for the enrichment pool (*Zaglossus attenboroughi*) to 98.4% (*Blastocerus dichotomus*). Those two extreme cases were from the same enrichment (Enr. 9), and both were historical museum samples (specimens of *Z. attenboroughi*, collected in 1961, and *B. dichotomus*, collected in 1941) derived from skin and museum 'tissue' material, respectively. The most dominant samples belonged to the orders Artiodactyla (5), followed by Rodentia (3) and Chiroptera (1). As expected, such variability in the number of raw reads per sample resulted in variation of mitochondrial genome coverage (Figs 1 and 2). Mean mitogenome coverage was 685X, ranging from 0X (*Zaglossus bartoni*) to 5421X (*B. dichotomus*). The percentage of mapped reads varied from 0% to 81.72%, with an average of 11.89%. From the total of 63 samples included in this study, 41 samples had average coverage higher than 10X, 37 samples higher than 30× and 31 samples >100X (including all 10 frozen tissue samples). A summary of the performance of the MMA is detailed in Tables 1 and 2. Interestingly, 14 of 63 (or ~30%) of the enrichments had 5× coverage or less across the entire mitogenome, and 22 samples had 10× or lower coverage (Table 2). This indicates poor enrichment for those samples. Of the 14 'poorest quality' enrichments, two were from bone (of 10, or 20% of bone samples), nine from museum 'tissue' (of 29, or 31%) and three from skin (of 8, or 38%). The percentage of the mitogenome recovered from BWA read mapping ranged from 1.7% to 100% (Table 2). We recovered 38 mitogenomes with >70% of the mitogenome represented, 32 with >80%, 25 with >90% and 23 with >95% of the mitogenome sequence. Gaps in the mitogenomes were not clustered in a single location, but were distributed across the mitochondrion. Many samples had gaps in the control region, more so than other genes. We decided to combine the coverage with percent recovered for our final determination of a successful mitogenome (>70% at 10X coverage). Additional detailed comparisons from the MEGA comparisons are detailed in Table S2 (Supporting Information).

Nuclear copies of mitochondrial DNA (NuMT's)

For all samples with 10X average coverage and 70% complete sequences we translated all protein coding genes to evaluate the presence of nuclear copies of mtDNA (NuMTs). After evaluating 32 samples, we detected few NuMTs. As recommended in Mason *et al.* (2011), we

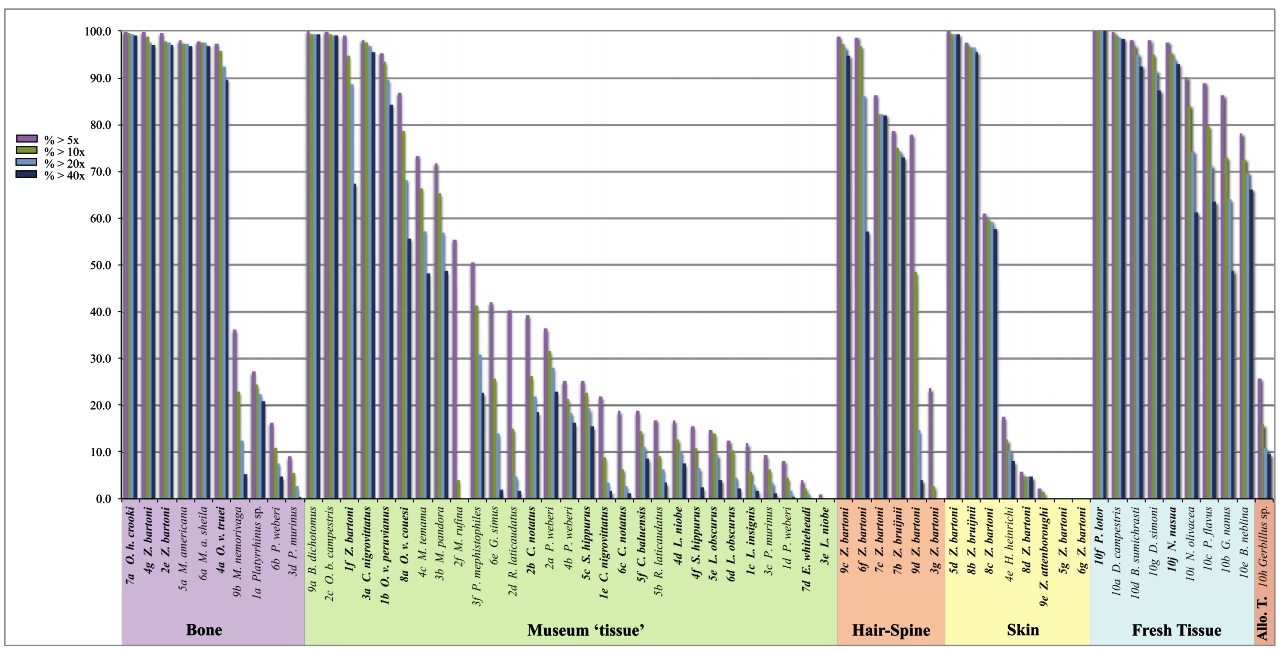

Fig. 2 Percentage of mitogenome (*y*-axis) sequenced to different levels of coverage (5, 10, 20 and 40X). The different types of sample are indicated along the *x*-axis. Bold sample names along the *x*-axis indicate taxa for which within-genus probes were included in the MMA probe set. Sample names have been abbreviated to match Fig. 1 and Table 2. The sample labelled 'Allo. T.' represents the liver tissue which was previously used for allozyme analysis and homogenized then frozen for approximately 30 years (hence the lower coverage is due to the degradation of the tissue).

Table 2 Detailed results of mammal mitogenome array enrichments, with BWA average coverage, and percentage of mitogenome recovered. Enr no. corresponds to the same IDs presented in Fig. 1. The column '% of mitogenome recovered' was obtained by determining the number of recovered bases divided by the length of the reference sequence. Samples in bold met our metric for 'success' (10X average coverage and at least 70% of the mitogenome recovered)

Enrichment	Enr no.	Species	Catalog no.	No. raw reads	% of Total no. reads	No. reads mapped w/BWA	Average coverage	% Recovered
Enr 1	1a	*Platyrrhinus* sp.	AMNH 92247	6 158 018	50.8	52 098	463	38.77
	1b	***Odocoileus virginianus peruvianus***	**FMNH 78421**	**3 590 072**	**29.6**	**77 589**	**547.7**	**97.40**
	1c	*Lariscus insignis*	ZRC 4 3088	930 129	7.7	440	2.4	21.29
	1d	*Prosciurillus weberi*	MZB 6256	821 737	6.8	244	1.3	18.37
	1e	*Callosciurus nigrovittatus*	ZRC 4131	595 652	4.9	766	4	70.41
	1f	***Zaglossus bartoni***	**KMH 2935**	**16 416**	**0.1**	**6592**	**60.8**	**97.78**
Enr 2	2a	*P. weberi*	MZB 6254	9 894 041	76.8	207 316	1890.5	47.86
	2b	***Callosciurus notatus***	**ZRC SMN108**	**2 249 185**	**17.5**	**26 728**	**151.3**	**79.92**
	2c	***Ozotoceros bezoarticus campestris***	**FMNH 28297**	**533 134**	**4.1**	**80 275**	**879.7**	**99.99**
	2d	*Rhinosciurus laticaudatus*	ZRC 3173	151 065	1.2	955	5.6	84.21
	2e	***Z. bartoni***	**AMNH 195373**	**44 683**	**0.3**	**32 818**	**334.8**	**99.10**
	2f	*Mazama rufina*	FMNH 70563	4664	0.0	965	5.1	99.90
Enr 3	**3a**	***C. nigrovittatus***	**ZRC 4093**	**9 132 095**	**78.2**	**479 122**	**3087.4**	**98.85**
	3b	***Mazama pandora***	**KU 93857**	**1 401 489**	**12.0**	**52 913**	**464.2**	**82.36**
	3c	*Prosciurillus murinus*	MZB 5977	484 178	4.1	376	2.1	21.22
	3d	*P. murinus*	MZB 5973	368 736	3.2	323	1.7	19.82
	3e	*Lariscus niobe*	ZRC 48486	160 256	1.4	45	0.2	10.96
	3f	***Pudu mephistophiles***	**AMNH 181506**	**122 282**	**1.0**	**3740**	**30.4**	**70.66**
	3g	*Z. bartoni*	AMNH 194702	5654	0.0	375	2.8	73.11
Enr 4	**4a**	***Odocoileus virginianus truei***	**KU 149129**	**8 342 731**	**53.3**	**355 838**	**3391.6**	**98.85**
	4b	*P. weberi*	MZB 6255	3 664 756	23.4	30 288	189.4	36.49
	4c	***Mazama temama***	**KU 82215**	**1 281 287**	**8.2**	**25 522**	**202.5**	**85.64**
	4d	*L. niobe*	ZRC 48477	1 158 270	7.4	3236	16.9	27.10
	4e	*Hyosciurus heinrichi*	MZB 34908	900 481	5.7	5305	46.9	27.17
	4f	*Sundasciurus hippurus*	SM 2371	252 938	1.6	693	3.9	27.24
	4g	***Z. bartoni***	**RMNH 23319**	**66 175**	**0.4**	**51 399**	**406.1**	**99.94**
Enr 5	**5a**	***Mazama americana***	**AMNH 67109**	**6 836 872**	**66.1**	**297 703**	**3013.7**	**98.62**
	5b	*R. laticaudatus*	ZRC 3551	1 869 862	18.1	950	5.3	68.28
	5c	*S. hippurus*	SM NH19	822 777	8.0	8218	47.9	33.78
	5d	***Z. bartoni***	**AMNH 104020**	**607 079**	**5.9**	**462 767**	**3383.9**	**99.96**
	5e	*Lariscus obscurus*	ZRC 48469	157 710	1.5	946	5.2	23.93
	5f	*Callosciurus baluensis*	SM NH1	36 282	0.4	3353	21.6	44.32
	5g	*Z. bartoni*	AMNH 195146	8062	0.1	10	0.1	5.85
Enr 6	**6a**	***Mazama americana sheila***	**USNM 443588**	**9 721 044**	**61.2**	**461 084**	**4871.1**	**99.99**
	6b	*P. weberi*	MZB 6252	2 031 581	12.8	1278	7.3	27.54
	6c	*C. notatus*	ZRC No. 33	1 724 509	10.9	593	3.4	70.11
	6d	*L. obscurus*	ZRC 48471	1 185 996	7.5	493	2.7	23.22
	6e	*Glyphotes simus*	NH 1832 SM	1 018 043	6.4	1405	9.4	76.73
	6f	***Z. bartoni***	**AMNH 157072**	**202 244**	**1.3**	**6243**	**49.8**	**99.95**
	6g	*Z. bartoni*	AMNH 195147	7827	0.0	3	0	1.70
Enr 7	**7a**	***Odocoileus hemionus crooki***	**USNM 99455**	**9 564 580**	**61.3**	**71 497**	**612.4**	**99.98**
	7b	***Zaglossus bruijni***	**USNM 268763**	**174 228**	**1.1**	**31 424**	**225.1**	**76.98**
	7c	***Z. bartoni***	**AMNH 190859**	**106 011**	**0.7**	**23 700**	**238.8**	**88.05**
	7d	*Exilisciurus whiteheadi*	SM NH1440	43 293	0.3	106	0.6	8.56
Enr 8	**8a**	***Odocoileus virginianus couesi***	**USNM 99351**	**5 408 492**	**59.8**	**70 886**	**838.9**	**94.49**

Table 2 (Continued)

Enrichment	Enr no.	Species	Catalog no.	No. raw reads	% of Total no. reads	No. reads mapped w/BWA	Average coverage	% Recovered
	8b	***Z. bruijnii***	**AMNH 249921**	**235 259**	**2.6**	**164 121**	**1159.9**	**98.27**
	8c	*Z. bartoni*	RMNH 325	155 340	1.7	43 911	219.3	63.24
	8d	*Z. bartoni*	AMNH 190863	12 944	0.1	1878	9.8	9.87
Enr 9	**9a**	***Blastocerus dichotomus***	**FMNH 52329**	**8 991 338**	**98.4**	**613 409**	**5420.9**	**100.00**
	9b	*Mazama nemorivaga*	AMNH 96171	107 411	1.2	1002	8.3	68.05
	9c	***Z. bartoni***	**AMNH 66194**	**29 016**	**0.3**	**23 712**	**170.5**	**99.63**
	9d	***Z. bartoni***	**AMNH 190862**	**6167**	**0.1**	**2846**	**19.4**	**92.46**
	9e	*Zaglossus attenboroughi*	RMNH 17301	1873	0.0	62	0.4	9.77
Enr 10	**10a**	***Dipodillus campestris***	**TK40900**	**1 981 049**	**31.0**	**217 850**	**3101.9**	**100.00**
	10b	***Gerbillus nanus***	**TK40880**	**1 103 874**	**17.3**	**22 187**	**275.7**	**97.34**
	10c	***Potos flavus***	**H015**	**2 111 048**	**16.5**	**1 076 959**	**7979.6**	**99.10**
	10d	***Bassariscus sumichrasti***	**BS**	**1 569 690**	**12.3**	**795 536**	**6566**	**99.89**
	10e	***Bassaricyon neblina***	**H021**	**1 130 326**	**8.9**	**642 156**	**5158.9**	**89.05**
	10f	***Procyon lotor***	**NDM 3842**	**661 514**	**5.2**	**433 356**	**3556.3**	**99.93**
	10g	***Dipodillus simoni***	**TK40906**	**272 401**	**4.3**	**41 204**	**524.8**	**100.00**
	10h	*Gerbillus* sp.	TK25614	129 919	2.0	2650	30.5	69.30
	10i	***Nasuella olivacea***	**H010**	**175 200**	**1.4**	**16 401**	**124.5**	**94.86**
	10j	***Nasua nasua***	**89-325**	**148 564**	**1.2**	**70 848**	**577.6**	**99.33**

evaluated assemblies for indels and stop codons as primary evidence of nuclear copies proliferating in assemblies. High frequency SNPs can also be evidence of nonterminating NuMTs, which we detected more commonly in samples of low coverage (Table S3, Supporting Information).

De novo assembly for distantly related taxa

The consensus *Gerbillus* mitogenome generated using MIRA v.4.0.2 from the LR PCR amplification was ~19% divergent from the closest sequences available on GenBank (*Meriones unguiculatus*; KF425526). The number of enriched pygmy gerbil (*Gerbillus* sp.) reads that mapped to *M. unguiculatus* (KF425526) mitogenome using BWA ranged from 16 850 to 17 960; whereas the number of reads mapped to the de novo *Gerbillus* sp. assembly ranged from 22 187 to 217 850. Average coverage ranged from 30.5 to 3102× for enriched *Gerbillus* sp. mapped to *Meriones*, compared to 238–10 424× for enriched *Gerbillus* sp. mapped to the de novo assembly. The percentage of sequences 'on target' for gerbil sequences enriched with MMA mapped to *Meriones* ranged from 0.91% to 13.0%. The percentage of sequences on target was slightly higher (2.0–15.1%) when enriched sequences were mapped to the de novo assembly derived from the *Gerbillus* sp. samples. To provide a comparable data set for all individuals, we have included the read mapping data for the *Gerbillus* samples. We realize that the mapping is directly affected by the genetic distance of the reference sequence to the samples of interest, and show here that the enrichment of *Gerbillus* was possible when including a distantly related mitogenome.

Clustering test

To assess the sequence similarity of our probe set compared to a broad array of taxa across the mammalian tree of life, 100 GenBank sequences were clustered with the sequences included in our probe set (Table S3, Supporting Information). We have also indicated whether these sequences would potentially hybridize to the current probe set based on empirical results of up to 13% sequence divergence successfully hybridizing (as published in Mason *et al.* 2011; Hancock-Hanser *et al.* 2013). From the clustering test of these GenBank sequences, we found 22 additional species that would likely enrich with slight modification of the hybridization incubation temperature (80% or greater similarity to the MMA probe set, Table 3). Of these, 10 (from species across five mammalian families) would likely enrich with little or no modification to the manufacturer's protocol (85–90% nucleotide similarity). We included 17 WMG of rodent and cervid species in the MMA design, and consequently found several additional closely

Table 3 Clustering test results for the 22 taxa that resulted in clusters above 80% sequence similarity. The full results from all 100 sequences are presented in Table S3

No.	Subclass	Order	Family	Species	GenBank	Gene	90%	85%	80%
1	Monotremata	Monotremata	Tachyglossidae	*Tachyglossus aculeatus*	NC_003321.1	Complete genome	Y- 93%	Y	Y
2	Placentalia	Artiodactyla	Giraffidae	*Okapia johnstoni*	AY012146.1	Partial 12S rRNA and tRNA-VAL	x	Y- 87%	Y
3	Placentalia	Carnivora	Canidae	*Canis simensis*	AF028216.1	Complete COII	Y- 95%	Y	Y
4	Placentalia	Carnivora	Viverridae	*Paguma larvata*	AF125151.2	Complete cytochrome *b*	x	x	Y- 81%
5	Placentalia	Cetacea	Ziphiidae	*Hyperoodon ampullatus*	KF281660.1	Partial CO1 gene	x	x	Y- 81%
6	Placentalia	Chiroptera	Mystacinidae	*Mystacina tuberculata*	AY197327.1	Partial 12S rRNA	x	Y- 86%	Y
7	Placentalia	Chiroptera	Myzopodidae	*Myzopoda aurita*	AF345926.1	Complete 12S, 16S, tRNA	x	x	Y- 81%
8	Placentalia	Chiroptera	Pteropodidae	*Desmalopex microleucopterus*	EU339339.1	Partial 12S rRNA	x	x	Y- 82%
9	Placentalia	Perissodactyla	Tapiridae	*Tapirus* sp.	GU593676.1	Partial COII gene	x	x	Y- 81%
10	Placentalia	Rodentia	Cricetidae	*Cricetus cricetus*	KC953838.1	Partial CO1 gene	x	x	Y- 80%
11	Placentalia	Rodentia	Gliridae	*Graphiurus murinus*	U67287.1	partial 12S rRNA	x	x	Y- 83%
12	Placentalia	Rodentia	Muridae	*Apodemus agrarius*	AB303226.1	Complete cytochrome *b*	x	x	Y- 81%
13	Placentalia	Rodentia	Muridae	*Lemniscomys macculus*	AF141268.2	Partial 12S rRNA	x	Y - 89%	Y
14	Placentalia	Rodentia	Muridae	*Maxomys* sp.	GU294890.1	Partial CO1 gene	x	x	Y- 82%
15	Placentalia	Rodentia	Muridae	*Rhombomys opimus*	KF182214.1	Partial CO1 gene	x	Y- 86%	Y
16	Placentalia	Rodentia	Muridae	*Sekeetamys calurus*	AJ851246.1	Compete 12S rRNA	Y -92%	Y	Y
17	Placentalia	Rodentia	Muridae	*Taeromys celebensis*	KF164226.1	Partial cytochrome *b*	x	x	Y- 80%
18	Placentalia	Rodentia	Muridae	*Tatera indica*	FJ790672.1	Partial CO1 gene	x	Y- 85%	Y
19	Placentalia	Rodentia	Muridae	*Zelotomys hildegardeae*	JQ844108.1	Partial 16S rRNA	x	Y- 89%	Y
20	Placentalia	Rodentia	Pedetidae	*Pedetes surdaster*	U59171.1	Partial 12S rRNA	x	x	Y-82%
21	Placentalia	Scandentia	Ptilocercidae	*Ptilocercus lowii*	AY862166.1	Complete 12S rRNA	x	x	Y- 83%
22	Placentalia	Soricomorpha	Soricidae	*Sorex hoyi*	AF7982	Partial cytochrome *b*	x	Y- 86%	Y

related species that should enrich with the MMA. A detailed workflow of the steps required to test additional taxa for usage with these probes is outlined in the Supporting Information. We should also note that we did not include representative marsupials in our probe design and consequently this resulted in zero marsupial sequences >80% similar to the MMA probes. Therefore, we caution that our probes are not likely to enrich marsupials even with modifications to the hybridization incubation temperature.

Discussion

Here, we show the utility of a mitogenome capture array designed with sequences from a diverse set of mammalian taxa for capturing target sequences from

degraded DNA samples. We show that by performing additional modifications to the manufacturer's protocol (probe dilution and multiplexing reactions), complete mitogenomes can be efficiently generated with a significant reduction in cost. For example, following the methods we employed in this study, Illumina sequencing and capture cost approximately $35/sample (excluding extraction, library preparation and QC costs which vary by protocol) compared to a cost of approximately $200/sample following the manufacturer's protocol. However, we caution that while multiplexing samples reduces costs, it also resulted in dramatically skewed read counts for different individuals within an enrichment pool, especially with degraded samples. Although the samples were pooled in equimolar ratios within a capture, we found that degraded samples had varying ratios of endogenous:exogenous DNA (by evaluating the resulting raw reads alone compared to the skewed number of reads among samples in each pool), which may explain the bias of a greater number of reads in samples with more endogenous DNA. In addition to the ratio of endogenous DNA confounding concentration estimation, certain samples appear to preferentially enrich or amplify and result in a single individual per pool recovering most of the reads per enrichment. Additional quantitative PCR methods of library prepared products may assist in the identification of samples better suited to be multiplexed. For example, samples with similar take-off-points during qPCR may be better suited for multiplexing as this would be a more reliable indicator of similar endogenous DNA concentration than the methods we used here to estimate total DNA concentration.

Based on our results, we do not recommend multiplexing more than five degraded samples and 10 fresh tissue samples in a single hybridization reaction (with an in-solution hybridization kit). Hancock-Hanser *et al.* (2013) demonstrated effective enrichment when multiplexing a larger number of samples, however, this study used a chip-based capture array rather than in-solution, which may explain the difference in capture efficiencies. Instead, if a greater number of enriched samples is desired, we recommend increasing the probe dilution to maximize the number of samples per kit. Additionally, we found that increasing the number of samples in a multiplexing capture pool led to an increased risk of producing chimeric sequences (see Supporting Information). These problems can be overcome with appropriate filtering techniques such as the ones developed here, and in combination with the use of dual indices during library preparation (Kircher *et al.* 2012). Therefore, by combining an appropriate number of individual libraries in pools and enriching them following our protocols and filtering pipelines, we demonstrate a time and cost effective method to obtain WMG without substantial loss in capture efficiency. For example, even though we recovered 38 successful mitogenomes (by our definition of 10X average coverage and recovery rate of at least 70% of the mitogenome), this method still yields considerable savings even though some samples will require a second hybridization (= $70/sample vs. $200/sample for single-sample enrichments).

Methods for mitogenome reconstruction

Here, we describe results from read mapping of mitogenomes to a closely related mitogenome. The diversity of mammalian taxa tested here prevented a standardized method from which to determine the exact effectiveness of the probe set on each species tested. We have used the best available mitogenome as a reference for each sample, which will affect the performance of the read mapping. For some samples (particularly the gerbils), we tested de novo assembly to validate that the sequenced molecules were mitochondrial. One sample that would particularly benefit from additional read mapping and de novo assembly is the *Platyrrhinus* sp., which had over 52 000 quality filtered reads, yet with BWA only 37.9% of the mitogenome was reconstructed. The reference for this sequence (*Sturnira tildae*, HG003314) is from the same family of bats, Phyllostomidae, which contains over 190 species and 52 genera. However, the mapped *Platyrrhinus* sp. reads were 89.5% similar (discounting ambiguous sites) to the reference. This sample is a good candidate for alternative mapping software, or de novo assembly. The availability of closely related reference sequence was very important for recovering molecules. For example, in many of the highly degraded *Zaglossus* samples, 10 of the 16 samples recovered at least 70% of the mitogenome, which is impressive when considering the low number of starting reads in some samples (e.g. *6f*, *Zaglossus bartoni*, detailed below).

Based on our results (Table 2), we recommend aiming for at least 15 000 reads per sample (postquality filtering steps). This is based on our data from museum specimens, and takes into account the percent of endogenous DNA, which is highly variable per sample (as documented when comparing the number of sequenced reads to the number of mapped reads as a crude estimate). A single sample (*6f*, *Z. bartoni*), yielded only 6243 mapped reads, yet had a nearly complete mitogenome due to the close relationship between the sample and reference sequence and inclusion of the reference sequence in the probe design. Most of our successful mitogenomes were assembled from over 100 000 reads per sample, and with that sequencing depth more divergent mitogenomes can be recovered with confidence. From this experiment we

have shown that if a closely related mitogenome is available you can aim for many fewer reads than if there is not a closely related mitogenome available.

Utility for additional mammalian diversity

We found that the baits included in this array successfully enriched mammalian taxa <20% divergent from probe sequences (Tables 1 and 2). However, it is unlikely that the array would successfully enrich more distantly related taxa that were not included in the probe design. For species with <80% genetic similarity (results from clustering test, Table 3), we suggest changing stringency of the enrichment conditions by lowering the hybridization temperature to account for a larger percent divergence from the probe sets. Previous ancient DNA studies have used hybridization temperatures as low as 48 °C (Enk *et al.* 2014). We did not test the efficiency of different hybridization temperatures in our study; however, we propose that additional mammalian taxa not included in this study may be enriched with this same probe set by optimizing the hybridization temperature. Another option would be to redesign the MMA by adding more probes for the taxa of interest, if sequences for those variable regions are available. These potential modifications of the MMA can expand its application to an even larger diversity of mammals. Additionally, selection of samples that would perform the best during a multiplexed hybridization depends on the following factors; percent endogenous DNA, the divergence levels of the target samples, and the usage of dual-indexed adapters. We chose to pool together samples from more divergent taxa. This would prevent competition for the same probes by samples of closely related taxa. We also believe that pooling together samples from more distantly related taxa allows better detection of chimeric molecules.

Strategies to improve accuracy of multiplex genotyping

Based on visual and bioinformatic evidence, we determined that chimeric sequences were forming during postenrichment amplification. Dubbed 'jumping PCR', this phenomenon occurs when DNA polymerase jumps from one template to another during amplification and creates a continuous DNA strand from hybrid origins (Meyerhans *et al.* 1990; Pääbo *et al.* 1990; Odelberg *et al.* 1995; Lahr & Katz 2009). In a multiplexed PCR, incomplete primer extension followed by annealing of the incompletely extended DNA strand to a region of similarity in a different template might result in the formation of chimeras. This phenomenon appears to be more common in highly degraded samples (Pääbo *et al.* 1990). However, it has also been documented in high quality DNA (obtained from freshly preserved tissue samples) on HTS platforms (Edgar *et al.* 2011; Haas *et al.* 2011).

We identified chimeric reads as those containing sequences identifiable to two or more distantly related species. After careful examination of each enrichment, we determined that multiplexing samples with single indices combined with more cycles of PCR than recommended likely caused molecules from one species to 'jump' to another molecule and resulted in the observed chimeric reads. In fact, Kircher *et al.* (2012) estimated that jumping PCR generated 0.4% chimeric reads in their dual indexed experiments on two Neanderthal bones when captures were multiplexed, vs. 0.03% chimeric reads when they were singleplexed. They concluded that singleplex methods reduce chimera formation, and the inclusion of dual indices allow for the detection of chimeras, even when samples are singleplexed. Their data indicate that dual indexing samples before capture substantially decreases the possibility of errors due to jumping PCR and index false assignment. Therefore, we advocate dual indexing any samples that will be captured in multiplex reactions.

Another factor that may increase the rate of chimeric molecule formation is by exceeding the number of PCR cycles recommended for enrichment in the MYbaits protocol (14 cycles; MYcroarray™). Amplified captured libraries prepared from degraded samples may have unincorporated adapter dimer that need to be eliminated to increase sequencing efficiency of targeted regions. However, because we routinely experienced a 20–30% loss of DNA during the gel purification step (see Qiagen product specifications), we increased the number of cycles in the enrichment step to increase the amount of library concentration for sequencing. We attributed this increase in cycles, coupled with the use of single indices, to a higher than expected formation of chimeric sequences than if we had used dual indices and performed the recommended number of cycles.

Once we detected that jumping PCRs were occurring in our multiplexed libraries, we developed a pipeline of additional quality filters to remove suspected chimeric sequences (see Fig. 3). We tested this pipeline in silico by generating a data set of nonchimeric and chimeric molecules of varying source proportions and evaluated the ability of this pipeline to detect the chimeric molecules. In this test, we were able to identify 100% of chimeric molecules in which more than 60 bp of the total read length were from a second source. Chimeric molecules with <60 bp were more difficult to detect and may be classified as 'good' reads, and therefore higher coverage is necessary to prevent their inclusion into the final consensus sequences. However, we ran the pipeline on samples that had low, medium and high coverage and detected chimeric sequences in all cases.

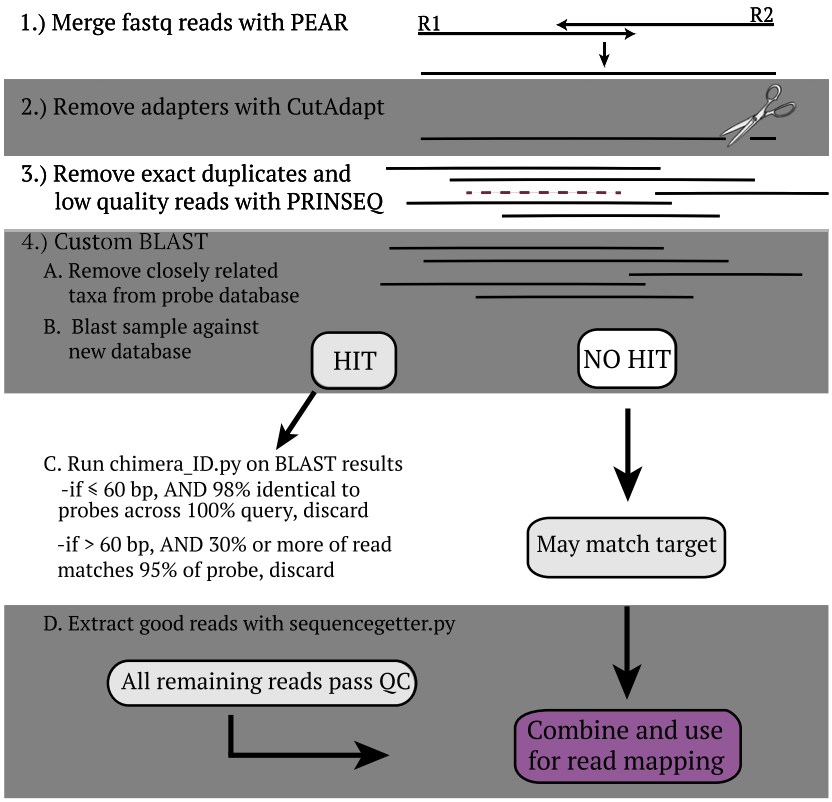

Fig. 3 A flow chart of the steps used in our chimera checking pipeline. The dashed line in step 3 represents a low quality read which would be discarded in quality filtering steps.

We compared the final mitogenome consensus sequences for low, medium and high coverage samples with and without running the chimera detection pipeline. After performing this pipeline, we found that the samples with low coverage (<10X) were affected by inclusion of chimeric molecules into the resulting consensus sequence (approximately 80 mismatches), or ~0.005% difference, between consensus sequences when reads were or were not run through the chimera detection pipeline. However, when samples had a moderate amount of coverage (between 30 and 50X) we found only 10 mismatches (~0.0006%) between consensus sequences when reads were not run through the chimera detection pipeline. At high coverage (over 100X) we observed only four mismatches (~0.0002%) incorporated across the mitogenomes. Based on these results, we suggest exercising caution when coverage is deemed 'low' or 'moderate' as this could lead to erroneous mitogenome sequences.

Samples with low or moderate coverage should be subjected to the chimera detection pipeline. While this pipeline conservatively removed some nonchimeric molecules, we would recommend the removal of additional nonchimeric molecules rather than risk including chimeric sequence into downstream consensus sequences. Because chimeric sequences can be detected more easily by using dual indexing (Kircher *et al.* 2012), we do not recommend the use single indices for multiplexed enrichments of degraded DNA samples.

Prevalence of nuclear copies of mitochondrial DNA in sequences

Overall, the inclusion of NuMTs into mitochondrial consensus sequences did not appear to be a major problem with our protocol. While NuMTs enrichment did occur during hybridization using the MMA, the application of multiple NuMT detection techniques allowed us to filter them from final inclusion in the mitogenome. These include translating protein coding genes for stop codons, visually searching for indels, and computing SNP frequencies. In the cases where SNPs were detected, which could indicate the presence of nuclear copies, either an ambiguity code or the more dominant nucleotide was called. Higher coverage will also prevent the inclusion of NuMTs by making the correct (mitochondrial) genotype easier to detect, especially if PCR cycles are limited to reduce clonal amplification of NuMT sequences and limit the proportion of NuMTs in the resulting data (although clonal sequences were removed with PRINSEQ).

Enrichment success

Enrichment success varied considerably between samples. Skin and museum 'tissue' material showed the

most variation, with samples ranging from almost 100% to ~1% of the mapped bases having at least 5X coverage (Table 2). We did not observe a correlation between quality and age of sample. Our results suggest that at least 15 000 quality filtered reads should be targeted when reconstructing novel mitogenomes. We found that 38 of the 63 individuals (60%) included in this study recovered at least 70% of the mitogenome. We decided to combine the 70% recovery threshold with at least 10X average coverage from mapping (in this case, with BWA) to determine if a sample was successful. This reduced the successful 38 to 32 individuals that had at least 70% of the mitogenome and at least 10X coverage (51%). In addition, it is possible that by relaxing hybridization stringency, more divergent taxa could more efficiently hybridize, but with the risk of incorporating nontarget molecules. A touchdown enrichment, where hybridization begins at 65 °C incubation and the annealing temperature slowly decreases over time (some studies report hybridization as low as 48 °C, Enk *et al.* 2014) should theoretically work in samples with greater level of divergence from the probe sequences.

Another recent study (Slon *et al.* 2015), based on similar methodology presented here reported a much lower enrichment success rate (zero of 42 samples). It should be noted that their study focused on much older samples, which were used to validate their method, but no authentic mitogenome sequences with the expected deamination patterns inherent in ancient bone samples were recovered. They also experienced lower enrichment success with the general mitochondrial tool developed in comparison to the cave bear-specific probe set used to validate their results. This should not be unexpected, as the number of cave bear sequences in the general mitochondrial tool is much less than that in a species-specific array.

Conclusions

The MMA probes successfully enriched mitogenomes for five orders of mammals from bone, museum 'tissue', spines and desiccated skin clips despite diluting probes and multiplexing reactions. Researchers looking to save time (in probe array design) and funds (covering multiple projects with the same array) can follow our design steps to design an array that is more specific for capturing molecular markers across taxa that meet the specifications of a particular lab group. We suggest multiplexing with dual indexed samples to enable chimera detection (as dual-indexed reads of chimeric origin would be discarded during demultiplexing, as the two indices would not match). We also recommend conducting multiple PCRs of the original enrichment product instead of adding additional PCR cycles to reduce chimera formation. After applying extra quality filters following chimera detection, we generated 32 successful mitogenomes. Our MMA probe set presents a cost effective alternative to generate complete mitogenomes from degraded samples for a large diversity of mammal taxa for projects spanning a range of interests.

Acknowledgements

We are grateful to the CCEG Genetics Laboratory for use of the facilities, and especially to Nancy Rotzel McInerney and Rob Fleischer for advise and support. Funding for this project came from two Peter Buck Postdoctoral Fellowships (to EEG, MMM) and the Smithsonian Small Grants Program (to KMH, JEM, EEG), both provided by the National Museum of Natural History, Smithsonian Institution; from the Systematics Research Fund Program provided by both the Systematics Association and the Linnean Society of London (to JEM, EEG, KMH); and from the Department of Biology of George Mason University (to MTRH). Museum specimens were sampled from a variety of collections, including; the National Museum of Natural History, Smithsonian Institution, Washington, DC (D. Lunde, E. Langan, N. Edmison, S. Peurach), The Field Museum, Chicago (B. Patterson, J. Phelps and W. Stanley), The American Museum of Natural History, New York (N. Simmons, R. Voss, E. Westwig, and N. Duncan), The Museum Zoologicum Bogoriense, Research Center for Biology, Indonesia Institute of Sciences, Cibinong, Indonesia (A. Achmadi), The Lee Kong Chian Natural History Museum, Singapore (K. Lim), Kansas University (R. Timm), Texas Tech University (H. Garner, K. MacDonald, and R. Baker), and The Naturalis Biodiversity Center, Leiden (C. Smeenk). Joel Callicrate contributed programming expertise and assisted with custom pipeline scripts. Gratitude is extended to Ross Furbush and the CCEG writing group for assistance with revisions of this manuscript (especially Loren Sackett for organizing). Jake Enk and Alison Devault provided valuable advice regarding troubleshooting and optimization of MYbaits kits and comments that helped improve the quality of our manuscript. Gratitude is extended to the three anonymous reviewers, and editor Travis Glenn whose comments greatly improved the quality of this manuscript.

References

Adler CJ, Dobney K, Weyrich LS *et al.* (2013) Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature Genetics*, **45**, 450–455, 455e1.

Ahlering MA, Eggert LS, Western D *et al.* (2012) Identifying source populations and genetic structure for savannah elephants in human-dominated landscapes and protected areas in the Kenya-Tanzania borderlands. *PLoS ONE*, **7**, e52288.

Asher RJ, Helgen KM (2010) Nomenclature and placental mammal phylogeny. *BMC Evolutionary Biology*, **10**, 102.

Bi K, Linderoth T, Vanderpool D *et al.* (2013) Unlocking the vault: next-generation museum population genomics. *Molecular Ecology*, **22**, 6018–6032.

Bozarth CA, Hailer F, Rockwood LL, Edwards CW, Maldonado JE (2011a) Coyote colonization of northern Virginia and admixture with Great Lakes wolves. *Journal of Mammalogy*, **92**, 1070–1080.

Bozarth CA, Lance SL, Civitello DJ, Glenn JL, Maldonado JE (2011b) Phylogeography of the gray fox (*Urocyon cinereoargenteus*) in the eastern United States. *Journal of Mammalogy*, **92**, 283–294.

Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.

Chevreux B, Wetter T, Suhai S (1999) Genome sequence assembly using trace signals and additional sequence information computer science and biology. *Proceedings of the German Conference on Bioinformatics*, **99**, 45–56.

Clark H, Hartl D (1997) *Principles of Population Genetics*. Sinauer, Sunderland, Massachusetts, USA.

Duchêne 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.

Edgar R, Haas B, Clemente J (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Enk JM, Devault AM, Kuch M *et al.* (2014) Ancient whole genome enrichment using baits built from modern DNA. *Molecular Biology and Evolution*, **31**, 1292–1294.

Faircloth BC, Glenn TC (2012) Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS ONE*, **7**, e42543.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Green RE, Krause J, Ptak SE *et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature*, **444**, 330–336.

Guschanski K, Krause J, Sawyer S *et al.* (2013) Next-generation museomics disentangles one of the largest primate radiations. *Systematic Biology*, **62**, 539–554.

Gutiérrez EE, Jansa SA, Voss RS (2010) Molecular systematics of mouse opossums (Didelphidae: Marmosa): assessing species limits using mitochondrial DNA sequences, with comments on phylogenetic relationships and biogeography. *American Museum Novitates*, **3692**, 1–22.

Gutiérrez EE, Anderson RP, Voss RS *et al.* (2014) Phylogeography of *Marmosa robinsoni*: insights into the biogeography of dry forests in northern South America. *Journal of Mammalogy*, **95**, 1175–1188.

Haas B, Gevers D, Earl A (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, **21**, 494–504.

Hancock-Hanser BL, Frey A, Leslie MS *et al.* (2013) Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Molecular Ecology Resources*, **13**, 254–268.

Hofman CA, Rick TC, Hawkins MTR *et al.* (2015) Mitochondrial genomes suggest rapid evolution of dwarf California Channel Islands foxes (*Urocyon littoralis*). *PLoS ONE*, **10**, e0118240.

Hofreiter M, Serre D, Poinar H, Kuch M, Pääbo S (2001) Ancient DNA. *Nature Reviews Genetics*, **2**, 353–359.

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics (Oxford, England)*, **26**, 680–682.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kircher M (2012) Analysis of high-throughput ancient DNA sequencing data. In: *Ancient DNA- Methods and Protocols* (ed. Shapiro B & Hofreiter M), pp. 197–228. Springer, New York.

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

Krause J, Fu Q, Good JM *et al.* (2010) The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, **464**, 894–897.

Lahr D, Katz L (2009) Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques*, **46**, 857–866.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Larsen RJ, Knapp MC, Genoways HH *et al.* (2012) Genetic diversity of neotropical Myotis (chiroptera: vespertilionidae) with an emphasis on South American species (D Steinke, Ed). *PLoS ONE*, **7**, e46578.

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

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lunter G, Goodson M (2010) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

Maddison WP (1997) Gene trees in species trees. *Systematic Biology*, **46**, 523–536.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.

Mason VC, Li G, Helgen KM, Murphy WJ (2011) Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Research*, **21**, 1695–1704.

McCormack JE, Faircloth BC (2013) Next-generation phylogenetics takes root. *Molecular Ecology*, **22**, 19–21.

McDonough MM, Sotero-Caio CG, Ferguson AW *et al.* (2013) Mitochondrial DNA and karyotypic data confirm the presence of *Mus indutus* and *Mus minutoides* (Mammalia, Rodentia, Muridae, Nannomys) in Botswana. *ZooKeys*, **359**, 35–51.

Meredith RW, Janečka JE, Gatesy J *et al.* (2011) Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science (New York, N.Y.)*, **334**, 521–524.

Meyerhans A, Vartanian J-P, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Research*, **18**, 1687–1691.

Miller W, Drautz DI, Janelka JE *et al.* (2009) The mitochondrial genome sequence of the Tasmanian Tiger (*Thylacinus cynocephalus*). *Genome Research*, **19**, 213–220.

Odelberg SJ, Weiss RB, Hata A, White R (1995) Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic Acids Research*, **23**, 2049–2057.

Pääbo S, Higuchi RG, Wilson AC (1989) Ancient DNA and the polymerase chain reaction. *Journal of Biological Chemistry*, **264**, 9709–9712.

Pääbo S, Irwin D, Wilson A (1990) DNA damage promotes jumping between templates during enzymatic amplification. *Journal of Biological Chemistry*, **265**, 4718–4721.

Pääbo S, Poinar H, Serre D *et al.* (2004) Genetic analyses from ancient DNA. *Annual Review of Genetics*, **38**, 645–679.

Pesole G, Gissi C, De Chirico A, Saccone C (1999) Nucleotide substitution rate of mammalian mitochondrial genomes. *Journal of Molecular Evolution*, **48**, 427–434.

Petrova TV, Zakharov ES, Samiya R, Abramson NI (2014) Phylogeography of the narrow-headed vole Lasiopodomys (Stenocranius) gregalis (Cricetidae, Rodentia) inferred from mitochondrial cytochrome *b* sequences: an echo of Pleistocene prosperity. *Journal of Zoological Systematics and Evolutionary Research*, **53**, 97–108.

Prüfer K, Racimo F, Patterson N *et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505**, 43–49.

Rogaev E, Moliaka Y (2006) Complete mitochondrial genome and phylogeny of Pleistocene mammoth *Mammuthus primigenius*. *PLoS Biology*, **4**, e73. doi: 10.1371/journal.pbio.0040073.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, **27**, 863–864.

Shapiro B, Hofreiter M (2012) *Ancient DNA: Methods and Protocols*. Springer, New York.

Siles L, Brooks DM, Aranibar H *et al.* (2013) A new species of Micronycteris (Chiroptera: Phyllostomidae) from Bolivia. *Journal of Mammalogy*, **94**, 881–896.

Slon V, Glocke I, Barkai R *et al.* (2015) Mammalian mitochondrial capture, a tool for rapid screening of DNA preservation in faunal and undiag-

nostic remains, and its application to Middle Pleistocene specimens from Qesem Cave (Israel). *Quaternary International*, 1–9, doi: 10.1016/j.quaint.2015.03.039.

Taberlet P, Luikart G (1999) Non-invasive genetic sampling and individual identification. *Biological Journal of the Linnean Society*, **68**, 41–55.

Taberlet P, Camarra J-J, Griffin S *et al.* (1997) Noninvasive genetic tracking of the endangered Pyrenean brown bear population. *Molecular Ecology*, **6**, 869–876.

Toews DPL, Brelsford A (2012) The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, **21**, 3907–3930.

Voss RS, Hubbard C, Jansa SA (2013) Phylogenetic relationships of New World porcupines (Rodentia, Erethizontidae): implications for taxonomy, morphological evolution, and biogeography. (American Museum novitates, no. 3769).

Wilson DE, Reeder DM (2005) *Mammal Species of the World: A Taxonomic and Geographic Reference, 2-Volume Set* (eds Wilson D & Reeder D). The Johns Hopkins University Press, Baltimore, Maryland, USA.

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)*, **30**, 614–620.

M.T.R.H. designed the array and performed the experiments. E.E.G., M.M.M. and M.T.N.T. assisted with laboratory experiments. T.E.C. and C.A.H. assisted with data analysis and troubleshooting. J.E.M. and K.M.H. assisted with experimental design and all authors participated in drafting this manuscript.

Data accessibility

All custom scripts, probe sequences and novel mitogenomes (including assembled and raw reads) have been deposited on Dryad at: http://dx.doi.org/10.5061/dryad.gq883. All resulting enrichment assemblies are also deposited on Dryad, and all raw reads have been uploaded to the GenBank SRA via the Bioproject SUB 1022917, and under the following Biosamples: SUB1022920, SUB1022922–SUB1022925, SUB1022927–SUB1022932, SUB1022935–SUB1022955, SUB1022957–SUB1022967, SUB 1022969–SUB1022973 and SUB1022975–SUB1022989.

Supporting Information

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

Table S1. A complete list of the species included in the MMA probe design.

Table S2. Additional metrics not included in Table 2.

Table S3. Number of SNPs detected in all samples by calculating the minimum variant frequency at 0.2, and represented by at least 5× read depth.

Table S4. A complete list of the 100 GenBank sequences tested for sequence similarity to the MMA probe set.