Received: 7 January 2021 | Revised: 9 May 2021 | Accepted: 14 May 2021

DOI: 10.1111/1755-0998.13433

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens

Nicolas Straube[1,2] | Mariana L. Lyra[3,4] | Johanna L. A. Paijmans[5] | Michaela Preick[5] | Nikolas Basler[5] | Johannes Penner[6,7] | Mark-Oliver Rödel[6] | Michael V. Westbury[8] | Célio F. B. Haddad[3] | Axel Barlow[5] | Michael Hofreiter[5]

[1]University Museum of Bergen, Bergen, Norway

[2]SNSB Bavarian State Collection of Zoology, München, Germany

[3]Departamento de Biodiversidade, Instituto de Biociências and Centro de Aquicultura (CAUNESP), Laboratório de Herpetologia, Universidade Estadual Paulista - UNESP, Rio Claro, SP, Brazil

[4]Zoological Institute, Braunschweig University of Technology, Braunschweig, Germany

[5]Department of Mathematics and Natural Sciences, Evolutionary Adaptive Genomics, Institute for Biochemistry and Biology, University of Potsdam, Potsdam, Germany

[6]Museum für Naturkunde– Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany

[7]Chair of Wildlife Ecology and Management, Albert Ludwigs University Freiburg, Freiburg, Germany

[8]Section for Evolutionary Genomics, The GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

Correspondence
Nicolas Straube, University Museum of Bergen, Allégaten 41, 5020 Bergen, Norway.
Email: nicolas.straube@uib.no

Mariana L. Lyra, Laboratório de Herpetologia, Departamento de Biodiversidade, Instituto de Biociências and Centro de Aquicultura (CAUNESP), Universidade Estadual Paulista - UNESP, Av. 24A, N 1515 Bela Vista, CEP 13506-900, Rio Claro, SP, Brazil.
Email: marillyra@gmail.com

Present address
Johanna L. A. Paijmans, Department of Zoology, Cambridge University, Cambridge, UK
Nikolas Basler, Department of Microbiology, Immunology, and Transplantation, Division of Clinical and Epidemiological Virology, Rega Institute for Medical Research, Leuven, Belgium
Axel Barlow, School of Science and Technology, Nottingham Trent University, Nottingham, UK

Abstract

Millions of scientific specimens are housed in museum collections, a large part of which are fluid preserved. The use of formaldehyde as fixative and subsequent storage in ethanol is especially common in ichthyology and herpetology. This type of preservation damages DNA and reduces the chance of successful retrieval of genetic data. We applied ancient DNA extraction and single stranded library construction protocols to a variety of vertebrate samples obtained from wet collections and of different ages. Our results show that almost all samples tested yielded endogenous DNA. Archival DNA extraction was successful across different tissue types as well as using small amounts of tissue. Conversion of archival DNA fragments into single-stranded libraries resulted in usable data even for samples with initially undetectable DNA amounts. Subsequent target capture approaches for mitochondrial DNA using homemade baits on a subset of 30 samples resulted in almost complete mitochondrial genome sequences in several instances. Thus, application of ancient DNA methodology makes wet collection specimens, including type material as well as rare, old or extinct species, accessible for genetic and genomic analyses. Our results, accompanied by detailed step-by-step protocols, are a large step forward to open the DNA archive of museum wet collections for scientific studies.

Straube and Lyra - Joint first authors. Barlow and Hofreiter - joint senior authors.

Funding information
Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Number: #2013/50741-7, #2017/2616-8 and #2018/15425-0; Deutsche Forschungsgemeinschaft, Grant/Award Number: 351649567; Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 431589/2016-0 and 306623/2018-8

KEYWORDS
ancient DNA, archival DNA, biological collection, formalin, single-stranded DNA library, target capture

1 | INTRODUCTION

Natural history museums and other biological collections worldwide house millions of specimens that document changes in biodiversity, are key references for species descriptions and identifications, and are often the only available resource for the reconstruction of the tree of life. Hence, they represent invaluable and irreplaceable assets for reconstructing patterns and processes of evolution across time and space (Habel et al., 2014; Kemp, 2015; Peacock et al., 2017; Wandeler et al., 2007; Yeates et al., 2016) and are unique archives for biodiversity (Rocha et al., 2014). These collections have played key roles in many scientific discoveries that changed our knowledge about the environment and our place in the natural world (Funk, 2018). During the past few decades, biological tissue collections and DNA banks have become the most important repository of genetic information as the molecular revolution continues to overhaul our understanding of the diversity of life. However, millions of preserved specimens in the world's natural history museums originate from times where tissue samples for DNA analyses were not routinely taken before preservation. Consequently, most of these have been excluded from genetic analyses for decades.

In recent years, advances in extraction protocols and next-generation sequencing methods have allowed us to not only obtain DNA sequence information, but also whole genomes from historical specimens (e.g., Burrell et al., 2015; Hykin et al., 2015; Li et al., 2015; Speidel et al., 2015; Sproul & Maddison, 2017). The application of these techniques allowed for obtaining DNA information from both type material and old or rare species from remote localities (e.g., Rancilhac et al., 2020). This opens up new perspectives for studies on evolution, ecology, taxonomy, phylogeography or conservation strategies (e.g., Evans et al., 2019; Turvey et al., 2019) and greatly enhances the relevance of natural history collections in modern biodiversity studies (Yeates et al., 2016).

While DNA sequencing of dry collection specimens (dried skin, pinned insect specimens, bird feathers, etc.) has been widely successful (Haran et al., 2018; Li et al., 2015; McCormack et al., 2016), wet collection specimens remain challenging for numerous reasons (Gilbert et al., 2007; McGuire et al., 2018; Ruane & Austin, 2017; Tang, 2006). The necessity of minimally invasive sampling to preserve morphological characters, especially in small specimens, and the usual lack of information on the fixation and preservation history are limiting factors. Moreover, fixation and preservation chemicals may cause DNA damage.

Until the end of the 19th century, a solution of ethanol often including denaturants and water was most commonly used for the preservation of specimens in wet collections (Simmons, 2014). This mode of preservation can be problematic for DNA (e.g., McGuire et al., 2018; Ruane & Austin, 2017) since the presence of free water accelerates DNA damage by hydrolysis (Lindahl, 1993). In addition, problems with DNA recovery are exacerbated if samples are initially fixed with diluted formalin (4% formaldehyde in aqueous solution). Accidentally discovered as a fixative in 1893 (Blum, 1893), formaldehyde quickly found its way into fixation protocols. In vertebrate collections it is common to fix specimens using formaldehyde at low concentration and subsequently preserve them in 70% ethanol for long-term storage (Neumann et al., 2010; Simmons, 2002, 2014). Alterations of DNA induced by formaldehyde fixation include disruption of base-pairing, promotion of denaturation, as well as cross-linking between DNA and proteins (Hoffman et al., 2015). Additionally, formaldehyde can react with the DNA bases adenine, guanine and cytosine, resulting in methylol adducts (Hoffman et al., 2015; Karlsen et al., 1994; Tang, 2006). Methylol adducts, in addition to other blocking lesions, inhibit techniques that employ amplification of DNA, including PCR, and consequently almost all currently available DNA sequencing techniques (Gilbert et al., 2007). Therefore, DNA extraction and sequencing from formaldehyde fixed wet-collection animal samples has proven to be difficult and has a high failure rate (Gilbert et al., 2007; Licht et al., 2012; Tang, 2006; Wu et al., 2002). An exception to this rule are formalin fixed and paraffin embedded (FFPE) clinical human samples, successfully analysed in multiple studies (e.g., Einaga et al., 2017; Lin et al., 2009; Paireder et al., 2013; Snow et al., 2014) even recovering whole genomes (e.g., Munchel et al., 2015; Robbe et al., 2018; Zar et al., 2019).

To date, obtaining genetic data from formaldehyde fixed and/or ethanol preserved museum specimens can best be described as "hit and miss". The application of ancient DNA methods seems a promising approach for extraction and sequencing of archival DNA from wet-collection specimens, as many of the challenges are similar, such as highly fragmented DNA and the presence of contaminants and inhibitors. Recent advances in ancient DNA research resulted in particularly recognised technical achievements allowing for sequencing DNA from a diversity of ancient samples preserved under different conditions (Barlow et al., 2016; Green et al., 2010; Gutiérrez-García et al., 2014; Meyer et al., 2012; Orlando et al., 2013; Prüfer et al., 2014; Sheng et al., 2019). So far, only few studies successfully

accessed DNA sequences from preserved specimens from wet collections, most by adopting ancient DNA methodologies for DNA extraction and/or library preparation (Evans et al., 2019; Gansauge et al., 2017; Hykin et al., 2015; Li et al., 2015, 2016; Lyra et al., 2020; Rancilhac et al., 2020; Ruane & Austin, 2017; Turvey et al., 2019). However, these successful studies on formaldehyde fixed samples obtained data from only one or few samples and may thus be rather taxon specific. Therefore, generalizations on the efficiency of the methods for wet collections, quantification of contamination and documenting characteristics of archival DNA are not possible with the available data.

Here, we tested the efficacy of published ancient DNA protocols applied to vertebrate wet collection specimens, including a wide range of different taxa and samples with poorly known fixation and preservation history. By developing a pipeline to obtain genetic information from these samples, we are taking an important step towards understanding the characteristics of wet collection archival DNA. Specifically, we (i) tested different ancient DNA extraction protocols for recovering archival DNA fragments, (ii) incorporated these fragments into single-stranded libraries, (iii) collected information on the properties and characteristics of archival DNA, and (iv) aimed to recover mitochondrial DNA using hybridisation capture.

2 | MATERIALS AND METHODS

2.1 | Samples

Samples and data sets analysed herein were combined by the authors from different projects with different scopes over a time range of three years (2017–2019) reflected in unequal sample sizes, different DNA extraction approaches and downstream analyses. Therefore, we adjusted our data analysis and interpretation of results accordingly. In total, we investigated 33 museum wet collection specimens of Euteleostei, Chondrichthyes, Lissamphibia, and Squamata, including several type specimens (Table 1). Specimens generally lacked detailed fixation histories. From these 33 specimens, we took 57 tissue samples ranging from 0.25 to 144 mg, comprising muscle (N = 45), brain (N = 1), teeth (N = 3), liver (N = 5), cartilage and bones (N = 3) (Table S1). Tissue sampling was carried out using sterile scalpels and tweezers to make microdissections or using punch biopsy instruments.

All laboratory procedures prior to PCR amplification were carried out in a dedicated clean laboratory at the University of Potsdam, Germany, following standard procedures (e.g., decontamination procedures for all materials and reagents, negative controls included during DNA extraction and library preparation, see Fulton & Shapiro, 2019). Prior to extraction, all samples were weighted and washed with 1 ml Qiagen PE or PBS buffer two times in an attempt to decrease the amount of formaldehyde, ethanol, and other potential inhibitors. A graphical summary of the laboratory pipeline and data analyses steps is provided in Appendix S1: Figure S1.

2.2 | DNA extraction

A widely adopted DNA extraction protocol in the field of ancient DNA research has been developed by Dabney, Knapp et al. (2013). This method is based on the binding of DNA to a silica membrane in the presence of a chaotropic salt (guanidine hydrochloride) buffer. An important aspect of this method is the use of an extension reservoir fitted to commercial silica spin columns allowing the ratio of binding buffer to sample to be increased (13:1), which enhances DNA recovery, in particular for short DNA fragments typical for ancient or archival samples. The original tissue lysis buffer described by Dabney, Knapp et al. (2013) was optimised for the digestion of subfossil bone. We therefore tested five alternative digestion methods to assess their suitability for tissues obtained from museum wet collection specimens, combined with the DNA purification method of Dabney, Knapp et al. (2013) in an attempt to maximise recovery of DNA from each sample. These were:

2.2.1 | Guanidine treatment (N = 34)

We incubated tissue samples in 1 ml of guanidinium thiocyanate buffer (5 M GuSCN, 50 mM Tris pH 8.0, 25 mM NaCl, 20 mM EDTA, 1% Tween 20, 1% 2-mercaptoethanol) adapted from the GuSCN-based buffer described by Rohland et al. (2004). Samples were incubated in buffer for ~18 hr rotating at room temperature. After centrifugation, we followed the procedure described in Dabney, Knapp et al. (2013), that is, we added 1 ml of the supernatant to 13 ml of binding buffer (5 M guanidine hydrochloride, 40% isopropanol, 0.05% Tween 20, 90 mM sodium acetate). DNA was purified using the MinElute silica spin column (Qiagen) and eluted two times in TET buffer (10 mM Tris-HCL, 1 mM EDTA, 0.05% Tween 20) for a total of 25 µl.

2.2.2 | Proteinase K re-digestion treatment (Re-Prot K; N = 11)

Since the guanidine treatment leaves a substantial pellet of undigested material, we investigated the potential for redigesting this pellet, to further enhance DNA recovery, using a standard lysis buffer containing proteinase K (100 mM NaCl, 10 mM Tris-Cl, 25 mM EDTA pH 8.0, 0.5% sodium dodecyl sulphate, 0.1 mg/ml proteinase K; adapted from Sambrook & Russell, 2001). Tissue lysis was carried out for ~18 hr at 37°C with rotation, and DNA purified as described for the guanidine treatment.

2.2.3 | Proteinase K treatment (N = 8)

Samples were subjected directly to proteinase K lysis buffer. For tissue digestion, we used ~18 hr of incubation with rotation at 37°C. After centrifugation, DNA was purified as in previous treatments. Two of the eight samples included in this treatment represent

TABLE 1 Specimens analysed and metadata; See Table S1 for details. *Approximate dates

Class	Species	Voucher ID	Collect. date	Extraction protocol	Sample ID	Tissue type	mtDNA capture
Chondrichthyes	*Etmopterus bullisi*	ZMH103111	1979	Guanidine	C_Ebul_10	Muscle	Yes
	Etmopterus granulosus	ZSM30786	1979	Prot K	C_Egra_37	Tooth	No
	Etmopterus granulosus	ZSM30795	1979	Guanidine	C_Egra_11	Bone/cartilage	No
	Etmopterus granulosus	ZSM37668	2007	Guanidine, Prot K	C_Egra_12,13,38	Muscle, brain, tooth	Yes
	Etmopterus granulosus	ZSM37670	1979	Guanidine, Prot K, Re-Prot K, heat-prot K, prot K–65	C_Egra_14, 40, 49, 54, 56	Muscle, tooth	Yes
	Etmopterus hillianus	ZMH121764	1981	Guanidine	C_Ehil_15	Muscle	Yes
	Etmopterus litvinovi	ZMH24994	1987	Guanidine	C_Elit_16	Muscle	Yes
	Etmopterus lucifer	ZSM37622	1979	Guanidine	C_Eluci_17	Muscle	Yes
	Etmopterus pusillus	NHM7852615	1903	Guanidine	C_Epus_18	Muscle	No
	Leucoraja sp.	ZMH26260	1988	Guanidine	C_Leu_20	Muscle	Yes
Euteleostei	*Chiloglanis niloticus*	ZSM43646	2008	Guanidine	E_Cnil_7	Muscle	No
	Gobio gobio	ZSM30674	2004	Guanidine	E_Ggob_19	Muscle	No
	Synodontis frontosus	ZSM35204	2007	Guanidine	E_Sfron_31	Muscle	No
Lissamphibia	*Allobates alagoanus*	MZUSP78193	1957	Guanidine, Re-Prot K	A_Aala_1, 43	Muscle	Yes
	Allobates capixaba	MZUSP76628	1964	Guanidine, Re-Prot K	A_Acap_2, 44	Muscle	Yes
	Allobates carioca	MZUSP73753	1967	Guanidine, Re-Prot K	A_Acar_3, 45	Muscle	Yes
	Allobates olfersioides	MZUSP75655	1975	Guanidine, Re-Prot K	A_Aolf_4, 46	Muscle	Yes
	Aplastodiscus musicus	EI7530	1982	Guanidine, Re-Prot K	A_Amus_5, 47	Muscle	Yes
	Boana claresignata	MNRJ54331	1953	Guanidine, Re-Prot K	A_Bcla_6, 48	Muscle	Yes
	Dendropsophus minutus	CFBH361	1987	Guanidine	A_Dmin_8, 9	Muscle, liver	Yes
	Paratelmatobius lutzii	MZUSP94627	1964	Guanidine, Re-Prot K	A_Plut_29	Liver	Yes
	Paratelmatobius mantiqueira	MZUSP15134	1953	Guanidine, Re-Prot K	A_Pman_30	Liver	Yes
	Thoropa lutzi	MNRJ23525	1945	Guanidine, Re-Prot K	A_Tlut_32	Muscle	Yes
	Thoropa petropolitana	MNRJ25686	1941	Guanidine, Re-Prot K	A_Tpet_33	Muscle	Yes
	Mantidactylus grandidieri	ZSM2632005	2005	Guanidine	A_Mgra_21, 22	Muscle, bone/cartilage	No

(Continuous)

TABLE 1 (Continued)

Class	Species	Voucher ID	Collect. date	Extraction protocol	Sample ID	Tissue type	mtDNA capture
Squamata	*Ophiophagus hannah*	ZMB15879	1895–1900*	Guanidine	S_Ohan_23	Muscle	No
	Ophiophagus hannah	ZMB2816	1860	Guanidine	S_Ohan_24	Muscle	No
	Ophiophagus hannah	ZMB29674	1920th*	Guanidine	S_Ohan_27	Muscle	No
	Ophiophagus hannah	ZMB33961	1930	Guanidine	S_Ohan_28	Muscle	No
	Ophiophagus hannah	ZMB4299	1860–1880*	Guanidine	S_Ohan_25	Muscle	No
	Ophiophagus hannah	ZMB5984	1860–1870*	Guanidine	S_Ohan_26	Muscle	No
	Uroplatus giganteus	ZSM21022007	2007	Guanidine	S_Ugig_34	Muscle	No
	Zonosaurus maximus	ZSM3542014	2014	Guanidine, Prot K, heat-protK, protK−65	S_Zmax_35, 36, 41, 42, 55, 57	Muscle, bone/cartilage	No

Note: Collection codes: CFBH: "Célio F. B. Haddad amphibian collection, Universidade Estadual Paulista, Rio Claro, SP, Brazil; EI: Eugenio Izecksohn Amphibian collection, Universidade Federal Rural, Seropédica, RJ, Brazil; ZMH: Zoologisches Museum Hamburg; MNRJ: Museu Nacional, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil; MZUSP: Museu de Zoologia da Universidade de São Paulo, São Paulo, SP, Brazil; ZMB: Museum für Naturkunde, Berlin, Germany; ZSM: Zoologische Staatssammlung München, Munich, Germany.

subsamples of the same specimens undergoing the proteinase K 65°C and Proteinase K 95°C treatments (see below).

2.2.4 | Proteinase K 65°C treatment (*N* = 2)

Two samples were subjected to proteinase K lysis with an increased incubation temperature of 65°C for ~18 hr. Increased incubation temperature has been suggested to have a positive effect on DNA yield by potentially reversing formaldehyde induced cross-links (Gilbert et al., 2007; Stiller et al., 2016). DNA purification was performed as described above.

2.2.5 | 95°C prior proteinase K treatment (*N* = 2)

Two samples were subjected to proteinase K lysis but with an initial heating for 15 min at 95°C in the extraction buffer prior to the addition of proteinase K. This was another test for increasing DNA yield by cross-link reversal (Gilbert et al., 2007). Samples were further processed as described for the other treatments.

Purified DNA from all extractions was quantified using a Qubit fluorometer with high sensitivity reagents, using 1 µl of DNA extract. This assay is able to detect DNA concentrations higher than 0.05 ng/µl. See Table S1 for details of which samples were extracted with which method, and Appendix S1: Supplementary Material 1 for detailed archival DNA extraction protocols.

We tested for correlations between total DNA yield and amount of tissue, total DNA yield and the age of specimens (i.e., time since the collection event) as well as average insert size (in bp) and age of specimens for the guanidine treated samples (*N* = 34). A MANOVA was performed to test for differences within the guanidine treated samples grouping for samples by DNA content (≥ and <0.05 ng/µl). We also tested for significant differences in total DNA yield between extracts obtained from the same specimens using the guanidine treatment, followed by Re-Prot K of the undigested pellet, using a two-sample *t* test. Comparisons between other treatments were done qualitatively due to low sample sizes. Correlation and *t* tests were performed in Past3 (Hammer et al., 2001).

2.3 | Single-stranded library preparation

Single-stranded DNA libraries have been shown to be more efficient in recovering molecules from highly degraded DNA compared to double-stranded libraries (Gansauge et al., 2017, 2020; Gansauge & Meyer, 2013; Meyer et al., 2012). This method processes both strands of the DNA fragments independently (Gansauge et al., 2020), minimizes the loss of DNA fragments with single-stranded breaks, allows conversion of very small fragments (shorter than 100 base pairs) with high efficiency, and is also not dependent on high input amounts of DNA (Gansauge & Meyer, 2013). Endogenous DNA from ancient or degraded samples was successfully recovered using single-stranded

libraries, for example, in Barlow et al. (2016), Gansauge and Meyer (2013), Meyer et al. (2012), Prüfer et al. (2014), and Stiller et al. (2016).

Considering that, we prepared dual-indexed single-stranded libraries from each extract following the protocol of Gansauge and Meyer (2013) with the modifications described in Basler et al. (2017). As the efficiency of single-strand ligation decreases with larger input amounts (Gansauge & Meyer, 2013) we used a maximum of 13 ng DNA as input for most libraries (N = 36), including dilution steps, if necessary. For N = 21 we used the complete DNA extract (exceeding 13 ng; Table S1).

Extracts were initially treated with uracil-DNA glycosylase, to remove deoxyuracils resulting from cytosine deamination, and Endonuclease VIII, to cleave abasic sites. Following heat denaturation, biotinylated adapter oligos were ligated to the 3' ends of the single stranded DNA fragments and immobilized on magnetic beads. In subsequent steps, the strand complementary to the original template is filled in incorporating the proximal sections of 5' and 3' Illumina adapter sequences, which serve as priming sites for dual indexing PCR. The optimal number of dual indexing PCR cycles was individually determined for each library by qPCR prior to amplification. The qPCR was carried out with three replicates per library in a 1:20 dilution adding 9 µl of mastermix (Thermofisher SYBR GREEN) to 1 µl of library. The cycling regime consisted of 2 min at 95°C, followed by variable numbers of cycles of 15 s denaturation at 95°C, 30 s annealing at 60°C and 1 min extension at 68°C. The amplified libraries were then purified using the QIAGEN MinElute PCR Purification kit and quantified using Qubit Fluorometer and Agilent Tapestation instruments.

Mean cycle threshold (Ct) values determined using qPCR were used as a measure of the relative conversion rate of DNA molecules into libraries. Specifically, we compared Ct values between libraries and between libraries and blanks. Considering samples with approximately the same amount of template, lower Ct values indicate a greater conversion of DNA molecules into the library, with a reduction of one PCR cycle corresponding, approximately, to a two-fold increase in conversion rate. Relative comparisons were only carried for libraries where the input DNA amount was ≤13 ng (N = 36 in total, N = 22 extracts from the guanidine treatment, N = 10 from the ReProt K, N = 3 from the Proteinase K, and N = 1 from the Proteinase K 65 treatments; Table S1). See Appendix S1: Supplementary Material 2 for the single-stranded library preparation protocol applied herein.

The composition of the libraries was assessed by shotgun-sequencing approximately one million 75 bp single-end reads using an Illumina Nextseq 500/550 sequencing platform, using 500/550 High Output v2.5 (75 cycles, Illumina 20024906) kits, following the procedure described in Paijmans et al. (2017).

2.4 | Analysis of DNA fragmentation and data recovery

We investigated the extent of DNA fragmentation in the archival samples by calculating the average library insert size. For that we trimmed the adapters using cutadapt v. 1.16 (Martin, 2011), with a minimum overlap length between read and adapter of 4 bp. Average insert sizes of the trimmed fragments (excluding untrimmed reads as well as reads that have a length of zero) were computed using standard bash utilities.

Advanced fragmentation of archival DNA reduces the overall sequence data yield compared to an equivalent, high quality modern sample. This occurs as a result of the removal of very short reads (typically <30 bp) which may map ambiguously to a reference genome (de Filippo et al., 2018), and because a large proportion of the remaining reads may be shorter than the read length used (75 bp in this case). We therefore investigated this relative reduction in data retrieval by calculating the proportion of usable data obtained from the museum specimens (including both target and contaminant DNA). Reads shorter than 30 bp were removed using cutadapt and duplicate reads (identical sequences) removed using Tally v. 14-020 (Davis et al., 2013). The usable data proportion ("data recovery per read") was then calculated by dividing the remaining bp by the starting 75 bp. In addition, if archival DNA extracts contain very small amounts of DNA, the resulting libraries can be of low complexity and the data will become increasingly saturated with duplicates at higher sequencing depths. We therefore also estimated library complexity, dividing the number of trimmed and non-duplicated reads by the total number of trimmed reads.

2.5 | Composition of archival DNA

A key parameter in the analysis of archival DNA is the ratio of target, or endogenous (i.e., representing the nuclear and mitochondrial genomes of the sequenced individual) to contaminant DNA. This ratio is inversely related to the cost of whole genome shotgun sequencing, and samples with extremely low target DNA content may only be suitable for hybridisation capture and other target enrichment approaches. In addition, the taxonomic composition of the contaminating DNA can be of key importance for both shotgun and target enrichment experiments. For example, contamination with closely related species may hinder target capture, while prokaryotic contamination may be eliminated more easily using mapping approaches.

To investigate the presence of target archival DNA and potential sources of contamination, we mapped the trimmed, de-duplicated shotgun reads against a selected set of reference genomes using FASTQSCREEN v0.13.0 (Wingett & Andrews, 2018; default parameters). The Burrows-Wheeler Alignment tool with seeding alignments with maximal exact matches (BWA-MEM) was specified as alignment algorithm (Li & Durbin, 2009).

Mapping-based estimates of endogenous DNA content are very sensitive to the reference genome used (Vieira et al., 2020), as with increasing phylogenetic distance, fewer and fewer reads map, giving the misleading impression that target DNA is absent (Prüfer et al., 2010). For most taxa investigated in this study, there are no closely related genomes available for improving target DNA detection estimates (except for the king cobra samples, see below). Therefore, we also used primary transcriptomic data of a more closely related species as reference when available (assembled RNA sequence data;

Table S1). Even though transcriptomic data does not represent the entire genome, primary transcriptomic data may represent as much as 80% of the genome, at least in humans (Dunham et al., 2012).

We considered that target archival DNA is present if we found reads mapping uniquely to the phylogenetically closest reference genome or transcriptome. Here, the usage of transcriptomic data was tested as alternative approach in case genomes of phylogenetically close species are unavailable as references. Custom configuration files were therefore created for our samples including a set of reference genomes scanned for all taxa as well as the phylogenetically closest and a set of possibly contaminating organisms (Appendix S1: Supplementary Material 3).

In order to further characterize the DNA content of extracts we used metaPhlAn 3.0 (Beghini et al., 2020; default parameters plus --add_viruses) for profiling the microbial DNA composition, including reference genomes of bacteria, archaea, viruses and microscopic eukaryotes.

2.6 | Precise quantification of target DNA in archival king cobra samples

The only taxon investigated in this study for which a conspecific reference genome is available is the king cobra, *Ophiophagus hannah* (Cantor 1836) (Vonk et al., 2013). This enabled the precise quantification of target and contaminant DNA in the six analysed archival king cobra samples. We estimated the proportion of target nuclear and mitochondrial DNA separately. For nuclear DNA, we first removed any scaffolds from the reference nuclear genome showing significant BLASTn hits to king cobra mitochondrial DNA. We then mapped ~8.5 million trimmed reads ≥30 bp from the modern king cobra individual used for genome assembly back to this reference genome using the BWA-ALN (Li & Durbin, 2009) alignment algorithm with default parameters. Samtools (Li et al., 2009) was then used to remove reads with low mapping quality (-q 30) and potential PCR duplicates (rmdup), and to determine the number of mapped reads (idxstats). This analysis provided an estimate of the maximum proportion of reads that can be mapped when using a high-quality modern sample, which we assumed to comprise 100% target nuclear DNA. We then mapped data from the archival samples and used the modern sample mapping proportion to convert the archival sample mapping proportion to an estimated proportion of target nuclear DNA. We also estimated the extent of cytosine deamination of the king cobra nuclear DNA using MAPDAMAGE 2.0 (Jónsson et al., 2013) with default settings.

For mitochondrial DNA, preliminary analyses indicated that the mitochondrial sequences were highly divergent, complicating mapping analysis using standard methods. We therefore utilised MITOBIM V. 1.9.1 (Hahn et al., 2013), which uses MIRA V.4.0.2 (Chevreux et al., 1999), to assemble the mitochondrial sequences using iterative mapping. The published king cobra mitochondrial sequence was used as the initial seed and the mismatch parameter was set to 3, which initial tests showed to produce the most complete assemblies. Iterations were run until no additional reads could be incorporated into the assembly. We then added the proportion of target mitochondrial DNA to the estimated proportion of target nuclear DNA and assumed that the remaining data fraction represented unknown contaminants. Cytosine deamination of the king cobra mitochondrial DNA was estimated with MAPDAMAGE 2.0 using the consensus sequences under default settings.

2.7 | Recovery of mitochondrial sequences using hybridisation capture

We investigated the potential for recovery of mitochondrial sequences from the archival specimens using hybridisation capture for $N = 30$ samples (Table S1). In-solution hybridisation capture enrichment was performed using home-made baits (e.g., Gonzalez-Fortes & Paijmans, 2019; Horn, 2012; Li et al., 2015; Maricic et al., 2010). To generate baits, we extracted DNA from high quality, modern tissues of species closely related to the archival specimens using commercial kits (Machery-Nagel blood and tissue DNA or Qiagen DNeasy Blood and Tissue extraction kits). Their mitochondrial DNA was then amplified using long-range PCR of two overlapping fragments of the mitochondrial genome, using either published or novel primers, with the latter designed in Geneious (Appendix S1: Table S2). The long-range PCR was carried out using Takara LA Taq DNA polymerase. The temperature profile comprised an initial denaturation and heat activation step at 94°C for 3 min; 35 PCR cycles including denaturation at 94°C for 15 s, annealing at 50°C–62°C for 30 s and elongation at 72°C for 10 min. Amplification success was verified using gel electrophoresis. The concentration of each amplicon was then measured using a Qubit Fluorometer using the broad range DNA kit. Amplicons for each species were pooled in equimolar ratios and sheared using a Covaris Sonicator to approximately 150 to 200 bp. Biotinylated DNA baits were prepared from the sheared product following protocols from Li et al. (2015) and Gonzalez-Fortes and Paijmans (2019) comprising blunt-end repair, adapter ligation, and purification with the MinElute PCR clean-up kit (Qiagen).

Hybridisation capture was performed based on Gonzalez-Fortes and Paijmans (2019). A maximum of 500 ng of target library was used for hybridisation. In a first step, blocking oligos and Human Cot1 DNA were hybridised to the library at 37°C, and after a denaturation step (95°C for 4 min), the bait library mix was incubated for 24 hr at 65°C. Thereafter, the captured product was immobilized with Streptavidin coated beads and washed. The captured libraries were amplified after evaluating the optimal number of amplification cycles as described in section 2.3. Captured libraries were subsequently purified using the MinElute PCR clean-up kit (QIAGEN®) and DNA quantified using a Qubit Fluorometer (high sensitivity DNA kit). The whole procedure was repeated to further increase target content (Li et al., 2013, 2015; Paijmans et al., 2016; Springer et al., 2015; Templeton et al., 2013). Libraries were then sequenced on either an Illumina Nextseq (single end, 75 bp read length) or Miniseq instrument (paired end, 250 bp read length). Libraries were pooled aiming for approximately 1 million sequencing reads per sample for libraries sequenced on the Nextseq instrument and 2 million reads on the Miniseq. Reads shorter than 30 bp and adapter sequences

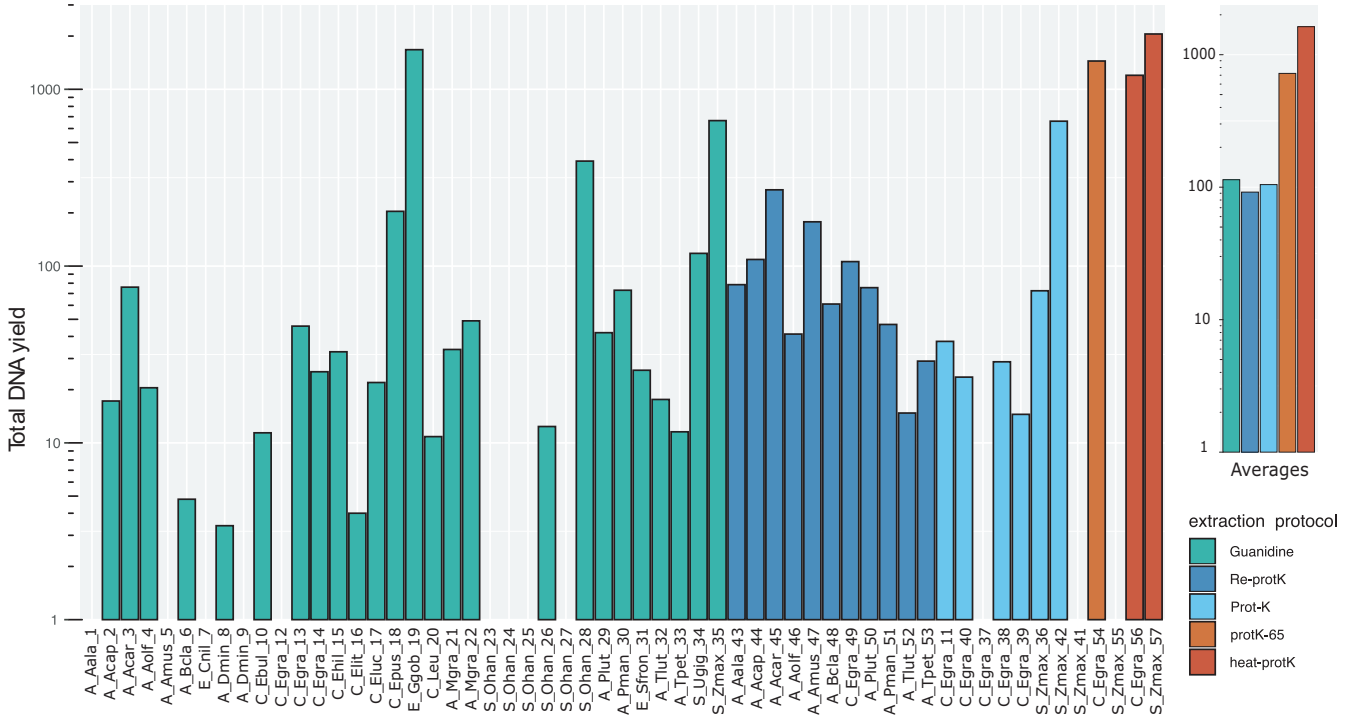

FIGURE 1 Amount of DNA extracted (DNA yield in ng). The Y-axis is in logarithmic scale. DNA of 12 samples was undetectable. Further sample information and sample metadata are given in Table S1

were trimmed using cutadapt as in section 2.3. To evaluate success of recovering mitochondrial sequences using hybridisation capture we used two strategies: we mapped reads to corresponding reference mitochondrial genomes of target species (Table S1) using BWA (Li & Durbin, 2009) and we assembled mitochondrial genomes from trimmed reads >30 bp using MITObim with the initial taxon-specific seeds listed in Table S1. The mismatch parameter was set to 3 and iterations were run until no additional reads could be incorporated into the assembly. In some instances, we lowered the k-mer size after initial experiments from the default of 31 bp to 21 bp for baiting.

For estimating cytosine deamination of the captured mitochondrial DNA, we used the mapped reads from the BWA analysis and corresponding consensus sequences assembled using MITObim as references in MAPDAMAGE 2.0. (Jónsson et al., 2013) with default settings. This step was performed for $N = 5$ samples, from which the consensus sequence allowed for damage pattern estimates. A detailed pipeline summary showing laboratory and data analysis is provided in Appendix S1: Figure S1.

3 | RESULTS

3.1 | Overall success of archival DNA extraction and library preparation

Application of ancient DNA extraction protocols to museum wet collection specimens produced DNA extracts with measurable DNA concentrations (≥0.05 ng/μl) for 45 of the 57 samples tested, ranging from 0.14 to 82.2 ng/μl (Figure 1, Table S1). The total yield of DNA in these extracts ranged from 3.5 to 2055 ng, which is equivalent in mass to approximately 1000 to 620,000 copies of the human genome. We found no correlation between tissue amount used for extraction and DNA yield (Figure 2a). Samples with approximately the same amount of initial tissue input showed both very high DNA yields (e.g., E_Ggob_19; 67 ng/μl) as well as DNA concentrations below the detection threshold (e.g., A_Amus_5; <0.05 ng/μl) (Figure 2a; Table S1). Similarly, specimen age showed no correlation with DNA yield (Figure 2b). Very low yields were observed across the complete range of specimen ages; however, some of the highest yields were obtained from specimens with less than 40 years of age or more than 90 years old (Figure 2b).

The Ct-values of quantitative PCRs suggest the incorporation of extracted DNA in the majority of sample libraries, including most samples for which DNA concentration was too low to be detected (Table S1). Average Ct-values of negative controls, both from extraction blanks and library blanks (extraction blank average 17.53, range 13.00–26.01 and library blank average 19.43, range 15.77–26.00) differed by 5.02 and 6.92 from the sample Ct-value average (12.51, range: 6.32–26.00), respectively. Only two of the 57 sample libraries showed higher Ct-values than the lowest Ct-values of extraction negative controls.

3.2 | Archival DNA fragmentation and data recovery

The average insert size incorporated in the libraries was 31 bp (range 20.02–48.23; Table S1), indicating advanced DNA fragmentation in all

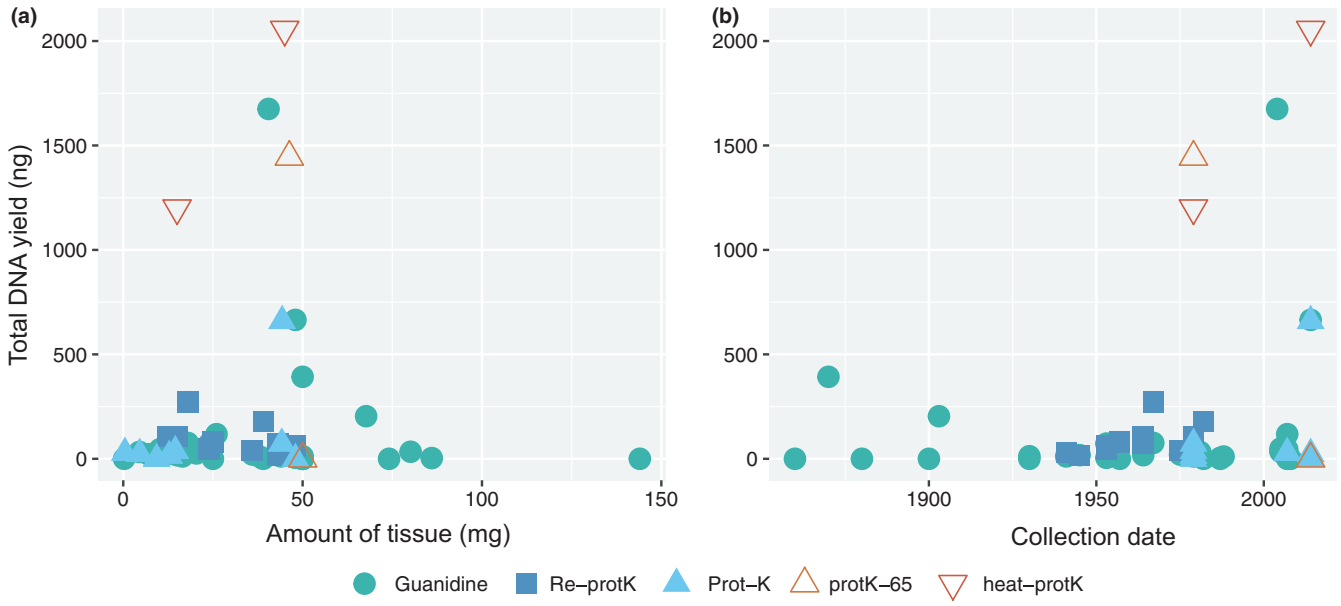

FIGURE 2 Relationship between (a) initial tissue input amount and DNA yield (total ng) and (b) collection date and DNA yield (total ng). The four king cobra samples with an imprecise collection date were included considering the most recent possible collection date

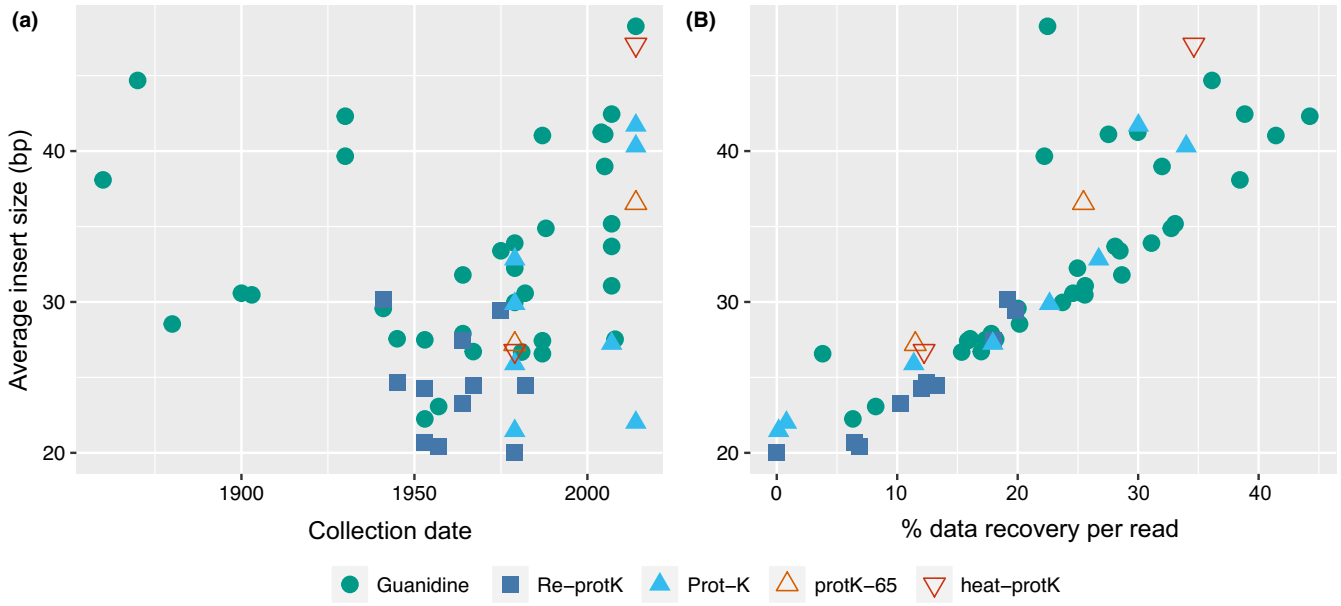

FIGURE 3 Relationship between (a) collection date and average insert size; and (b) % data recovery per read and average insert size. The four king cobra samples with an imprecise collection date were included considering the most recent possible collection date

samples. Additionally, electrophoresis using the Agilent Tapestation and high sensitivity DNA ScreenTape assay performed for some samples also showed highly fragmented DNA in the extracts.

Specimen age and average insert length are not correlated if we considered all the guanidine treatment samples (Figure 3a). The data recovery per read, that is, proportion of nucleotides per read useful for analyses, was on average 21% across all samples (range: 0.02% to 44.2%; Table S1) and was correlated with average insert size, after excluding reads smaller than 30 bp and duplicates (Figure 3b; Spearman's rho = 0.95 [$p < 0.000$]; Kendall's tau = 0.85 ($p < 0.000$)].

3.3 | DNA extraction methods

We successfully recovered DNA in all extraction methods. When comparing the guanidine treatment samples, we found no differences on data recovery per read or average Ct-values between samples that showed DNA concentration higher than 0.05 ng/µl and below this value ($N = 25$ and $N = 9$, respectively) (Appendix S1: Figure S2A, B). However, the average insert sizes were slightly shorter for extracts with low DNA concentration (31 bp vs. 33 bp; Appendix S1: Figure S2C).

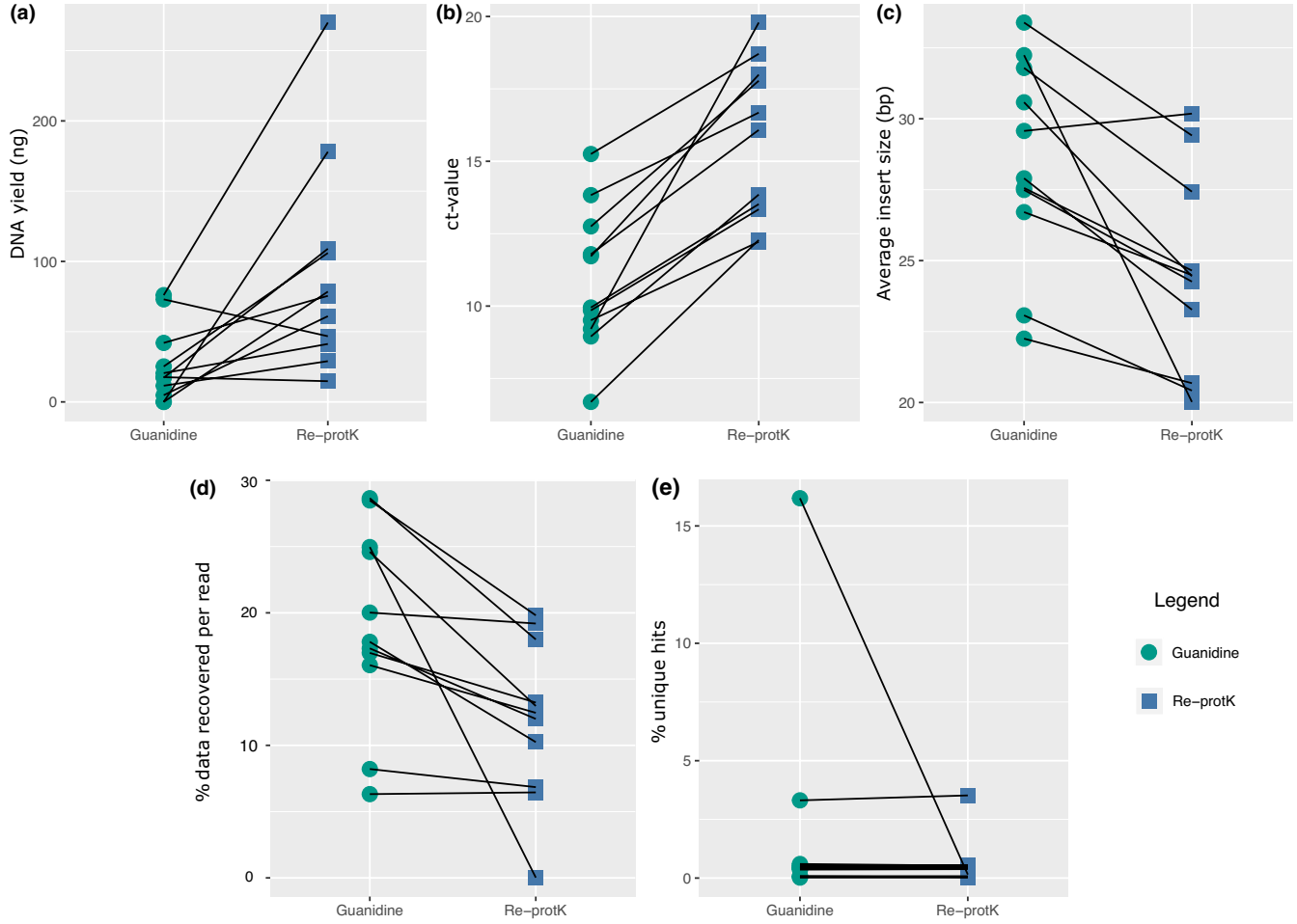

FIGURE 4 Comparison between samples extracted with the guanidine treatment and proteinase K redigestion treatment (re-Prot K) (*N* = 11 for each group). (a) Total amount of extracted DNA (ng), (b) Ct-values, (c) Average insert size (bp), (d) % data recovery per read, and (e) % of unique hits

The two-stage DNA extraction procedure of the guanidine treatment followed by Re-Prot K produced DNA extracts with notable differences. For all specimens, the Re-Prot K treatment produced significantly higherDNA yields than the initial guanidine treatment (Figure 4a; p = 0.03). The initial guanidine treatment extraction recovered on average 26.32 ng of DNA (range: below 0.05 to 76.00 ng); and Re-Prot K recovered on average 91.79 ng of DNA (range: 14.75 to 270.00 ng). However, when single-stranded libraries were prepared from these extracts, guanidine treatment consistently showed lower Ct values (guanidine treatment extraction: 6.69 to 15.25 [average 11.03] cycles; Re-Prot K: 12.22 to 19.79 [average 15.66] cycles, Figure 4b; p = 0.0003). This was even the case for the samples for which the DNA concentration of the incomplete digestion extract was too low to be measurable, necessitating the input DNA amount for library preparation to be substantially lower than for the Re-Prot K treatment. In addition, both data recovery per read and average insert sizes were higher for libraries prepared from the initial guanidine treatment (Figure 4c,d; data recovery per reads: guanidine = 19%; Re-Prot K = 12%; Average insert size: guanidine treatment = 28 bp; Re-Prot K = 24 bp). We found no differences in the relative endogenous content between these treatments (Figure 4e; p = 0.34), except for the re-extraction sample C_Egra_14, that showed a notable decrease in endogenous DNA content (C_Egra_49; Figure 4e).

Although robust interpretation of the results of proteinase K treatments with increased incubation temperatures or heating of samples prior to lysis are limited by small sample sizes, it is noteworthy that higher DNA yields were obtained in these extractions compared with the 37°C incubation treatment of the same samples (Figure 1, Figure 2; Appendix S1: Figure S3). For the sample *Etmopterus granulosus* ZSM37670, both extraction experiments applying increased temperature steps resulted in elevated DNA yields. The sample *Zonosaurus maximus* ZSM3542014 showed an elevated DNA yield only after the sample was heated to 95°C for 10 min. However, information on data recovery per read, average insert size and percentage of unique hits to reference genomes collected from the sequenced libraries did not show differences between extraction approaches (Appendix S1: Figure S3).

It is noteworthy that we were able to construct libraries from all 12 samples from which we did not detect DNA comprising the three different extraction approaches: guanidine treatment, proteinase K treatment and proteinase K 65°C treatment. These libraries showed

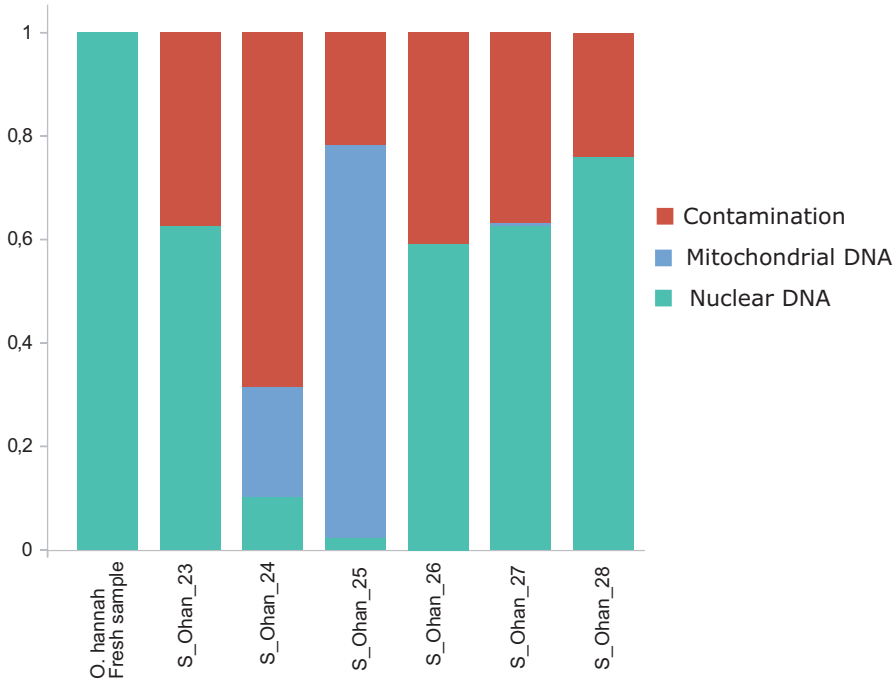

FIGURE 5 Proportions of nuclear, mitochondrial and putative contaminant DNA in six king cobra (*Ophiophagus hannah*) museum wet collection samples

relatively high endogenous DNA contents, with unique hits estimated up to 61.22% (Table S1).

3.4 | Composition of archival DNA

FastqScreen analysis identified between 0.02% and 70.17% of the reads mapping uniquely to the closest available reference genome or transcriptome for all samples, indicating the presence of target endogenous DNA (Table S1). This includes the libraries prepared from extracts with no detectable DNA, which showed a similar range of endogenous DNA content from 0.29% to 61.22%.

Using transcriptomic data as reference in the FastQscreen analyses, instead of complete genomes, allowed us to detect notably higher amounts of target DNA, that is, numbers of unique hits, in several instances. For example, when mapping lantern shark reads (*Etmopterus* spp.) to the whale shark genome, the average percentage of mapped reads is 0.44% (Table S1). Mapping the same reads to the transcriptome of *Etmopterus spinax*, the percentage of mapped reads rises on average to 9.61%. The importance of using a comparatively closely related reference for mapping is further supported by data from *Leucoraja* sp. (sample C_Leu_20; Table S1) where more than 34% of trimmed and deduplicated reads map to the congeneric *Leucoraja erinacea* genome (Table S1).

We found that human contamination estimates are generally low ranging from 0% to 2.30% of reads mapping uniquely to the human genome. The specimen with the maximum number of reads mapping to the human genome (2.30%; *Etmopterus pusillus*, C_Epus_18) was sampled at the surface of the specimen and not internally. We also found very low contents of identifiable microbial DNA according to MetaPhlAn3 analyses, ranging from 0% to 1% for bacteria and 0% to 2.47% for viruses (Table S1).

3.5 | Quantification of target DNA in archival king cobra samples

Estimated proportions of target endogenous nuclear DNA in the six archival king cobra specimens, the only species for which a conspecific reference genome is available, were highly variable, ranging from 2.9%–97.0% (Figure 5). Estimated contamination levels were similarly variable, from 2.9%–65.7%.

Estimates of mitochondrial DNA content indicated a low proportion of mitochondrial DNA (0.01%–0.5%) in four archival king cobra samples, broadly in line with the modern king cobra data (0.01%). However, data from two archival king cobra specimens showed remarkably high proportions of mitochondrial DNA: 21.1% and 76.1%, respectively. These mitochondrial contents are so high that for one specimen (S_Ohan_25), it was possible to assemble an ~50x coverage mitochondrial genome from only 20,000 shotgun sequencing reads. This massive mitochondrial enrichment was not observed for any other data set, even those generated using identical methodology (Table S1).

MapDamage2 analysis indicated comparatively low levels of cytosine deamination in the endogenous nuclear data of the king cobra samples, with a range of 1% to 8% of cytosine residues at the 5' fragment ends showing evidence of C to T substitutions. Damage estimates from the mitochondrial data show a range of 1% to 7% cytosine residues at the 5' ends (Appendix S1: Figure S4).

3.6 | Recovery of mitochondrial sequences using hybridisation capture

We recovered mitochondrial DNA for all tested samples (N = 30). On average 1,489,582 reads were available after trimming and

filtering the target capture data for further analyses. An average of 2.82% of reads mapped to the corresponding mitochondrial reference genome using BWA (range: 0.69% to 5.41%), which represented a 47-fold enrichment on average compared to the shotgun sequencing data. Subsequent reconstructions of parts of the mitochondrial genome from reads using MITObim were successful for 29 samples and resulted in an average number of 370,269 baited reads for iterative mapping (range: 1734 to 2,729,752). Mitochondrial RNAs were in general overrepresented, especially the fragments containing the 12S rRNA, interleaving tRNA-Val and 16S rRNA, while coding mtDNA genes were generally not as well covered or not recovered at all. DNA damage estimates from the mitochondrial data show a range of 2% to 5% cytosine residues at the 5' ends (Table S1).

4 | DISCUSSION

We were able to successfully recover endogenous DNA from 57 wet collection museum specimens with a 100% success rate. Thus, the results of our study show that ancient DNA extraction and subsequent single-stranded library preparation techniques are suitable for sequencing museum vertebrate wet collection material, including formalin preserved samples. Indeed, Stiller at al. (2016) have already documented a significant increase in library complexity comparing single-stranded libraries with double-stranded libraries built from formalin fixed paraffin embedded clinical biopsy samples (FFPE samples).

A lack of information regarding fixation and preservation is not a limiting factor in deciding whether or not to analyse a sample, as we were able to construct single-stranded libraries from all our samples and were able to detect target DNA in all single-stranded libraries. These results help to further open the DNA archive of millions of preserved specimens housed in wet collections both for mitochondrial as well as nuclear DNA analyses. Thus, the often-unique specimens of species groups that have largely been preserved in ethanol solutions (e.g., ichthyological and herpetological samples, besides numerous invertebrate specimens and some mammals such as bats and rodents) can be incorporated into genetic studies, similar to groups that are mostly preserved dried.

One important result of our study is the successful sequencing of target DNA in libraries constructed from DNA extracts with undetectable amounts of DNA using fluorometric quantification (Appendix S1: Figure S2). Discarding samples without measurable DNA amounts is a common strategy to pre-select samples with supposedly higher chance of yielding successful results (e.g., Hykin et al., 2015) but can prevent the inclusion and analysis of rare or precious samples, such as holotypes. Our results are promising, as sampling of wet collection specimens may also be limited in cases where the specimen is simply physically too small for obtaining multiple or large tissue samples without increasing the risk of damage to morphology. This is further supported by the absence of a correlation between the amount of tissue used for DNA extraction and the final DNA yield. Small tissue amounts may still provide enough DNA for successful library preparation.

The guanidine treatment was applied to most samples in this study and was widely successful with all types of tissues tested. The re-extraction of tissue remains from the guanidine extraction treatment (Re-Prot K treatment) resulted in higher DNA concentrations than in the initial extracts. We were expecting a decrease of Ct-values based on the assumption of a higher portion of DNA conversion into libraries; however, the opposite is the case. The Re-Prot K libraries consistently showed higher Ct-values compared to the initial round of extraction (Figure 4b). Experiments to specifically test why Re-Prot K treatments result in a lower conversion of DNA molecules into the library need to be conducted in the future, for example by constructing double stranded libraries from Re-Prot K samples and comparing insert sizes and data recovery per read from different libraries. We found no difference in the relative endogenous content recovered comparing the two extraction approaches (Figure 4e), which contrasts with results from ancient DNA studies suggesting that contaminating DNA may be preferentially removed in a first extraction, improving the overall endogenous content of the second extraction (Basler et al., 2017; Boessenkool et al., 2017).

The digestion of hard tissue using Proteinase K containing extraction buffer further worked for library constructions (Figures 1 and 2). However, the less invasive guanidine extraction treatment may be favourable for collection specimens, as it can preserve morphological information of hard tissue samples (Rohland et al., 2004) potentially avoiding physical damage to small specimens. Heating samples prior to extraction or increasing the digestion temperature also resulted in higher DNA concentrations (Figure 1) aligning with results of other studies (Gilbert et al., 2007; Stiller et al., 2016). Again, the high DNA concentration did not result in an improvement on data recovery per read, average insert size or relative endogenous content (Appendix S1: Figure S3). Based on our results, DNA extraction methods should not be rated simply on DNA yields.

Once samples underwent the complete workflow from DNA extraction to shotgun test-sequencing, we were able to compare the archival DNA characteristics of the wet-collection samples from this study to ancient DNA, which reveals both similarities and differences. The archival DNA fragmentation level is very high and similar to ancient DNA samples. This suggests that, even when using relatively short read lengths of 75 bp, the default expectation is that museum samples will incur more than double the sequencing expenditure of an equivalent high-quality modern sample, even when endogenous DNA levels are disregarded. The fragmentation of these wet collection samples appears to be much greater than DNA decay in fossil bones (Allentoft et al., 2012), or bones and dry skin samples of mammals stored in museums (Sawyer et al., 2012). Sawyer et al. (2012) analysed samples with a wide range of ages (from 18 to 2400 years) and reported on increased adenine frequencies at position −1 of mtDNA in samples younger than 100 years. We can confirm this observation in our samples (Table S1, Appendix S1: Figure S5). Sawyer

et al. (2012) also found the median length of mtDNA fragments to range between 44 bp and 170 bp. In contrast, none of our libraries had an average size larger than 49 bp, suggesting that fixation and preservation chemicals in wet collection samples probably speed up fragmentation. In some cases, high temperatures in collection facilities as well as the presence of water in storage containers probably further fuel DNA fragmentation. Storage containers represent (almost) closed systems, in which chemical reactions persist. Ethanol is a lipid solvent and lipids are extracted from wet-collection specimens, which allows for fatty acid formation. The increased acidity leads to further specimen tissue damage (Simmons, 2014) probably also affecting its DNA. In addition, the Uracil-DNA glycosylase and endonuclease VIII treatment results in further, albeit probably limited (Briggs et al., 2010) fragmentation and underestimation of deamination.

The *O. hannah* archival nuclear DNA samples analysed here show only weak increases of C to T substitutions at the 5′ ends (Appendix S1: Figure S4). Damage patterns estimated for mitochondrial archival DNA using reads from the king cobra samples and target enriched libraries show partially higher levels of deamination, but not reaching levels reported in some ancient DNA studies which reach up to 60% (e.g., Dabney et al., 2013; Fortes et al., 2016; Rizzi et al., 2012). Ancient DNA characteristics such as short read length and terminal cytosine deamination are often used to distinguish ancient DNA sequences from present-day DNA contamination (Peyrégne & Prüfer, 2020). Based on our results, damage caused by cytosine deamination may not be useful to distinguish archival DNA from modern DNA. However, we cannot rule out that the observed patterns for the nuclear DNA of the *O. hannah* samples are an effect of the enzymatic treatment during single-stranded library construction (Meyer et al., 2012). Further analyses, including deep sequencing, no enzymatic treatment of DNA before construction of single-stranded libraries and sequencing of further wet-collection specimens with a conspecific reference genome, are required to test for the range of terminal cytosine deamination in wet collection samples.

We could detect endogenous DNA in all libraries, but the percentage of mapping reads was low for a substantial number of our samples. Using the endogenous DNA content of a sample as a quality measure is a standard practice in ancient DNA studies (e.g., Meyer et al., 2012; Orlando et al., 2013; Skoglund et al., 2014; Westbury et al., 2017). However, mapping efficiency decreases fast with increasing genetic distances (Vieira et al., 2020) and using phylogenetically close reference genomes is important for estimating this parameter precisely (Prüfer et al., 2010). Sequenced vertebrate genomes are distributed unevenly across the vertebrate tree of life, with numerous mammalian genomes being available, but far fewer non-mammalian genomes, especially relative to the number of extant, described species. We tried to mitigate this problem by using transcriptomic data from phylogenetically closer species for taxa analysed here. Our results, which showed on average seven times more reads mapping to the phylogenetically closest transcriptome compared to the closest available genome, show that transcriptomic data is indeed informative. This result also emphasizes that the endogenous content is a relative concept, strongly influenced by the reference used for mapping. Thus, the assessment of sample quality in terms of mappable data, especially when comparing different species mapped to highly divergent references, should be interpreted with care.

As mapping gets increasingly difficult with evolutionary distance between target species and the reference genome, a solution to improve mapping efficiency could be estimating optimal mapping parameters in the light of different reference genomes. For example, the mapping parameters may be estimated using algorithms as in TAPAS tools (Taron et al., 2018) for reference genomes spanning a range of target species. It is also possible to explore different in silico strategies such as the use of hybrid genomes (e.g., Vieira et al., 2020). At the same time, based on our results, sequencing the transcriptome, or performing genome skimming of a more closely related species from which high-quality DNA is available may be a cost-efficient alternative if the goal is to detect relative endogenous DNA. However, it is important to take into account that short reads can result in large numbers of unassigned reads, that is, underprediction of endogenous content, regardless of the reference used (Huson et al., 2007).

As mitochondrial DNA sequence data is of large interest for biodiversity studies, we tested its presence in wet-collection specimens. Outstanding are the two king cobra samples showing mitochondrial enrichment. Its causes remain enigmatic, however, the high amount of mtDNA allows for reconstructing the full mitochondrial genome from a comparatively low number of reads. As two of the *O. hannah* samples show exceptionally high (up to 70% of mapped reads) mitochondrial DNA contents relative to expected amounts (0.01%–0.1% mapped reads), the assessment of mitochondrial DNA amounts in single-stranded libraries from museum samples based on a limited number of reads is crucial for evaluating the necessity of subsequent costly and labour-intensive target capture or to guide the decision of whole genome sequencing.

The target capture approaches of mitochondrial DNA reveal high copy numbers of 12S and 16S rRNAs as well as tRNAs. These are very conserved regions and may have been preferentially captured by the use of phylogenetically divergent baits (Paijmans et al., 2020). Nevertheless, fragments of the ribosomal genes are commonly used for phylogenetic inferences of vertebrates in non-model organisms (e.g., Frost et al., 2006; Iglésias et al., 2005; Straube et al., 2010) and are often the only available reference sequences. The availability of these fragments may guide the taxonomic analyses of museum samples in the future, as DNA fragments available in high copy numbers may be easier to sequence and assemble from highly fragmented archival DNA of wet-collection specimens and are also used in genome skimming analyses (e.g., Grandjean et al., 2017; Richter et al., 2015).

More recently, Lyra et al. (2020) as well as Rancilhac et al. (2020) successfully applied the approach discussed herein for taxonomic purposes. Their results allowed for clarification of complex taxonomic issues and resulted in the description of several amphibian

species new to science. This demonstrates not only the importance of scientific collections and deposited type material as genetic resources but further the applicability of the methods discussed here.

5 | CONCLUSIONS

We were able to convert extracted DNA into single-stranded libraries for all tested samples notably increasing the sample size of analysed wet collection specimens. To our knowledge, only few studies successfully applied single-stranded library construction approaches to formaldehyde fixed tissues of non-human museum wet collections specimens until early 2020 (e.g., Evans et al., 2019; Gansauge et al., 2017; Lyra et al., 2020; Rancilhac et al., 2020). We demonstrate that formaldehyde fixed wet collection samples with an unknown fixation history of different ages can be sequenced. This implies that DNA extraction using the guanidine treatment followed by single-stranded library construction is appropriate for sequencing DNA of wet collection specimens and can be performed even if DNA cannot be measured using fluorometric quantification (e.g., Qubit) in the extract. Additionally, we note that while a low Ct-value is an indicator for successful incorporation of DNA in the ss-library, only test sequencing will allow for the accurate measurement of endogenous DNA content. Our study is a large step forward in accessing archival DNA of vertebrate wet collection specimens. Based on our results, we suggest testing suitability of our pipeline on other wet collection specimens such as invertebrate samples.

ACKNOWLEDGEMENTS

We would like to express our sincere thanks to colleagues and biological collection curators for providing samples: Simon Weigmann (ZMH), Ralf Thiel (ZMH), Irina Eidus (ZMH), Ulrich Schliewen (ZSM), Dirk Neumann (ZSM), Frank Glaw (ZSM), Michael Franzen (ZSM), Timo Moritz (DMM), Gavin Naylor (FMNH), José Pombal Jr (MNRJ), Taran Grant (MZUSP), and Frank Tillack (ZMB). Further, we thank all the helping hands of the Evolutionary Adaptive Genomics group at the University of Potsdam: Stefanie Hartmann, Elisabeth Hempel, Federica Alberti and Remco Folkertsma. Miguel Vences (TUB) is thanked for providing transcriptomic reference data. We thank Délio Baêta and Julian Faivovich for suggestions on initial experiment design and Ariadne Sabbag and Andrés Brunetti for help with figure drafts. Leon Hilgers (ZFMK) is thanked for advice on transcriptomic data. We would further like to thank Gert Wörheide (LMU), Dirk Erpenbeck (LMU) and Sergio Vargas (LMU) for support with sequencing. Two anonymous reviewers are thanked for constructive criricism. This work was funded by: the German Research Foundation (DFG; project number 351649567 to NS and MH within the DFG SPP 1991 "Taxon-Omics"); São Paulo Research Foundation - FAPESP (grant #2013/50741-7, #2017/26162-8 and #2018/15425-0); and Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (Universal 431589/2016-0, research fellowship 306623/2018-8).

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

AUTHOR CONTRIBUTIONS

Nicolas Straube, Mariana L. Lyra, Célio F. B. Haddad, Michael Hofreiter and Axel Barlow designed the study. Johannes Penner and Mark-Oliver Rödel provided additional materials and samples. Laboratory work was carried out by Nicolas Straube, Mariana L. Lyra, Michaela Preick, Michael V. Westbury and Nikolas Basler based on protocols adapted by Johanna L. A. Paijmans and Axel Barlow. Data analysis was performed by Nicolas Straube and Mariana L. Lyra. Nicolas Straube, Mariana L. Lyra, Johanna L. A. Paijmans, Michael Hofreiter and Axel Barlow wrote the manuscript with input from all authors. All authors read and approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

Data presented in the manuscript are available in Table S1 and Appendix S1.

ORCID

Nicolas Straube https://orcid.org/0000-0001-7047-1084
Mariana L. Lyra https://orcid.org/0000-0002-7863-4965
Johanna L. A. Paijmans https://orcid.org/0000-0002-1938-7052
Michael V. Westbury https://orcid.org/0000-0003-0478-3930

REFERENCES

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., Campos, P. F., Samaniego, J. A., Gilbert, M. T. P., Willerslev, E., Zhang, G., Scofield, R. P., Holdaway, R. N., & Bunce, M. (2012). The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B*, *279*(1748), 4724–4733. https://doi.org/10.1098/rspb.2012.1745

Barlow, A., Fortes, G. G., Dalén, L., Pinhasi, R., Gasparyan, B., Rabeder, G., & Hofreiter, M. (2016). Massive Influence of DNA Isolation and Library Preparation Approaches on Palaeogenomic Sequencing Data. *bioRxiv*, 75911. https://doi.org/10.1101/075911

Basler, N., Xenikoudakis, G., Westbury, M. V., Song, L., Sheng, G., & Barlow, A. (2017). Reduction of the contaminant fraction of DNA obtained from an ancient giant panda bone. *BMC Research Notes*, *10*(1), 754. https://doi.org/10.1186/s13104-017-3061-3

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Segata, N. (2020). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *bioRxiv*. , https://doi.org/10.1101/2020.11.19.388223

Blum, F. (1893). Der formaldehyd als antisepticum. *Münchner Medicinische Wochenschrift*, *40*(32), 601–602.

Boessenkool, S., Hanghøj, K., Nistelberger, H. M., Der Sarkissian, C., Gondek, A. T., Orlando, L., Barrett, J. H., & Star, B. (2017). Combining bleach and mild predigestion improves ancient DNA recovery from bones. *Molecular Ecology Resources*, *17*, 742–751. https://doi.org/10.1111/1755-0998.12623

Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., & Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research*, *38*(6), e87. https://doi.org/10.1093/nar/gkp1163

Burrell, A. S., Disotell, T. R., & Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *Journal of Human Evolution*, *79*, 35–44. https://doi.org/10.1016/j.jhevol.2014.10.015

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 (GCB), 99*, 45–56.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., Garcia, N., Paabo, S., Arsuaga, J.-L., & Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences USA*, *110*(39), 15758–15763. https://doi.org/10.1073/pnas.1314445110

Dabney, J., Meyer, M., & Pääbo, S. (2013). *Cold Spring Harbor Perspectives in Biology*, *5*(7), a012567. https://doi.org/10.1101/cshperspect.a012567.

Davis, M. P. A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., & Enright, A. J. (2013). Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods*, *63*, 41–49. https://doi.org/10.1016/j.ymeth.2013.06.027

de Filippo, C., Meyer, M., & Prüfer, K. (2018). Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences. *BMC Biology*, *16*, 121. https://doi.org/10.1186/s12915-018-0581-9

Dunham, I., Kundaje, A., & The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*, 57–74. https://doi.org/10.1038/nature11247

Einaga, N., Yoshida, A., Noda, H., Suemitsu, M., Nakayama, Y., Sakurada, A., Kawaji, Y., Yamaguchi, H., Sasaki, Y., Tokino, T., & Esumi, M. (2017). Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation. *PLoS One*, *12*(5), e0176280. https://doi.org/10.1371/journal.pone.0176280

Evans, B. J., Gansauge, M.-T., Stanley, E. L., Furman, B. L. S., Cauret, C. M. S., Ofori-Boateng, C., Gvoždík, V., Streicher, J. W., Greenbaum, E., Tinsley, R. C., Meyer, M., & Blackburn, D. C. (2019). *Xenopus fraseri*: Mr. Fraser, where did your frog come from? *PLoS One*, *14*(9), e0220892. https://doi.org/10.1371/journal.pone.0220892

Fortes, G. G., Grandal-d'Anglade, A., Kolbe, B., Fernandes, D., Meleg, I. N., García-Vázquez, A., Pinto-Llona, A. C., Constantin, S., de Torres, T. J., Ortiz, J. E., Frischauf, C., Rabeder, G., Hofreiter, M., & Barlow, A. (2016). Ancient DNA reveals differences in behaviour and sociality between brown bears and extinct cave bears. *Molecular Ecology*, *25*, 4907–4918. https://doi.org/10.1111/mec.13800

Frost, D. R., Grant, T., Faivovich, J., Bain, R. H., Haas, A., Haddad, C. F. B., & Wheeler, W. C. (2006). The amphibian tree of life. *Bulletin of the American Museum of Natural History*, *297*, 1–370.

Fulton, T. L., & Shapiro, B. (2019). Setting Up an Ancient DNA Laboratory. In B. Shapiro, A. Barlow, P. Heintzman, M. Hofreiter, J. Paijmans, & A. Soares (Eds.), *Ancient DNA: Methods in Molecular Biology* (pp. 1–13). Humana Press. https://doi.org/10.1007/978-1-4939-9176-1_1

Funk, V. A. (2018). Collections-based science in the 21st century. *Journal of Systematics and Evolution*, *56*, 175–193. https://doi.org/10.1111/jse.12315

Gansauge, M.-T., Aximu-Petri, A., Nagel, S., & Meyer, M. (2020). Manual and automated preparation of single-stranded DNA libraries for the sequencing of DNA from ancient biological remains and other sources of highly degraded DNA. *Nature Protocols*, *15*, 2279–2300. https://doi.org/10.1038/s41596-020-0338-0

Gansauge, M. T., Gerber, T., Glocke, I., Korlevic, P., Lippik, L., Nagel, S., & Meyer, M. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Research*, *45*(10), e79. https://doi.org/10.1093/nar/gkx033

Gansauge, M. T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, *8*(4), 737–748. https://doi.org/10.1038/nprot.2013.038

Gilbert, M. T., Haselkorn, T., Bunce, M., Sanchez, J. J., Lucas, S. B., Jewell, L. D., & Worobey, M. (2007). The isolation of nucleic acids from fixed, paraffin-embedded tissues- which methods are useful when? *PLoS One*, *2*(6), 3537. https://doi.org/10.1371/journal.pone.0000537

Gonzalez-Fortes, G., & Paijmans, J. L. A. (2019). Whole-genome capture of ancient DNA using homemade baits. In B. Shapiro, A. Barlow, P. Heintzman, M. Hofreiter, J. Paijmans, & A. Soares (Eds.), *Ancient DNA: Methods in molecular biology* (pp. 93–105). Humana Press. https://doi.org/10.1007/978-1-4939-9176-1_11

Grandjean, F., Tan, M. H., Gan, H. M., Lee, Y. P., Kawai, T., Distefano, R. J., Blaha, M., Roles, A. J., & Austin, C. M. (2017). Rapid recovery of nuclear and mitochondrial genes by genome skimming from northern hemisphere freshwater crayfish. *Zoologica Scripta*, *46*(6), 718–728. https://doi.org/10.1111/zsc.12247

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., & Pääbo, S. (2010). A draft sequence of the Neanderthal genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Gutiérrez-García, T. A., Vázquez-Domínguez, E., Arroyo-Cabrales, J., Kuch, M., Enk, J., King, C., & Poinar, H. N. (2014). Ancient DNA and the tropics: A rodent's tale. *Biology Letters*, *10*(6), 20140224. https://doi.org/10.1098/rsbl.2014.0224

Habel, J. C., Husemann, M., Finger, A., Danley, P. D., & Zachos, F. E. (2014). The relevance of time series in molecular ecology and conservation biology. *Biological Reviews*, *89*, 484–492. https://doi.org/10.1111/brv.12068

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research*, *41*(13), e129. https://doi.org/10.1093/nar/gkt371

Hammer, Ø., Harper, D. A. T., & Ryan, P. D. (2001). PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*, *4*(1), 9.

Haran, J., Delvare, G., Vayssieres, J.-F., Benoit, L., Cruaud, P., Rasplus, J.-Y., & Cruaud, A. (2018). Increasing the utility of barcode databases through high-throughput sequencing of amplicons from dried museum specimens, an example on parasitic Hymenoptera (Braconidae). *Biological Control*, *122*, 93–100. https://doi.org/10.1016/j.biocontrol.2018.04.001

Hoffman, E. A., Frey, B. L., Smith, L. M., & Auble, D. T. (2015). Formaldehyde cross linking: A tool for the study of chromatin complexes. *The Journal of Biological Chemistry*, *290*(44), 26404–26411. https://doi.org/10.1074/jbc.R115.651679

Horn, S. (2012). Target enrichment via DNA hybridization capture. In B. Shapiro, & M. Hofreiter (Eds.), *Ancient DNA: Methods in molecular biology (methods and protocols)* (pp. 177–188). Humana Press.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/gr.5969107

Hykin, S. M., Bi, K., & McGuire, J. A. (2015). Fixing formalin: A method to recover genomic-scale DNA sequence data from formalin-fixed museum specimens using high-throughput sequencing. *PLoS One*, *10*(10), e0141579. https://doi.org/10.1371/journal.pone.0141579

Iglésias, S. P., Lecointre, G., & Sellos, D. Y. (2005). Extensive paraphylies within sharks of the order Carcharhiniformes inferred from nuclear and mitochondrial genes. *Molecular Phylogenetics and Evolution*, *34*(3), 569–583. https://doi.org/10.1016/j.ympev.2004.10.022

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, *29*, 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Karlsen, F., Kalantari, M., Chitemerere, M., Johansson, B., & Hagmar, B. (1994). Modifications of human and viral deoxyribonucleic acid by formaldehyde fixation. *Laboratory Investigation*, *71*, 504–611.

Kemp, C. (2015). Museums: The endangered dead. *Nature*, *518*(7539), 292–294. https://doi.org/10.1038/518292a

Li, C., Corrigan, S., Yang, L., Straube, N., Harris, M., Hofreiter, M., White, W. T., & Naylor, G. J. P. (2015). DNA capture reveals transoceanic gene flow in endangered river sharks. *Proceedings of the National Academy of Sciences*, *112*(43), 13302–13307. https://doi.org/10.1073/pnas.1508735112

Li, C., Hofreiter, M., Straube, N., Corrigan, S., & Naylor, G. J. P. (2013). Capturing protein-coding genes across highly divergent species. *BioTechniques*, *54*, 321–326. https://doi.org/10.2144/000114039

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Kuang, T., & Li, C. (2016). Determining mitochondrial genome sequences from formalin fixed paddlefish (*Polyodon spathula*) samples. *Journal of Shanghai Ocean University*, *25*(5), 661–669.

Licht, M., Schmuecker, K., Huelsken, T., Hanel, R., Bartsch, P., & Paeckert, M. (2012). Contribution to the molecular phylogenetic analysis of extant holocephalan fishes (Holocephali, Chimaeriformes). *Organisms Diversity and Evolution*, *12*(4), 421–432. https://doi.org/10.1007/s13127-011-0071-1

Lin, J., Kennedy, S. H., Svarovsky, T., Rogers, J., Kemnitz, J. W., Xu, A., & Zondervan, K. T. (2009). High-quality genomic DNA extraction from formalin-fixed and paraffin-embedded samples deparaffinized using mineral oil. *Analytical Biochemistry*, *395*(2), 265–267. https://doi.org/10.1016/j.ab.2009.08.016

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, *362*(6422), 709–715. https://doi.org/10.1038/362709a0

Lyra, M. L., Lourenço, A. C., Pinheiro, P. D. P., Pezutti, T. L., Baêta, D., Barlow, A., & Faivovich, J. (2020). High throughput DNA sequencing of museum specimens shed light on the long missing species of the *Bokermannohyla claresignata* group (Anura: Hylidae: Cophomantini). *Zoological Journal of the Linnean Society*, *190*(4), 1235–1255. https://doi.org/10.1093/zoolinnean/zlaa033

Maricic, T., Whitten, M., & Pääbo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One*, *5*(11), e14004. https://doi.org/10.1371/journal.pone.0014004

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet Journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

McCormack, J. E., Tsai, W. L., & Faircloth, B. C. (2016). Sequence capture of ultraconserved elements from bird museum specimens. *Molecular Ecology Resources*, *16*, 1189–1203. https://doi.org/10.1111/1755-0998.12466

McGuire, J. A., Cotoras, D. D., O'Connell, B., Lawalata, S. Z. S., Wang-Claypool, C. Y., Stubbs, A., & Iskandar, D. T. (2018). Squeezing water from a stone: High-throughput sequencing from a 145-year old holotype resolves (barely) a cryptic species problem in flying lizards. *PeerJ*, *6*, e4470. https://doi.org/10.7717/peerj

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prufer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Paabo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, *338*(6104), 222–226. https://doi.org/10.1126/science.1224344

Munchel, S., Hoang, Y., Zhao, Y., Cottrell, J., Klotzle, B., Godwin, A. K., Koestler, D., Beyerlein, P., Fan, J. B., Bibikova, M., & Chien, J. (2015). Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics. *Oncotarget*, *6*(28), 25943–25961. https://doi.org/10.18632/oncotarget.4671

Neumann, D. (2010). Preservation of freshwater fishes in the field. In J. Eymann, J. Degreef, C. H. Häuser, J. C. Monje, Y. Samyn, & D. van den Spiegel (Eds). *Manual on field recording techniques and protocols for all taxa biodiversity inventories and monitoring* (pp. 587–631). ABC Taxa.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L. F., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliussen, T., Malaspinas, A.-S., Vogt, J., Szklarczyk, D., Kelstrup, C. D., … Willerslev, E. (2013). Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, *499*, 74–81. https://doi.org/10.1038/nature12323

Paijmans, J. L. A., Baleka, S., Henneberger, K., Taron, U., Trinks, A., Westbury, M., & Barlow, A. (2017). Sequencing single-stranded libraries on the Illumina NextSeq 500 platform. https://arxiv.org/abs/1711.11004

Paijmans, J. L. A., Barlow, A., Henneberger, K., Fickel, J., Hofreiter, M., & Foerster, D. W. G. (2020). Ancestral mitogenome capture of the Southeast Asian banded linsang. *PLoS One*, *15*(6), e0234385. https://doi.org/10.1371/journal.pone.0234385

Paijmans, J. L. A., Fickel, J., Courtiol, A., Hofreiter, M., & Förster, D. W. (2016). Impact of enrichment conditions on cross-species capture of ancient, archival and fresh DNA. *Molecular Ecology Resources*, *16*, 42–55. https://doi.org/10.1111/1755-0998.12420

Paireder, S., Werner, B., Bailer, J., Werther, W., Schmid, E., Patzak, B., & Cichna-Markl, M. (2013). Comparison of protocols for DNA extraction from long-term preserved formalin fixed tissues. *Analytical Biochemistry*, *439*(2), 152–160. https://doi.org/10.1016/j.ab.2013.04.006

Peacock, M. M., Hekkala, E. R., Kirchoff, V. S., & Heki, L. G. (2017). Return of a giant: DNA from archival museum samples helps to identify a unique cutthroat trout lineage formerly thought to be extinct. *Royal Society Open Science*, *4*, 171253. https://doi.org/10.1098/rsos.171253

Peyrégne, S., & Prüfer, K. (2020). Present-Day DNA Contamination in Ancient DNA Datasets. *BioEssays*, *42*, 202000081. https://doi.org/10.1002/bies.202000081

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., & Pääbo, S. (2014). The complete genome of a Neanderthal from the Altai Mountains. *Nature*, *505*, 4349. https://doi.org/10.1038/nature12886

Prüfer, K., Stenzel, U., Hofreiter, M., Pääbo, S., Kelso, J., & Green, R. E. (2010). Computational challenges in the analysis of ancient DNA. *Genome Biology*, *11*, R47. https://doi.org/10.1186/gb-2010-11-5-r47

Rancilhac, L., Bruy, T., Scherz, M. D., Pereira, E. A., Preick, M., Straube, N., & Vences, M. (2020). Targeted-enrichment DNA sequencing from historical type material enables a partial revision of the Madagascar giant stream frogs (genus *Mantidactylus*). *Journal of Natural History*, *54*, 87–118. https://doi.org/10.1080/00222933.2020.1748243

Richter, S., Schwarz, F., Hering, L., Böggemann, M., & Bleidorn, C. (2015). The utility of genome skimming for phylogenomic analyses as demonstrated for glycerid relationships (Annelida, Glyceridae). *Genome Biology and Evolution*, *7*, 3443–3462. https://doi.org/10.1093/gbe/evv224

Rizzi, E., Lari, M., Gigli, E., De Bellis, G., & Caramelli, D. (2012). Ancient DNA studies: new perspectives on old samples. *Genetics Selection Evolution*, *44*(1), 21. https://doi.org/10.1186/1297-9686-44-21

Robbe, P., Popitsch, N., Knight, S. J. L., Antoniou, P., Becq, J., He, M., Kanapin, A., Samsonova, A., Vavoulis, D. V., Ross, M. T., Kingsbury, Z., Cabes, M., Ramos, S. D. C., Page, S., Dreau, H., Ridout, K., Jones, L. J., Tuff-Lacey, A., Henderson, S., … Schuh, A. (2018). Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. *Genetics in Medicine*, *20*, 1196–1205. https://doi.org/10.1038/gim.2017.241

Rocha, L. A., Aleixo, A., Allen, G., Almeda, F., Baldwin, C. C., Barclay, M. V. L., Bates, J. M., Bauer, A. M., Benzoni, F., Berns, C. M., Berumen, M. L., Blackburn, D. C., Blum, S., Bolanos, F., Bowie, R. C. K., Britz, R., Brown, R. M., Cadena, C. D., Carpenter, K., … Witt, C. C. (2014). Specimen collection: An essential tool. *Science*, *344*(6186), 814–815. https://doi.org/10.1126/science.344.6186.814

Rohland, N., Siedel, H., & Hofreiter, M. (2004). Nondestructive DNA extraction method for mitochondrial DNA analyses of museum specimens. *BioTechniques*, *36*(5), 814–821. https://doi.org/10.2144/04365ST05

Ruane, S., & Austin, C. (2017). Phylogenomics using formalin-fixed and 100+ year old intractable natural history specimens. *Molecular Ecology Resources*, *17*, 1003–1008. https://doi.org/10.1111/1755-0998.12655

Sambrook, J., & Russell, D. (2001). *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press.

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., & Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One*, *7*(3), e34131. https://doi.org/10.1371/journal.pone.0034131

Sheng, G. L., Basler, N., Ji, X. P., Paijmans, J. L. A., Alberti, F., Preick, M., & Barlow, A. (2019). Paleogenome reveals genetic contribution of extinct giant Panda to extant populations. *Current Biology*, *29*, 1695–1700. https://doi.org/10.2139/ssrn.3316802

Simmons, J. E. (2002). Herpetological collecting and collections management. *Society for the Study of Amphibians and Reptiles*, *Herpetological Circular*, *31*, 1–153.

Simmons, J. E. (2014). *Fluid preservation. A comprehensive reference*. Rowman and Littlefield.

Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., & Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neanderthal. *Proceedings of the National Academy of Sciences USA*, *111*, 2229–2234. https://doi.org/10.1073/pnas.1318934111

Snow, A. N., Stence, A. A., Pruessner, J. A., Bossler, A. D., & Ma, D. (2014). A simple and cost-effective method of DNA extraction from small formalin-fixed paraffin-embedded tissue for molecular oncologic testing. *BMC Clinical Pathology*, *14*(30). 1. https://doi.org/10.1186/1472-6890-14-30

Speidel, W., Hausmann, A., Muller, G. C., Kravchenko, V., Mooser, J., Witt, T. J., & Hebert, P. D. N. (2015). Taxonomy 2.0: Sequencing of old type specimens supports the description of two new species of the *Lasiocampa decolorata* group from Morocco (Lepidoptera, Lasiocampidae). *Zootaxa*, *3999*, 401–412. https://doi.org/10.11646/zootaxa.3999.3.5

Springer, M. S., Signore, A. V., Paijmans, J. L. A., Vélez-Juarbe, J., Domning, D. P., Bauer, C. E., He, K., Crerar, L., Campos, P. F., Murphy, W. J., Meredith, R. W., Gatesy, J., Willerslev, E., MacPhee, R. D. E., Hofreiter, M., & Campbell, K. L. (2015). Interordinal gene capture, the phylogenetic position of Steller's sea cow based on molecular and morphological data, and the macroevolutionary history of Sirenia. *Molecular Phylogenetics and Evolution*, *91*, 178–193. https://doi.org/10.1016/j.ympev.2015.05.022

Sproul, J. S., & Maddison, D. R. (2017). Sequencing historical specimens: successful preparation of small specimens with low amounts of degraded DNA. *Molecular Ecology Resources*, *17*, 1183–1201. https://doi.org/10.1111/1755-0998.12660

Stiller, M., Sucker, A., Griewank, K., Aust, D., Baretton, G. B., Schadendorf, D., & Horn, S. (2016). Single-strand DNA library preparation improves sequencing of formalin-fixed and paraffin-embedded (FFPE) cancer DNA. *Oncotarget*, *7*(37), 59115–59128. https://doi.org/10.18632/oncotarget.10827

Straube, N., Iglésias, S. P., Sellos, D. Y., Kriwet, J., & Schliewen, U. K. (2010). Molecular phylogeny and node time estimation of bioluminescent Lantern Sharks (Elasmobranchii: Etmopteridae). *Molecular Phylogenetics and Evolution*, *56*(3), 905–917. https://doi.org/10.1016/j.ympev.2010.04.042

Tang, E. P. Y. (Ed) (2006). *Path to effective recovering of DNA from formalin-fixed biological samples in natural history collections: Workshop summary*. Retrieved from http://www.nap.edu/catalog/11712.html

Taron, U. H., Lell, M., Barlow, A., & Paijmans, J. L. A. (2018). Testing of alignment parameters for ancient samples: evaluating and optimizing mapping parameters for ancient samples using the TAPAS tool. *Genes*, *9*, 157. https://doi.org/10.3390/genes9030157

Templeton, J. E. L., Brotherton, P. M., Llamas, B., Soubrier, J., Haak, W., Cooper, A., & Austin, J. J. (2013). DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *BMC Investigative. Genetics*, *4*(26), https://doi.org/10.1186/2041-2223-4-26

Turvey, S. T., Marr, M. M., Barnes, I., Brace, S., Tapley, B., Murphy, R. W., Zhao, E., & Cunningham, A. A. (2019). Historical museum collections clarify the evolutionary history of cryptic species radiation in the world's largest amphibians. *Ecology and Evolution*, *9*, 10070–10084. https://doi.org/10.1002/ece3.5257

Vieira, F. G., Castruita, J. A. S., & Gilbert, M. T. P. (2020). Using in silico predicted ancestral genomes to improve the efficiency of paleogenome reconstruction. *Ecology and Evolution*, *10*(23), 12700–12709. https://doi.org/10.1002/ece3.6925

Vonk, F. J., Casewell, N. R., Henkel, C. V., Heimberg, A. M., Jansen, H. J., McCleary, R. J. R., Richardson, M. K. (2013). The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proceedings of the National Academy of Sciences*, *110*(51), 20651–20656. https://doi.org/10.1073/pnas.1314702110

Wandeler, P., Hoeck, P. E. A., & Keller, L. F. (2007). Back to the future: Museum specimens in population genetics. *Trends in Ecology and Evolution*, *22*(12), 634–642. https://doi.org/10.1016/j.tree.2007.08.017

Westbury, M., Baleka, S., Barlow, A., Hartmann, S., Paijmans, J. L. A., Kramarz, A., Forasiepi, A. M., Bond, M., Gelfo, J. N., Reguero, M. A., López-Mendoza, P., Taglioretti, M., Scaglia, F., Rinderknecht, A., Jones, W., Mena, F., Billet, G., de Muizon, C., Aguilar, J. L., … Hofreiter, M. (2017). A mitogenomic timetree for Darwin's enigmatic South American mammal *Macrauchenia patachonica*. *Nature Communications*, *8*, 15951. https://doi.org/10.1038/ncomms15951

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, *7*, 1338. https://doi.org/10.12688/f1000research.15931.2

Wu, L., Patten, N., Yamashiro, C. T., & Chui, B. (2002). Extraction and amplification of DNA from formalin-fixed, paraffin-embedded tissues. *Applied Immunohistochemistry and Molecular Morphology*, *10*(3), 269–274. https://doi.org/10.1097/00129039-200209000-00015

Yeates, D. K., Zwick, A., & Mikheyev, A. S. (2016). Museums are biobanks: unlocking the genetic potential of the three billion specimens in the world's biological collections. *Current Opinion in Insect Science*, *18*, 83–88. https://doi.org/10.1016/j.cois.2016.09.009

Zar, G., Smith, J. G., Smith, M. L., Andersson, B., & Nilsson, J. (2019). Whole-genome sequencing based on formalin-fixed paraffin-embedded endomyocardial biopsies for genetic studies on outcomes after heart transplantation. *PLoS One*, *14*(6), e0217747. https://doi.org/10.1371/journal.pone.0217747

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Straube, N., Lyra M. L., Paijmans J. L. A., Preick M., Basler N., Penner J., Rödel M.-O., Westbury M. V., Haddad C. F. B., Barlow A., & Hofreiter M. Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens. *Mol Ecol Resour.* 2021;00:1–17. https://doi.org/10.1111/1755-0998.13433