Unlocking inaccessible historical genomes preserved in formalin

Erin E. Hahn | Marina R. Alexander | Alicia Grealy | Jiri Stiller | Donald M. Gardiner | Clare E. Holleley

Abstract

Museum specimens represent an unparalleled record of historical genomic data. However, the widespread practice of formalin preservation has thus far impeded genomic analysis of a large proportion of specimens. Limited DNA sequencing from formalin-preserved specimens has yielded low genomic coverage with unpredictable success. We set out to refine sample processing methods and to identify specimen characteristics predictive of sequencing success. With a set of taxonomically diverse specimens collected between 1962 and 2006 and ranging in preservation quality, we compared the efficacy of several end-to-end whole genome sequencing workflows alongside a k-mer-based trimming-free read alignment approach to maximize mapping of endogenous sequence. We recovered complete mitochondrial genomes and up to 3x nuclear genome coverage from formalin-preserved tissues. Hot alkaline lysis coupled with phenol-chloroform extraction out-performed proteinase K digestion in recovering DNA, while library preparation method had little impact on sequencing success. The strongest predictor of DNA yield was overall specimen condition, which additively interacts with preservation conditions to accelerate DNA degradation. Here, we demonstrate a significant advance in capability beyond limited recovery of a small number of loci via PCR or target-capture sequencing. To facilitate strategic selection of suitable specimens for genomic sequencing, we present a decision-making framework that utilizes independent and nondestructive assessment criteria. Sequencing of formalin-preserved specimens will contribute to a greater understanding of temporal trends in genetic adaptation, including those associated with a changing climate. Our work enhances the value of museum collections worldwide by unlocking genomes of specimens that have been disregarded as a valid molecular resource.

Keywords

formalin, genome sequencing, hot alkaline lysis, museum, fixed
INTRODUCTION

Natural history collections are a window into the recent past, offering a view of historical biodiversity that is unparalleled in its detail. Collected over the last 250 years, voucher specimens document a period of time over which humans have had a devastating impact on the natural world (Shaffer et al., 1998). The comprehensive metadata associated with each specimen (collection date, location, sex, weight, age, etc.), phenotypic data (e.g., colour, size, gut contents) and genomic data can be used to monitor ecosystem health and study the mechanisms driving adaptation, evolution, speciation and extinction (Holmes et al., 2016; Meineke et al., 2018). The value of collections as sources of historical genetic material has been recognized for the past 30 years, with numerous pathways emerging to retrieve high-quality DNA from challenging archival vertebrate tissues such as skins (Martínková & Searle, 2006), feathers (Rawlence et al., 2009; Sefc et al., 2003), eggshells (Grealy et al., 2019, 2021) and toe pads (Tsai et al., 2020).

DNA degradation associated with preservation method and aging has limited most genetic studies of museum specimens to interrogation of relatively few loci, often targeting genes on the high copy mitochondrial genome via PCR amplification. For phylogenetic studies where a survey of many-fold more loci improves understanding of species’ evolutionary history (Hackett et al., 2008; McElroy et al., 2018; Morgan et al., 2014), genome-wide analyses are increasingly becoming common place. With demand for historical genome-wide data on the rise, newly-developed target-capture approaches now facilitate broader genomic survey from degraded museum specimens (Derkarabetian & Benavides, 2019; Linck et al., 2017; Wood et al., 2018). In some cases, recovery and assembly of whole historical genomes has been achieved (Parejo et al., 2020; Staats et al., 2013), including, extinct species (e.g., passenger pigeon [Murray et al., 2017] and Tasmanian tiger [Feigin et al., 2018]). While technological advances are enabling recovery of genomic data from many museum specimens, genomic study of those preserved with 10% formalin (3.4% w/v formaldehyde) has thus far been very limited (Raxworthy & Smith, 2021).

Formalin-fixation, followed by storage in ethanol, is a common curatorial method used to preserve soft tissue structure. Of the 1.9 million records of preserved chordates within the open-access Atlas of Living Australia (ALA) specimen database (https://www.ala.org.au; 28 January 2021), 33% are classified as “spirit-preserved” (preserved in ethanol with or without prior formalin-fixation). A search for “formalin” preparation within the ALA’s chordate records indicates at least 4% of specimens (N = 77,301) have been formalin-fixed. This is probably a severe underestimate because formalin-fixation is not consistently recorded by all collections. Notably, for fish, reptiles and amphibians, formalin-fixation has historically been the primary method used to preserve tissues long-term while mammals and birds are commonly dry-preserved. Most collections now archive frozen fresh tissue specifically as a genomic resource. However, prior to the 1980s, spirit-preservation was the only method used to preserve soft internal tissue. Thus, spirit collections offer the only opportunity to obtain genetic data from a large proportion of older specimens and holotypes. In particular, spirit collections are enriched for highly biodiverse fish, reptile and amphibian taxa, which, as ectothermic vertebrates may be particularly sensitive to the effects of climate change (Rohr & Palmer, 2013).

Genomic study of formalin-preserved museum specimens has lagged behind because DNA extracted from such tissues is typically low-yield, highly fragmented and heavily cross-linked. PCR amplification of formalin-degraded DNA templates is generally restricted to few, short genomic loci, which provide limited phylogenetic resolution (Appleyard et al., 2021). Formalin fixation presents further challenges by inducing numerous molecular lesions, such as strand breaks, base misincorporation, and both intra- and intermolecular cross-links (Do & Dobrovic, 2015; Srinivasan et al., 2002; Williams et al., 1999). Formaldehyde damage to DNA templates can result in sequencing artefacts that are difficult to differentiate from true genetic variants (Do & Dobrovic, 2015; Williams et al., 1999). Because PCR amplification of damaged DNA is particularly prone to sequencing artefacts, it is preferable to perform deep next-generation sequencing of amplicons (Appleyard et al., 2021) or to avoid amplicon approaches altogether through whole genome sequencing (WGS) of degraded templates (Burrell et al., 2015). Coupled with library preparation methods optimized for low-input and damaged DNA templates (Caraé et al., 2018; Gansauge et al., 2017), high-throughput sequencing can generate enough coverage to call genomic variants with high confidence (Parks & Lambert, 2015). Thus, WGS and reduced representations of genomes could provide a way to overcome the challenges associated with formalin damage and accurately reconstruct historical genetic variation from formalin-preserved tissues.

Promisingly, WGS of formalin-fixed paraffin-embedded (FFPE) archival tissues has become routine in clinical and medical contexts (Robbe et al., 2018; Stiller et al., 2016). However, museum specimens are often older, exposed to higher concentrations of formaldehyde, incubated in the fixative for longer (Simmons, 2014) and in most cases have not been preserved in ideal conditions. Common museum practices, such as failure to rinse specimens prior to permanent storage in ethanol, result in prolonged formaldehyde exposure (MacLeod, 2008). Indeed, many specimens can be in contact with formaldehyde (or its derivatives, such as formic acid) for the entirety of their tenure in a collection. Prolonged formaldehyde exposure, especially under acidic conditions, is thought to result in more extreme DNA degradation (Appleyard et al., 2021; Koshiba et al., 1993). The damage resulting from the preservation process compounds with DNA damage due to natural decomposition, which can be extensive and often precedes any obvious visual indicators of decomposition (Bär et al., 1988). Unfortunately, the time between death and preservation (post-mortem interval) is highly variable and rarely recorded. In light of these additional challenges, WGS methods used with FFPE tissues are relevant but not directly transferable to formalin-fixed museum specimens.

Of the few genetic studies of formalin-fixed museum specimens, most have targeted nuclear (Bibi et al., 2015; Joshi et al., 2013;
Lutterschmidt et al., 2010; Palmer, 2009; Scatena & Morielle-Versute, 2008; Shiozaki et al., 2021 and high copy mitochondrial (Appleyard et al., 2021; Boyle et al., 2004; Shedlock et al., 1997) loci via PCR amplification due to the difficulty and unpredictability of nuclear DNA extraction. There are few examples of broader-scale genomic sequencing of formalin-fixed museum specimens and none have recovered whole vertebrate genomes. WGS of 30-year-old formalin-preserved Anolis lizard (Hykin et al., 2015) and king cobra (Straube et al., 2021) has yielded sufficient coverage to reconstruct entire mitochondrial genomes. Using hot alkaline extraction, whole genomes were recovered for the bioluminescent bacterial symbionts contained within light organs of formalin-preserved cardinalfish (Gould et al., 2020). Using a proteinase K digestion method, sufficient gDNA was recovered to sequence a large number of genomic loci from formalin-preserved snakes (Ruane & Austin, 2017) and spring salamanders (O’Connell et al., 2021) using target capture and restriction-site associated sequencing methods. Hybridization capture baits have also been used to recover the mitochondrial genome from a 120-year-old formalin-preserved Crimean green lizard (Kehlmaier et al., 2019) and an assortment of cartilaginous fish and amphibians collected between 1941 and 2007 (Straube et al., 2021). Highlighting the difficulty of recovering gDNA from formalin-preserved specimens, numerous studies have reported failure to extract and amplify gDNA from formalin-preserved museum tissues (Appleyard et al., 2021; Díaz-Viloria et al., 2005; Pierson et al., 2020). In this context, it is unfortunate yet wise to be hesitant to conduct destructive sampling of formalin-preserved specimens for the purpose of costly WGS.

Recent reports of successful, albeit limited, genomic sequencing in formalin-preserved specimens indicate WGS of higher quality specimens is possible. However, without a framework to guide specimen selection, genomic work on formalin-preserved museum tissues will remain infeasible. It is probably impossible to fully know the numerous and interdependent factors driving sequencing success, for example, age of the specimen (McGaughran, 2020; Watanabe et al., 2017), method of preservation (Zimmermann et al., 2008), post-mortem interval (Bär et al., 1988) and heat and light exposure during storage. However, identification of metrics with which to prescreen specimens for sequencing suitability will improve yield of genomic data while reducing unnecessary destruction of specimens. With screening criteria in hand, museum curators will be less reluctant to grant destructive sampling (Freedman et al., 2018) and researchers will be more inclined to include historical specimens in their analyses.

To facilitate informed-selection of formalin-preserved museum specimens for WGS, we set out to further refine appropriate extraction and library preparation methods and to identify specimen characteristics predictive of DNA extraction and sequencing success. First, we investigated the relationship between residual formaldehyde concentration and pH in preservation media through a survey of specimens in the Australian National Wildlife Collection (ANWC; Crace, Australia). Next, in a phased approach, we compared DNA yield achieved with three extraction methods which have been previously applied to formalin-fixed museum tissues: (1) hot alkaline lysis digestion followed by phenol-chloroform extraction, (2) proteinase K digestion followed by phenol-chloroform extraction, and (3) proteinase K digestion followed by silica spin column purification. While hot alkaline lysis contends with formalin-induced crosslinking via short (25 min) high heat (120°C) treatment under severely alkaline (pH 13) conditions, the more traditional proteinase K digestion performs de-crosslinking of the DNA and peptide bond cleavage simultaneously through overnight enzymatic digestion at 56°C under slightly alkaline (pH 8) conditions. The hard-and-fast hot alkaline lysis method and the gentle-and-slow proteinase K digestion method have the potential to differ in terms of how effectively they recover highly-crosslinked DNA. Moreover, the comparison of phenol-chloroform extraction versus column purification addresses the small fragment sizes expected from formalin-fixed tissues as column purification in general yields highly pure DNA with lower contamination risk (Yang et al., 1998) but phenol-chloroform extraction has been shown to yield higher quantities of DNA from archival tissues (Tsai et al., 2020). After comparing performance of these three DNA extraction methods on a small subsample of three specimens, we then applied the best-performing method to terrestrial vertebrate specimens representing the broad range of tissue quality observed in museum specimens and tested the performance of two library preparation methods: (1) single-stranded method version 2.0 (ss2) (Gansauge et al., 2017), and (2) BEST double-stranded method (dsBEST) (Carøe et al., 2018). Both methods have been shown to effectively construct sequencing libraries from highly degraded DNA but, as of yet, have not been directly compared. Finally, we place our results into context with a comprehensive and unbiased survey of collection-wide spirit preservation conditions and present a decision-making framework to accelerate and facilitate genomic research using formalin-preserved specimens.

2 | MATERIALS AND METHODS

2.1 | Preservation media condition survey

We conducted an unbiased survey of the ANWC spirit vault to measure variation in preservation characteristics that can be sampled without disturbing the specimen. We randomly selected 149 specimen jars spanning a range of taxonomic groups and ages, and removed a 25 ml aliquot of preservation media. We measured pH using an Orion Versa Star Pro benchtop pH meter (Thermo Scientific) and residual formaldehyde concentration ([F]) using MQuant test strips (Merck). Where [F] was at the upper detection limit of the test strips, we diluted the aliquot 1:10 with ultrapure water and remeasured, extrapolating the neat concentration of the media by multiplying the measurement by the dilution factor.

2.2 | Specimen selection

To select specimens for genomic sequencing, we first identified those with a publicly available whole-genome reference for the
specimen species or closely related species. From these specimens, we selected 12 (Table 1) representing a range of taxonomic groups (five birds, three mammals, three reptiles and one amphibian), preservation conditions (three ethanol-preserved and nine formalin-preserved; [F] ranging 0–20,000 mg/l; pH ranging from 4.86–8.2) and ages (collection date ranging 1962–2006). We sampled liver when it was available and muscle or skin from specimens lacking gut contents (Figure 2). From the nine formalin-preserved specimens, we selected three from which liver was available to test the relative success of three DNA extraction methods. To represent “moderate” quality formalin-preserved specimens, we selected a cane toad (Rhinella marina) preserved in 2002. Visually, this specimen appeared minimally degraded and measurements of the storage media indicated low [F] and a neutral pH. To represent “poor” quality formalin-preserved specimens, we selected a tammar wallaby (Macropus eugenii) preserved in 1989 and a saltwater crocodile (Crocodylus porosus) preserved in 1973. Visually, these two “poor” specimens were reasonably well-preserved; however, measurements of the storage media indicated substantial [F] in both specimen jars and mildly acidic pH in that of the wallaby.

2.3 | Tissue preparation

Prior to DNA extraction, we liquid nitrogen pulverized all dissected tissue into a fine powder using a cryoPREP (Covaris) dry pulverizer (three impacts to a TT05 tissueTUBE on intensity setting three; 10 s in liquid nitrogen between impacts). We then stored the pulverized tissue powder in 70% ethanol at ~80°C until further processing. We rehydrated the pulverized tissue by stepping it into 50% ethanol, 30% ethanol then TE buffer with rocking for 10 min intervals. For the nine formalin-fixed tissues, we quenched excess formaldehyde by rocking for 2 h in 1 ml GTE buffer (100 mM glycine, 10 mM Tris-HCL, pH 8.0, 1 mM EDTA), followed by a further wash in fresh GTE for 2 h and a final fresh GTE wash overnight at room temperature. We removed the GTE buffer and washed by rocking in sterile water for 10 min.

2.4 | DNA extraction

With each specimen, we conducted hot alkaline lysis (HA) and two variations on a standard proteinase K (proK) digestion. For the HA extractions, we heated 50 mg (wet weight) tissue aliquots to 120°C for 25 min in 500 µl of alkali buffer (0.1 M NaOH with 1% SDS, pH 13) using an autoclave according to methods described in Campos and Gilbert (2012). We purified DNA from the lysate with three phenol-chloroform extractions followed by ethanol precipitation, resuspending the DNA in 30 µl TE. For the protK extractions, we digested two 50 mg (wet weight) aliquots of tissue overnight at 55°C with 30 µl of 20 mg/ml proteinase K in 970 µl lysis buffer (10 mM NaCl, 20 mM Tris-HCL, pH 8.0, 1 mM EDTA, 1% SDS). We isolated DNA from the proK lysates with either (1) three extractions of phenol-chloroform followed by ethanol precipitation (proK-PC),

TABLE 1 Specimen metadata and independently assessed preservation quality metrics for the samples selected for DNA extraction and sequencing. Twelve specimens (three ethanol-preserved and nine formalin-preserved) from the ANWC spirit vault were selected for DNA extraction and sequencing. Unique ANWC specimen IDs, species names, common name, recorded year of collection, residual formaldehyde concentration in the preservation media (mg/l), pH and tissue sampled for extraction are given as well as specimen quality as defined by criteria established in Figure 5.

| Preservation | Specimen ID | Species name | Common name | Collection year | [F] (mg/l) | pH | Tissue sampled | Specimen quality |
|--------------|-------------|--------------|-------------|----------------|-----------|----|----------------|-----------------|
| Ethanol      | ANWC B30438| Phalacrocorax carbo | Great black cormorant | 1977 | 0 | 8.2 | Skin | Very poor |
|              | ANWC B00001| Aquila audax  | Wedge-tailed eagle | 1973 | 0 | 7.68 | Liver | Good |
|              | ANWC M15492| Phascolarctos cinereus | Koala | 1971 | 0 | 7 | Muscle | Moderate |
| Formalin     | ANWC A02522| Rhinella marina | Cane toad | 2002 | 2050 | 6.41 | Liver | Moderate |
|              | ANWC M11465| Macropus eugenii | Tammar wallaby | 1989 | 8000 | 5.26 | Liver | Poor |
|              | ANWC R03280| Crocodylus porosus | Saltwater crocodile | 1973 | 4000 | 6.31 | Liver | Poor |
|              | ANWC B47838| Melopsittacus undulatus | Budgerigar | 1996 | 5000 | 6.3 | Liver | Moderate |
|              | ANWC R06312| Pogona minor | Dwarf dragon | 1986 | 1800 | 7.04 | Liver | Moderate |
|              | ANWC R01545| Pogona vitticeps | Central dragon | 1971 | 325 | 6.24 | Liver | Moderate |
|              | ANWC B40690| Taeniopygia guttata | Zebras finch | 1986 | 20,000 | 4.86 | Muscle | Very poor |
|              | ANWC B34691| Falco cenchroides | Australian kestrel | 2006 | 2000 | 5.45 | Liver | Poor |
|              | ANWC M03973| Ornithorhynchus anatinus | Platypus | 1962 | 10,000 | 5.79 | Muscle | Very poor |
resuspending the DNA in 30 µl TE, or (2) a QiAquick PCR purification column (Qiagen) (proK-col), following the manufacturer’s instructions and eluting the DNA in 30 µl TE. Alongside the museum tissues, we processed tissue-free controls. We quantified extracted dsDNA using a Qubit fluorometer and high sensitivity (HS) DNA kit (Invitrogen). With HA extractions yielding >200 ng total DNA, we analysed DNA fragmentation using the LabChip GXII (PerkinElmer) capillary electrophoresis system with a 5K HT DNA chip.

2.5 Library preparation methods

To avoid cross-contamination, we prepared all sequencing libraries in the Ecogenomics and Bioinformatics Laboratory trace facility at the Australian National University following standard anti-contamination procedures. We prepared libraries from all DNA extracts and tissue-free controls using two methods developed for high efficiency conversion of fragmented ancient DNA (aDNA): the single-stranded method version 2.0 (ss2) (Gansauge et al., 2017) and the BEST double-stranded method (dsBEST) (Carøe et al., 2018). Concurrently, we prepared DNA-free control libraries, pooled the libraries in approximate equimolar concentrations and measured the concentration of the final pooled library using a Qubit fluorometer and HS DNA kit. The Australian Genome Research Facility sequenced the pooled library on a 150 bp paired-end S4 flow cell on the Illumina NovaSeq 6000 platform. See Supporting Information for further details.

2.6 Quality control and alignment of raw reads

We computed quality control metrics for the raw reads using FastQC version 0.11.8 (Andrews, 2010). Our adapter content analysis included both default Illumina adapters and our custom library adapters. To rapidly detect library contamination by nontarget species’ DNA, we classified the taxonomic origin of reads using Kraken2 version 2.0.9b (Wood et al., 2019). We estimated the number of unique fragments present in the raw sequence libraries with the EstimateLibraryComplexity function of PICARD version 2.9.2 (http://broadinstitute.github.io/picard/). Species-specific reference genomes were not available for three of the specimens. For *Aquila audax*, *Falco cenchroides* and *Pogona minor*, we used the reference genomes of species in the same genera: *A. chrysaetos*, *F. perigirinus* and *P. vitticeps*, respectively (Table S1). The size of the reference genomes ranges from 1.1–3.4 Gb with the bird and reptile genomes generally being smaller than the mammal and amphibian genomes (Table S1). Notably, N50 of the reference genomes varied substantially from a low of 50 kb for *Phalacrocorax carbo* to a high of 387.1 Mb for *M. eugenii* (Table S1). We aligned raw reads with the kallign function of the ngskit4b tool suite version 200218 (https://github.com/kit4b), see Supporting Information for further details. We removed PCR and optical duplicates from the alignments using the MarkDuplicates function of PICARD enabling REMOVE_DUPLICATES=TRUE. For each deduplicated alignment, we computed a histogram of aligned insert lengths and calculated the mean aligned insert length using the CollectInsertSizeMetrics function of PICARD.

2.7 Genome coverage

We estimated nuclear genome coverage ($C_{nuc}$) as the number of unique aligned reads multiplied by the mean insert length divided by unmasked genome size. To estimate how much genomic coverage could be achieved by increasing sequencing depth, we calculated the sequenced proportion of the prepared library as the number of read pairs examined divided by the estimated library size. We estimated the number of possible reads represented in the prepared library by dividing the number of actual reads aligned by the sequenced proportion of the library. We then roughly estimated the potential genomic coverage represented in the full prepared library ($C_{lib}$) as: $$\text{(# possible reads} \times \text{mean insert length (bpi)} \div \text{genome size (bp)}}.$$ To calculate the proportion of mitochondrial genome sites with 30× or greater coverage ($C_{mt}$), we executed the Samtools depth function (Li et al., 2009) on SAM files for the mitochondrial contigs for each species combined across all libraries.

2.8 Statistical analyses

We performed statistical analyses in the R environment, version 4.0.2 (https://www.R-project.org) and produced figures using the packages ggplot2 version 3.3.5 (Wickham, 2016) and ggpubr version 0.4.0 (https://www.github.com/kassambra/ggpubr). To test if the residuals of data were normally distributed, we ran Shapiro–Wilk tests with the function shapiro.test. We conducted t tests with the function t.test, analyses of variance (ANOVA) with the function aov and computed confidence intervals using Tukey’s honest significant difference method (Tukey test) with the function TukeyHSD in the base package stats. We computed Pearson’s correlation coefficients with associated p-values with the ggpubr function stat_cor.

3 RESULTS

3.1 Preservation media condition survey

Within 149 ANWC specimen jars surveyed (23 amphibian, 40 mammal, 40 reptile, and 46 avian), preservation media pH ranged from 4.8–8.4 with 70 (47%), 61 (41%) and 18 (12%) having neutral (6.5–7.5), low (<6.5) and high (>7.5) pH, respectively. [F] ranged from 0–40,000 mg/l. High [F] (>1,000 mg/l) was detected in 61% of low pH jars, 6% of neutral pH jars and 0% of high pH jars. We assumed specimens in jars yielding [F] = 0 (n = 82) were preserved with ethanol and without exposure to formaldehyde. Consistent with the practice of fixing specimens with unbuffered formalin combined with the gradual degeneration of formaldehyde to formic acid, the pH of the formalin-preserved samples (range 4.8–7.1; mean = 6.2)
was significantly lower than for the ethanol-preserved samples (range 6.1–8.4; mean = 7.1) (t test; p < .0001; Figure S1A). The recorded collection date of the specimens ranged from 1936–2015. The time since collection (age) of the ethanol-preserved specimens (mean = 40.1 years) was not significantly different than the formalin-preserved specimens (mean = 36.1 years) (Figure S1B). Among the formalin-preserved samples, [F] and pH were negatively correlated ($R = -0.6, p < .001$; Figure 1). Age was not significantly correlated with either [F] or pH. Of the 12 specimens selected for sequencing, collection date ranged from 1962–2006 and pH ranged from 4.9–8.2. Three sequenced specimens were ethanol-preserved and nine sequenced specimens were formalin-preserved with nine sequenced specimens were ethanol-preserved and nine sequenced specimens were formalin-preserved samples, [F] and pH were negatively correlated ($R = -0.6, p < .001$; Figure 1). Age was not significantly correlated with either [F] or pH. Of the 12 specimens selected for sequencing, collection date ranged from 1962–2006 and pH ranged from 4.9–8.2. Three sequenced specimens were ethanol-preserved and nine sequenced specimens were formalin-preserved with [F] ranging from 325–20,000 mg/l (Table 1).

**3.2 | DNA quantification**

We compared DNA yield from liver with HA, proK-PC and proK-col extraction methods for the *R. marina*, *M. eugenii* and *C. porosus* specimens and observed no significant differences between extraction methods (one-way ANOVA; Figure S3A). However, the HA method produced more DNA from the two poor quality specimens (*M. eugenii* and *C. porosus*) compared to either of the proK methods (Table 2). Thus, we predicted the HA method would perform better on specimens ranging broadly in preservation quality and we used this method to extract the remaining nine specimens. HA extraction yielded DNA detectable by high sensitivity Qubit for all 12 specimens.

The extracted DNA was highly fragmented, particularly for extracts from formalin-preserved tissues (Figure S4). Two ethanol-preserved specimens (*A. audax* and *Phascolarctos cinereus*) and two formalin-preserved specimens (*R. marina* and *Melopsittacus undulatus*) yielded $>1,000$ ng total DNA from 50 mg of tissue (Table 2). Three specimens, *P. carbo*, *Taeniopygia guttata* and *O. anatinus*, yielded particularly low (<100 ng) total DNA from 50 mg of tissue (Table 2). We observed no significant difference in DNA yield between ethanol and formalin-preserved specimens (t test; Figure S3B). However, mean DNA yield from ethanol-preserved specimens was more than double that from formalin-preserved specimens. Mean DNA yield from formalin-preserved specimens in preservation media with low pH (<6) was not significantly different from those in media with neutral to high pH (>6) (Figure S3C). DNA yield was significantly higher from formalin-preserved liver tissue compared to nonliver tissue (t test; $p < .05$; Figure S3D). Both [F] and age showed a negative but nonsignificant correlation with DNA yield from formalin-preserved specimens (Figures S3E and F).

**3.3 | Prealignment library quality assessment**

All libraries contained a high proportion of adapter content and low read quality score beginning at roughly 50 bp, consistent with highly fragmented input DNA. Focusing on the first 75 bp of the raw reads, mean sequence quality was slightly but significantly higher for read 2 (mean Phred score = 34.3) than for read 1 (mean Phred score = 33.7) across all libraries (paired t test; $p<.001$). Likewise, the mean sequence quality was significantly higher in ss2 libraries compared to the corresponding dsBEST libraries for both read 1 (mean of the differences = 2.1; paired t test; $p<.001$) and read 2 (mean of the differences = 0.79; paired t test; $p<.01$). Mean sequence quality was not significantly different between reads derived from ethanol and formalin-preserved tissues, even when excluding libraries prepared from less than 200 ng of input DNA (paired t test). We found evidence of cross-contamination in several libraries prepared from low DNA yield extractions. Compared to negative controls, both *Ornithorhynchus anatinus* libraries and all but two *C. porosus* libraries showed a higher number of reads classified as genus *Mus* by Kraken2 (Table S2). The *O. anatinus* libraries also contained a high percentage of reads classified as *Homo sapiens* (9.7% and 25%). The *O. anatinus* and *C. porosus* tissues were among those that yielded the least DNA. The *O. anatinus* HA extraction yielded just 22 ng. The *C. porosus* HA and proK-PC extractions yielded 130 and 79 ng, respectively, while the proK-col extraction yielded no detectable DNA. The only other specimens to yield less than 500 ng were the *P. carbo*, *T. guttata* and *M. eugenii*.

**3.4 | Relative alignment quality from three extraction methods**

We used three indicators of alignment quality to compare the relative success of the three extraction methods on the *R. marina*, *M.*
TABLE 2  Sequencing and alignment statistics. For all specimens, DNA yield is given for all extractions. For the remaining metrics, the values shown were calculated having combined both the ss2 and dsBEST libraries. The number of raw reads is given as a sum of all single reads (R1 and R2) from the paired-end sequencing run. Reads aligned indicates the percent of raw reads aligned to reference genome after removal of PCR and optical duplicates. The mean aligned insert length is the mean length (in bp) of the aligned portion of the read. C_nuc is the coverage of the nuclear genome. C_mt is the proportion of mitochondrial genome with greater than 30x coverage. C_pot is the estimated potential genomic coverage if the full library had been sequenced, calculated from the estimated library complexity. MRM is the number of reads aligned to the mitochondrial genome per one million raw reads. GenBank accession numbers are provided for full mitochondrial sequences from specimens yielding C_mt > 0.5.

| Preservation | Species                  | Extraction method | DNA yield from 50 mg (ng) | Raw reads (million) | Reads aligned (%) | Mean aligned insert length (bp) | C_nuc | C_mt | C_pot | MRM | GenBank Accession |
|--------------|--------------------------|-------------------|---------------------------|---------------------|------------------|--------------------------------|-------|------|-------|-----|-------------------|
| Formalin     | Rhinella marina          | HA                | 1860                      | 434                 | 21               | 65                             | 2.2   | 0.78 | 6.2   | 52  | MZ895076          |
|              |                          | proK-col          | 666                       | 77                  | 40               | 81                             | 1     | 0    | 6.2   | 14  |                   |
|              |                          | proK-PC           | 2550                      | 321                 | 15               | 74                             | 1.2   | 0.42 | 11.4  | 29  |                   |
|              | Macropus eugenii         | HA                | 271                       | 306                 | 8                | 56                             | 0.5   | 0.59 | 2.7   | 50  |                   |
|              |                          | proK-col          | 4                         | 17                  | 1                | 67                             | 0     | 0    | 0.1   | 3   | MZ595099          |
|              |                          | proK-PC           | 33                        | 801                 | <1               | 65                             | 0     | 0    | 0.1   | 2   |                   |
|              | Crocodylus porosus       | HA                | 130                       | 23                  | <1               | 67                             | 0     | 0    | 0     | 11  |                   |
|              |                          | proK-col          | None detected             | 160                 | <1               | 70                             | 0     | 0    | 0.1   | 12  |                   |
|              |                          | proK-PC           | 79                        | 294                 | <1               | 62                             | 0     | 0    | 0     | 2   |                   |
|              | Melopsittacus undulatus  | HA                | 2400                      | 318                 | 20               | 60                             | 3.1   | 0.94 | 23.6  | 201 | MZ571484          |
|              | Pogona minor             | HA                | 521                       | 367                 | 7                | 58                             | 0.8   | 0.51 | 7.5   | 29  | MZ595323          |
|              | Pogona vitticeps         | HA                | 672                       | 432                 | 15               | 59                             | 2.1   | 0.85 | 7.9   | 52  | MZ595322          |
|              | Taeniopygia guttata      | HA                | 15                        | 62                  | <1               | 66                             | 0     | 0    | 0     | 1   |                   |
|              | Falco cenchroides        | HA                | 690                       | 303                 | 5                | 56                             | 0.7   | 0.12 | 2.1   | 14  |                   |
|              | Ornithorhynchus anatinus | HA                | 22                        | 520                 | <1               | 70                             | 0     | 0.13 | 0.8   | 20  |                   |
| Ethanol      | Phalacrocorax carbo      | HA                | 57                        | 292                 | <1               | 69                             | 0.10  | 0.90 | 0.60  | 50.00 | MZ595096          |
|              | Aquila audax             | HA                | 1932                      | 282                 | 67               | 76                             | 11.3  | 0.98 | 323   | 2515 | MZ595097          |
|              | Phascolarctos cinereus   | HA                | 1254                      | 423                 | 60               | 76                             | 5.4   | 0.94 | 93    | 2606 | MZ595098          |
eugenii and C. porosus specimens: percent of raw reads aligned to the genome (% alignment), the number of reads aligned to the mitochondrial genome expressed as the number of reads aligned per million raw reads (MRM) and the mean aligned insert length. Among these three specimens, we observed no significant differences between library preparation methods in any of the three alignment quality indicators (paired t tests). Therefore, we took the mean of the two library preparations to compare extraction methods across each alignment quality indicator. Again, we observed no significant difference in MRM between the extraction methods (one-way ANOVA; p < .05). We observed no significant difference in MRM between the proK-col and proK-PC methods (Tukey tests; p < .05). We observed no significant difference in % alignment or mean insert length between the three extraction methods (one-way ANOVA).

3.5 | Effect of specimen quality on sequencing success

The percentage of aligned reads removed by optical and PCR deduplication varied between 8.8% and 99.5% across all libraries. Among the HA alignments, deduplication reduced significantly more mapped reads from dsBEST libraries than from ss2 libraries (paired t test; p < .01). Combining the ss2 and dsBEST libraries for each HA extraction, deduplication removed more than double the
percentage of reads (69.8% vs. 32.8%) from poor quality specimens (those yielding <1% reads aligned) compared to better quality specimens (those yielding >1% reads aligned). However, this difference was not significant (t test). Deduplication removed significantly more reads from the formalin-preserved specimens (mean = 54.6%) than from the ethanol-preserved specimens (mean = 16.7%) (t test; p < .01). Following deduplication, the mean percent of mapped reads remaining was 44% and 59% for the dsBEST and ss2 HA libraries, respectively. Across all specimens extracted using the HA method, we observed no significant differences between library preparation methods in any of the three alignment quality indicators (paired t tests). Therefore, we conducted further comparison of the effect of specimen quality on alignment success taking the mean of each alignment quality indicator from the two HA library preps.

HA extraction of one of three ethanol-preserved specimens (P. carbo) and three of nine formalin-preserved specimens (C. porosus, T. guttata and O. anatinus) produced <1% aligned reads (Table 2), indicating equal rates of very poor sequencing success with ethanol- and formalin-preserved tissues. Excluding the specimens with <1% aligned reads, the ethanol-preserved specimens produced a significantly higher percentage of aligned reads (t test; p < .01). Two of the three ethanol-preserved specimens (A. audax and P. cinereus) produced >60% aligned reads while the remaining six formalin-preserved specimens (R. marina, M. eugenii, M. undulatus, P. minor, P. vitticeps and F. cenchroides) produced between 5 and 21% aligned reads (Table 2). Excluding the specimens with <1% aligned reads, the mean insert length was significantly longer for the ethanol-preserved specimens (mean = 76 bp) compared to the formalin-preserved specimens (mean = 59 bp) (t test; p < .0001). MRM was also significantly higher for the ethanol-preserved specimens (mean = 2,560) compared to the formalin-preserved specimens (mean = 43) (t test; p < .01).

The percent of reads aligned was highly and significantly correlated to DNA yield across all extracts (R = .78; p < .001) as well as just those from formalin-preserved tissues (R = .78; p < .001) (Figures S3g and h). The percentage of reads aligned increased with preservation media pH (R = .44; Figure 3a), decreased with preservation media F (R = -.53; Figure 3b) and decreased with specimen age (R = -.46; Figure 3c), although these correlations were not statistically significant. The percentage of aligned reads was significantly higher in specimens sampled with liver than those sampled with muscle and skin (t test; p < .05; Figure 3d). Of the specimens yielding poor sequencing success (<1% reads aligned), all but C. porosus were sampled with either muscle or skin as liver was not present. The only specimen sampled with a tissue other than liver to yield a percent of reads aligned >1% was the ethanol-preserved P. cinereus.

3.6 | Genome sequencing coverage

Nuclear genome coverage (C_nuc) of the deduplicated alignments was <1x for the majority of libraries. Since raw read yield was highly variable, C_nuc is not an appropriate measure with which to compare the extraction or library preparation methods. However, it is noteworthy that we achieved C_nuc >1x for two of the ethanol-preserved specimens and three of the formalin-preserved specimens. Combining all libraries for a given specimen, we achieved a total of 5.4x and 11.3x C_nuc for the ethanol-preserved P. cinereus and A. audax specimens, respectively (Table 2). Likewise, we achieved a total of 2.1x, 3.1x and 4.4x C_nuc for the formalin-preserved P. vitticeps, M. undulatus and R. marina specimens, respectively (Table 2). To estimate the potential for improving C_nuc through resequencing of the prepared libraries, we calculated potential genomic coverage (C_pot) (Table 2). Combining all libraries for a given specimen, C_pot exceeded 20x for the R. marina and M. undulatus and exceeded 75x for the P. cinereus and A. audax. Focussing on the mitochondrial genome, the proportion of sites with 30x or higher coverage (C_mit) was nearly complete (>0.9) for all three ethanol-preserved specimens (Table 2). C_mit for the formalin-preserved M. undulatus (0.94) was comparable to that of the ethanol-preserved specimens. C_mit was moderate to high (>0.5) for five of the formalin-preserved specimens (Table 2). Only the C. porosus, T. guttata, F. cenchroides and O. anatinus yielded very poor C_mit (<0.15).

3.7 | Read length periodicity

From the aligned insert lengths estimated with Picard, we plotted the frequency of reads between 50 and 100 bp (Figure 4). This plot revealed a pattern of read length periodicity in several specimens, notably those that resulted in higher mapping success. We observed prominent periodicity of approximately 10.1 bp in the R. marina specimen extracted with the proK-PC method. While less pronounced, we observed read length periodicity of approximately 10.8 bp in the HA extractions of R. marina, P. vitticeps, P. minor, F. cenchroides, A. audax and P. cinereus. The pattern of periodicity was observed in both the dsBEST and ss2 libraries, however, it was slightly more pronounced in the dsBEST libraries.

4 | DISCUSSION

Despite numerous reports of successful DNA sequencing from formalin-preserved tissues, it remains a common perception that formalin-preserved museum specimens are unreliable sources of accessible DNA. To challenge this perception and increase confidence among both researchers and museum curators, we conducted a systematic study of sequencing success in spirit-preserved museum specimens. Processed with a tailored molecular and bioinformatic workflow, formalin-preserved specimens had an overall sequencing success rate (success defined loosely as yielding >1% reads aligning to the genome) equivalent to ethanol-preserved specimens, albeit with recovery of a lower percentage of sequence reads mapping to the reference genome. Contrary to popular belief, we found genomewide nuclear data is retrievable from some formalin-preserved museum specimens, even with a moderate investment of sequencing.
effort (with 30% of formalin-preserved specimens, we achieved >2× nuclear genome coverage from 300–500 million raw reads). We also show reconstruction of large sections of the mitochondrial genome is possible even in poor quality specimens where limited nuclear data were recovered (with 55% of formalin-preserved specimens, we achieved >30× coverage of more than 50% of the mitochondrial genome). Investigating specimens covering a range of preservation quality, we also developed a decision-making framework to improve sequencing success rate and prioritize suitable specimens. Our findings support a considered and targeted sequencing approach that transforms thousands of spirit collection specimens into a new molecular resource. Improved access to genomic data held in these specimens has the potential to inform research into the mechanisms driving adaptation, evolution, speciation and extinction.

4.1 | Hot alkaline lysis effectively recovers gDNA from formalin-preserved archival tissues suitable for next generation sequencing

Originally developed for DNA extraction from FFPE sections, the HA method relies on high heat (120°C) under alkaline conditions (pH = 13) to break strong inter- and intramolecular cross links and utilizes organic extraction to maximize capture of fragmented gDNA from formalin-preserved tissues (Campos & Gilbert, 2012; Shi et al., 2002, 2004). This method has been applied to museum specimens to successfully recover sections of the mitochondrial genome in trout (Splendiani et al., 2017) and full mitochondrial genomes from lizards (Hykin et al., 2015) and bacterial symbionts (Gould et al., 2020). Here we show that the HA method yields gDNA in adequate quantities for WGS from higher-quality formalin-preserved museum specimens. Coupled with library preparation methods designed to efficiently convert degraded DNA, we produced complex sequencing libraries with the potential to recover full vertebrate genomes when mapped using a strategy optimized to maximize recovery of endogenous sequence. Our results indicate that the HA method is appropriate for DNA extraction from a broad range of taxa preserved under various conditions, making it well-suited for application in both museum and pathological settings.

In a small-scale comparison of the HA method to proK digestion with either phenol-chloroform extraction or column purification, we found the HA method to perform superiorly for poor quality formalin-preserved specimens. Specifically, the HA extraction from M. eugenii yielded substantially greater DNA, resulted in higher percent of reads mapping to the nuclear genome and recovered a greater proportion of the mitochondrial genome. Recently, application of mechanical energy via a vortex fluidic device has been shown to increase proteinase K activity, reduce processing time and improve recovery of DNA from formalin-preserved tissues (Totoiu et al., 2020). Thus, there is potential for further optimisation of the proK method,
including increased incubation temperature (Straube et al., 2021),
to increase its efficiency relative to the HA method in recovering
DNA from museum specimens. We experienced equal success rates
with the HA method in formalin and ethanol-preserved tissues. It is
not standard practice to apply the HA method to ethanol-preserved
specimens, which do not suffer from cross-linking, but we imple-
mented it in this study to serve as a comparison to formalin-fixed
tissues. Thus, while the HA method is probably unnecessarily harsh
for recovery of DNA from tissues not crosslinked with formalde-
hyde, we propose this extraction method is suitable across a wide
range of tissue qualities and preservation conditions observed in
museum spirit collections. Given that we achieved relatively high
yield from the ethanol-preserved tissues, we propose that the HA
method is appropriate in cases where contact with formalin cannot
be determined. We caution, however, the HA method’s success may
be limited to DNA-rich tissues such as liver as has been previously
observed for the proK method (O’Connell et al., 2021). Our HA ex-
tractions of formalin-preserved muscle and ethanol-preserved skin
tissue failed to yield adequate gDNA for sequencing, while our HA
extraction of ethanol-preserved muscle tissue was less successful
than our extraction of ethanol-preserved liver tissue. HA extraction
has been previously observed to perform poorly compared to
cetyltrimethylammonium bromide (CTAB) protocols on formalin-
preserved mammalian heart tissue (Paireder et al., 2013). We also
note that, preservation conditions being equal, DNA yield may differ
between taxonomic groups due to factors such as blood cell nucle-
lation. Due to low sample size, we were not able to test if the lack of
nucleated red blood cells in mammal tissues impacted DNA yield.

4.2 | aDNA library preparation methods effectively
capture DNA extracted from formalin-preserved
archival tissues

DNA degradation in museum specimens is a significant challenge to
genie sequencing. To improve our conversion of degraded DNA
from formalin-preserved tissues into high quality library molecules,
we utilized two library preparation methods developed specifi-
cally for degraded aDNA templates. We tested the ss2 (Gansauge
et al., 2017) and dsBEST (Carøe et al., 2018) methods on DNA
extracted from both ethanol and formalin-preserved archival tis-
sues. Compared to standard double-stranded DNA library prepa-
ration methods, the ss2 method performs superiorly in converting
highly fragmented DNA extracted from formalin-preserved tissues
(Gansauge et al., 2017). Here, we compared the ss2 method to an
improved double-stranded library preparation method, dsBEST, which
employs minimal tube transfers to retain high proportions of frag-
mented endogenous DNA (Carøe et al., 2018). Sequence quality was
significantly higher for libraries prepared using the ss2 method com-
pared to the dsBEST protocol. However, this quality difference did
not result in significantly lower rates of read alignment or reduced
mapped insert length for the dsBEST libraries. While we did not see
differences in contamination rates between the two methods, an
advantage of the dsBEST method is its reliance on fewer tube trans-
fers and additions of solution, thus reducing opportunities to lose
DNA and introduce contaminants. The ss2 and dsBEST methods
performed similarly on all 12 of our archival templates, indicating
both are well-suited to prepare libraries from DNA extracted from
ethanol and formalin-preserved tissues. Alternative library prepara-
tion methods developed specifically for degraded DNA may prove
equally effective. To maximize conversion of fragmented archival
DNA template, we advise using a library preparation method de-
signed to capture small fragments whilst minimising contamination.
Overall, we found DNA yield was highly predictive of mapping success and we observed samples with very low DNA yield (<200 ng from 50 mg of tissue) did not produce libraries with high rates of mapping success. Consistent with our observations, a recent study of formalin-preserved salamanders found that specimens yielding <200 ng of DNA contained high levels of exogenous DNA (O’Connell et al., 2021). Thus, as a cost-saving measure, we advise quantifying DNA templates prior to library preparation and focusing sequencing effort on higher yielding samples. However, we caution that total DNA yield is not always informative about endogenous DNA content.

4.3 High alignment rates of fragmented DNA are achieved through exhaustive match searching

Removal of adapter sequence and low-quality bases via read-trimming is a standard pre-processing procedure conducted on raw sequencing reads prior to mapping. In the context of libraries prepared from highly degraded templates, filtering and trimming can reduce the data set substantially. For example, preprocessing of the library prepared from a formalin-preserved Anolis lizard reduced the data set to 13.5% of the raw data (Hykin et al., 2015). Although filtering and trimming are effective at removing PCR duplicates and erroneous bases introduced through library preparation and sequencing, quality control parameters should be optimized to avoid removing informative endogenous sequence, particularly with data derived from highly fragmented low-input templates. Compared to DNA extractions from fresh tissue, our extractions from formalin-preserved specimens were highly fragmented as is typical of aDNA sources (Prüfer et al., 2010). We opted to trial a computationally efficient approach that eliminates loss of endogenous sequence during preprocessing. The kalign function from the open source kit4b toolkit performs alignments of raw reads by searching for the maximum length match within the read to the reference sequence regardless of the match’s position within the read. For each raw read, kalign performs a rapid complete exhaustive match search across the indexed reference genome. The match search is performed recursively through seed expansions generated along the read length. The longest match to endogenous sequence is retrieved while satisfying the
minimum length threshold of the match. Using this approach, we aligned up to 21% and 67% of raw reads from formalin and ethanol-preserved tissues, respectively. These alignment rates are consistent with the degree of degradation in the DNA we extracted from spirit-preserved museum specimens being intermediate between that of fresh and truly ancient tissues. A previous application of the ss2 method yielded a maximum of 11.3% mappable reads from libraries prepared from aDNA tissue sources (Gansauge et al., 2017). The same study yielded 60% and 68% mappable reads from libraries prepared from horse and pig liver stored in buffered formalin for 5 and 11 years, respectively (Gansauge et al., 2017). In comparison, our modest alignment rates may be the result of tissues of intermediate age and using a different metric of calculating the percent of mapped reads.

4.4 Sequencing success is strongly influenced by specimen integrity prior to fixation

To explore the effects of formalin-fixation on sequencing success, we selected three specimens preserved with ethanol only and nine specimens preserved with formalin. We found no significant difference in DNA yield between the ethanol and formalin-preserved specimens and the differences we observed in DNA fragment lengths were minimal. Furthermore, we observed equal rates of very poor sequencing success within ethanol and formalin-preserved specimens, indicating preservation method is not a strict determinant of sequencing success. Older, poor-quality ethanol-preserved specimens have previously been shown to be as problematic for genomic analyses as formalin-preserved specimens (McGuire et al., 2018; Ruane & Austin, 2017). This is not to say preservation method does not impact sequencing success. Two of our ethanol-preserved specimens (P. cinereus and A. audax) had much higher mapping rates (60% and 67% reads aligned, respectively) than even our most successful formalin-preserved specimens (R. marina, produced 21% reads aligned with the HA method) despite the expected reduction of mapping success from having mapped the A. audax reads to the genome of different species in the same genus (A. chrysaetos). Our findings indicate WGS of formalin-preserved museum specimens is possible using HA extraction paired with a library preparation optimized for conversion of degraded DNA. However, as with all potential DNA sources, the overall integrity of the tissue will ultimately determine sequencing success.

The specimens with poor sequencing success (<1% reads aligned) were largely older, their preservation media had lower pH and higher [F] and they were sampled with a tissue other than liver. On the contrary, the specimens with better sequencing success were preserved more recently, their preservation media had neutral pH and lower [F] and the tissue sampled was liver. We calculated the correlation between specimen quality measures ([F], pH, age and tissue type) and both DNA yield and mapping success. Tissue type was the only quality measure significantly associated with lower DNA yield, with liver yielding significantly more DNA than either muscle or skin. Our higher success with liver is consistent with previous studies comparing sequencing success from liver, muscle and tail-tip in a formalin-preserved Anolis lizard (Hykin et al., 2015) and from liver, bone and skin in spirit-preserved garter snakes (Zacho et al., 2021).

Post-mortem DNA degradation occurs more rapidly in liver relative to other bodily tissues including skeletal muscle, heart and brain (Itani et al., 2011; Johnson & Ferris, 2002). In the museum curatorial setting, specimens undergo varying degrees of post-mortem decay prior to fixation. As is the case for most museum specimens, the length of the post-mortem interval (PMI) was not recorded for the specimens used in this study. Given expected rapid decay of the viscera, we used the visual appearance of the gut contents as a reasonable proxy for the length of the PMI. The four specimens used in this study that lacked liver tissue were visibly more degraded than those with intact liver tissue (Figure S2). In the case of the P. cinereus, P. carbo and O. anatinaus, the complete absence of viscera indicated the internal organs were probably well-degraded and discarded prior to fixation. For specimens preserved after a long PMI, DNA integrity throughout the carcass would be lower than in specimens preserved after a short PMI. Therefore, we conclude that the higher yield from specimens sampled with liver is a reflection of overall specimen quality and DNA damage occurring post-mortem but prior to fixation.

Relative sequencing success between specimens is probably influenced by differences in genome assembly quality (Smits, 2019) as well as genetic divergence of the archival specimen from the reference genome (Günther & Nettelblad, 2019; Straube et al., 2021). While our reference genomes were all within the average for vertebrates, species with much larger genomes will probably prove more difficult to sequence. Species-specific genomes were not available for three of our specimens (A. audax, F. cenchroides and P. minor) and, thus, it is possible that the percent of reads aligned for these specimens could be improved further if a new reference were to be made available. It is difficult to assess the impact of reference genome quality and divergence on mapping success with only single specimens from each species. However, we did observe that although the P. vitticeps and P. minor specimens yielded roughly equal amounts of DNA (672 and 521 ng, respectively) the percentage of reads aligned and mitochondrial genome coverage was much lower in P. minor data mapped to a P. vitticeps reference (Table 2). With respect to genome assembly quality, the genome with the lowest scaffold contiguity (P. carbo, N50 = 0.05 Mb) yielded an extremely low percentage of reads aligned (<1%) despite relatively high mitochondrial genome coverage (Cmt = 0.90). These results warrant further in-depth investigation and underscore the importance of sourcing and vetting species-specific reference genomes.

4.5 Rethinking formalin damage

Formalin-preserved museum specimens have long been considered intractable sources of gDNA. Encouragingly, we found specimen contact with formaldehyde does not prohibit DNA sequencing if tissue decomposition occurring prior to fixation is minimized. With
appropriate sample vetting (Figure 5). HA extraction and DNA library preparation optimized for degraded DNA, historical genomic data may be extracted from many formalin-preserved specimens. These data will not be of similar quality to those recovered from fresh or ethanol-preserved tissues. However, higher sequencing volume and borrowing of analytical methods from the field of aDNA may facilitate reconstruction of historical genomes from formalin-preserved tissues. We found evidence that DNA damage in formalin-preserved specimens shares characteristics with that of aDNA. In addition to capturing shorter fragments with low mapping rates, we observed a pattern of read length periodicity of approximately 10 bp. This is consistent with observations in aDNA specimens (Pedersen et al., 2014) and is an interval that coincides with the length of a turn of the DNA helix. Pedersen et al. (2014) attributed the 10 bp read periodicity in specimens greater than 4,000 years old to protection of the DNA by nucleosomes preferentially positioned at 10 bp intervals. In a meta-analysis of ancient genomes, fragmentation bias consistent with protection of the DNA by nucleosomes was associated more strongly in specimens from regions with lower annual mean temperature suggesting that cellular architecture is better preserved in colder climates (Kistler et al., 2017). We observed a striking periodicity pattern averaging 10.8 bp in HA extracted samples and 10.1 bp in the proK-PC samples. The shorter periodicity in the proK treated samples may be due to reduced protection of the ends of DNA fragments by digestion of the nucleosomes during extraction. Our observation of 10 bp read length periodicity in the majority of our more successful specimens is consistent with rapid initial post-mortem DNA damage occurring prior to fixation followed by relative stability as has been proposed as a model of aDNA degradation (Pääbo, 1989). Furthermore, the appearance of 10 bp periodicity suggests it may be possible to infer nucleosome occupancy from patterns of DNA degradation observed in formalin-preserved specimens if higher coverage is achieved. This signal, as well as nucleosome occupancy signatures from degraded DNA serves as a window into historical genome architecture. While we did not observe a signal of nucleosome occupancy in read depth as did Pedersen et al. (2014) perhaps such a signal would be revealed if we were to have sequenced to higher depth.

4.6 | Managing expectations

We have shown WGS of formalin-preserved museum specimens is feasible and success can be improved through specimen quality vetting. We stress, however, measures of specimen quality are imperfect and the key parameters may vary between and within museum collections. Modern collection institutions aim to limit light exposure and temperature variation within their spirit vaults. With older specimens, the likelihood they have been exposed to undocumented DNA-degrading conditions increases. Consistent with a study of garter snake specimens collected between 1842 and 1964 (Zacho et al., 2021), we found the age of the specimen was not strongly predictive of sequencing success, however, we did not sample specimens collected prior to the 1960s. This warrants further investigation into the extent to which intact DNA can be extracted from much older formalin-preserved specimens.

While preservation media pH and [F] were not predictive of sequencing success in our specimens, we note these measures do not always accurately reflect preservation condition. Most institutions periodically top up the specimen jars in their spirit vaults to replace ethanol lost through evaporation. In some cases, the preservation media is replaced entirely. Thus, media pH and [F] values at the time of sampling for sequencing may not reflect preservation and long-term storage conditions. With additional sampling of older and more varied specimens, it may be possible to establish clear correlates of sequencing success associated with pH and [F].

Both researchers and museums would benefit from an improved set of guidelines for strategic decision-making based on independent quality metrics rather than qualitative ad hoc assessments. This will empower researchers to most effectively deploy their sequencing budgets and support museums in deciding when to grant requests for destructive sampling. A cost-benefit analysis should be conducted prior to genomic sequencing of museum specimens. From the perspective of the museum, destructive sampling should be avoided if the specimen is unlikely to yield sufficient DNA to achieve a project’s aims. From the perspective of the researcher, sequencing of high-quality specimens should be prioritized to generate high-quality data. To assist in making these assessments, we provide a decision-making tree (Figure 5) for use by both curators and researchers which describes how we will use the results from this study to vet specimens for future genomic analyses. We present this tree as a provisional framework to guide further refinement of specimen decision-making as additional evidence of sequencing success is generated for museum specimens.

Ultimately, museum curators decide if the potential benefit of sequencing outweighs the damage to the specimen through destructive sampling. Once sampling and DNA extraction has been completed, the decision to proceed with library preparation and sequencing can be made on the basis of DNA yield. We found specimens with high DNA yield (>1,500 ng/50 mg tissue) produced a high percentage (>20%) of mappable reads while specimens with low DNA yield (<200 ng/50 mg tissue) produced virtually no mappable reads. While specimens yielding between 200–1,500 ng of DNA per 50 mg tissue produced relatively low genomic coverage, they did produce high coverage of the mitochondrial genome. Thus, reconstruction of historical mitochondrial haplotypes may be possible from specimens yielding low quantities of DNA. When nuclear data is required, high-volume sequencing should be reserved for high-quality specimens. Generally speaking, most research projects aim to sequence a small number of museum specimens with which to provide a baseline for comparison to contemporary specimens. In light of the limited availability of historical specimens in collections, it is often reasonable and feasible to allocate a relatively large budget to conduct deep sequencing of a small number of specimens.
ACKNOWLEDGEMENTS
We thank Olly Berry and Andrew Young for their leadership within the Environomics Future Science Platform. We thank the director of the Australian National Wildlife Collection, Leo Joseph, and the ANWC staff (specifically, Margaret Cawsey, Alex Drew, Tonya Haff, Dave Spratt and Chris Wilson) for their contributions of curatorial expertise, metadata management and sampling assistance. We thank Kerensa McElroy for her assistance and guidance in data management. We thank Ondrej Hlinka and CSIRO IM&T Client Services for their assistance in utilising the CSIRO Pearcey supercomputing system. We thank Niccy Aitkin for her guidance in utilising the Australian National University’s Ecogenomics and Bioinformatics Laboratory for library preparation. We thank the Australian Genome Research Facility for their conversations around sequencing. We thank Sharon Appleyard, Meghan Castelli, Andrew Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Appleyard, S. A., Maher, S., Pogonoski, J. P., Bent, S. J., Chua, X.-Y., & McGrath, A. (2021). Assessing DNA for fish identifications from reference collections: The good, bad and ugly shed light on formalin fixation and sequencing approaches. Journal of Fish Biology, 98, 1421–1432. https://doi.org/10.1111/jfb.14687

Bär, W., Kratzer, A., Mächler, M., & Schmid, W. (1988). Postmortem stability of DNA. Forensic Science International, 39(1), 59–70. https://doi.org/10.1016/0379-0738(88)90118-1

Bibi, S. S., Rehman, A., Minhas, R. A., & Janjua, S. (2015). Evaluation of DNA extraction method from formalin preserved skin samples of Panthera pardus for molecular genetic assessment. The Journal of Animal & Plant Sciences, 25, 1196–1199.

Boyle, E. E., Zardus, J. D., Chase, M. R., Etter, R. J., & Rex, M. A. (2004). Strategies for molecular genetic studies of preserved deep-sea macrofauna. Deep Sea Research Part I: Oceanographic Research Papers, 51(10), 1319–1336. https://doi.org/10.1016/j.dsr.2004.04.003

Burrell, A. S., Disotell, T. R., & Bergery, 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

Campos, P. F., & Gilbert, T. M. P. (2012). DNA extraction from formalin-fixed material. Ancient DNA: Methods and Protocols, 840, 81–85. https://doi.org/10.1007/978-1-61779-516-9_11

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

Derkarabetian, S., Benavides, L. R., & Gribet, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: Unlocking the rest of the vault. Molecular Ecology Resources, 19(6), 1531–1544. https://doi.org/10.1111/1755-0998.13072

Díaz-Viloria, N., Sanchez-Velasco, L., & Perez-Enriquez, R. (2005). Inhibition of DNA amplification in marine fish larvae preserved in formalin. Journal of Plankton Research, 27(8), 787–792. https://doi.org/10.1093/plankt/fbi052

Do, H., & Dobrovic, A. (2015). Sequence artifacts in DNA from formalin-fixed tissues: Causes and strategies for minimization. Clinical Chemistry, 61(1), 64–71. https://doi.org/10.1373/clinchem.2014.223040

Feigin, C. Y., Newton, A. H., Doronina, L., Schmitz, J., Hipsey, C. A., Mitchell, K. J., Gower, G., Llamas, B., Soubrier, J., Heider, T. N., Menzies, B. R., Cooper, A., O’Neill, R. J., & Pask, A. J. (2018). Genome of the Tasmanian tiger provides insights into the evolution and demography of an extinct marsupial carnivore. Nature Ecology & Evolution, 2(1), 182–192. https://doi.org/10.1038/s41559-017-0417-y

Freedman, J., van Dorp, L. B., & Brace, S. (2018). Destructive sampling natural science collections: An overview for museum professionals and researchers. Journal of Natural Science Collections, 5, 21–34.

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/gkw033
and Evolution, 12(12), 2535–2551. https://doi.org/10.1093/gbe/ evaa188

Parks, M., & Lambert, D. (2015). Impacts of low coverage depths and post-mortem DNA damage on variant calling: A simulation study. BMC Genomics, 16, 19. https://doi.org/10.1186/s12864-015-1219-8

Pedersen, J. S., Valen, E., Velazquez, A. M. V., Parker, B. J., Rasmussen, M., Lindgreen, S., Lilje, B., Tobin, D. J., Kelly, T. K., Yang, S., Andersson, R., Jones, P. A., Hoover, C. A., Tikhonov, A., Prokhorkhouchouk, E., Rubin, E. M., Sandelin, A., Gilbert, M. T. P., Krogh, A., ... Orlando, L. (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. Genome Research, 24(3), 454–466. https://doi.org/10.1101/gr.163592.113

Pierson, T. W., Kieran, T. J., Clause, A. G., & Castleberry, N. L. (2020). Preservation-induced morphological change in salamanders and failed DNA extraction from a decades-old museum specimen: Implications for Plathodon ainsworthi. Journal of Herpetology, 54(2), 137–143. https://doi.org/10.1670/19-012

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(5), R47. https://doi.org/10.1186/gb-2010-11-5-r47

Rawlence, N. J., Wood, J. R., Armstrong, K. N., & Cooper, A. (2009). DNA content and distribution in ancient feathers and potent to reconstruct the plumage of extinct avian taxa. Proceedings of the Royal Society B: Biological Sciences, 276(1672), 3395–3402. https://doi.org/10.1098/rspb.2009.0755

Raxworthy, C. J., & Smith, B. T. (2021). Mining museums for historical DNA: Advances and challenges in museums. Trends in Ecology and Evolution. https://doi.org/10.1016/j.tree.2021.07.009

Robbe, P., Popitsch, N., Knight, S. J. L., Antoniou, P., Becq, J., He, M., Kanapin, A., Samsonova, A., Vavouli, D. V., Ross, M. T., Ksibi, Z., Cabes, M., Ramos, S. D. C., Page, S., Dreau, H., Ridout, K., Jones, L. J., Totoiu, C. A., Phillips, J. M., Reese, A. T., Majumdar, S., Girguis, P. R., Watanabe, M., Hashida, S., Yamamoto, H., Matsubara, T., Ohtsuka, T., Watanabe, Y., Sato, S., & Sato, K. (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. Genome Research, 24(2), 137–143. https://doi.org/10.1101/gr.163592.113

Shi, S.-R., Cote, R. J., Wu, L., Liu, C., Datar, R., Shi, Y., & Taylor, C. R. (2002). DNA extraction from archival formalin-fixed, paraffin-embedded tissue sections based on the antigen retrieval principle: Heating under the influence of pH. The Journal of Histochemistry and Cytochemistry, 50(8), 1005–1011.

Shi, S.-R., Datar, R., Liu, C., Wu, L., Zhang, Z., Cote, R. J., & Taylor, C. R. (2004). DNA extraction from archival formalin-fixed, paraffin-embedded tissues: Heat-induced retrieval in alkaline solution. Histochemistry and Cell Biology, 122(3), 211–218. https://doi.org/10.1007/s00418-002-0996-3

Shiokazi, T., Itoh, F., Hirose, Y., Onodera, J., Kuwata, A., & Harada, N. (2021). A DNA metabarcoding approach for recovering plankton communities from archived samples fixed in formalin. PLoS One, 16(2), e0245936. https://doi.org/10.1371/journal.pone.0245936

Simmons, J. E. (2014). Fluid preservation: A comprehensive reference. Rowman & Littlefield.

Smits, T. H. M. (2019). The importance of genome sequence quality to microbial comparative genomics. BMC Genomics, 20(1), 662. https://doi.org/10.1186/s12864-019-6014-5

Splendiani, A., Fioravanti, T., Giovanniotti, M., Olivieri, L., Ruggeri, P., Nisi Cerioni, P., & Caputo Barucchi, V. (2017). Museum samples could help to reconstruct the original distribution of Salmo trutta complex in Italy. Journal of Fish Biology, 90(6), 2443–2451. https://doi.org/10.1111/jfb.13307

Srinivasan, M., Sedmak, D., & Jewell, S. (2002). Effect of fixatives and tissue processing on the content and integrity of nucleic acids. The American Journal of Pathology, 161(6), 1961–1971. https://doi.org/10.1016/S0002-9440(10)64472-0

Staats, M., Erkens, R. H. J., van de Vossenberg, B., Weringa, J. J., Kraaijveld, K., Stielow, B., Geml, J., Richardson, J. E., & Bakker, F. T. (2013). Genomic treasure troves: Complete genome sequencing of herbarium and insect museum specimens. PLoS One, 8(7), e69189. https://doi.org/10.1371/journal.pone.069189

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., Lyra, M. L., Pajimans, J. L. A., Preick, M., Basler, N., Penner, J., & Hofreiter, M. (2021). Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens. Molecular Ecology Resources, 21(7), 2299–2315. https://doi.org/10.1111/1755-0998.13433

Totoiu, C. A., Phillips, J. M., Reese, A. T., Majumdar, S., Girguis, P. R., Raston, C. L., & Weiss, G. A. (2020). Vortex fluidics-mediated DNA rescue from formalin-fixed museum specimens. PLoS One, 15(1), e0225807. https://doi.org/10.1371/journal.pone.0225807

Tsai, W. L. E., Schedl, M. E., Maley, J. M., & McCormack, J. E. (2020). More than skin and bones: Comparing extraction methods and alternative sources of DNA from avian museum specimens. Molecular Ecology Resources, 20(5), 1220–1227. https://doi.org/10.1111/1755-0998.13077

Watanabe, M., Hashida, S., Yamamoto, H., Matsubara, T., Ohtsuka, T., Suzawa, K., Maki, Y., Soh, J., Asano, H., Tsukuda, K., Tooyooka, S., & Miyoshi, S. (2017). Estimation of age-related DNA degradation from formalin-fixed and paraffin-embedded tissue according to the extraction methods. Experimental and Therapeutic Medicine, 14(3), 2683–2688. https://doi.org/10.3892/etm.2017.7497

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag.

Williams, C., Pontén, F., Moberg, C., Söderkvist, P., Uhlén, M., Pontén, J., & Arumugam, M. (2011). A high frequency of sequence alterations is due to formalin fixation of archival specimens. The American Journal of Pathology, 178(9), 2700–2707. https://doi.org/10.1016/j.ajpath.2011.06.034

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biology, 20(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, H. M., Gonzalez, V. L., Lloyd, M., Coddington, J., & Scharff, N. (2018). Next-generation museum genomics: Phylogenetic relationships among palomino spiders using sequence capture techniques.
(Araneae: Palpimanoidea). Molecular Phylogenetics and Evolution, 127, 907–918. https://doi.org/10.1016/j.mpev.2018.06.038

Yang, D. Y., Eng, B., Waye, J. S., Dudar, J. C., & Saunders, S. R. (1998). Improved DNA extraction from ancient bones using silica-based spin columns. American Journal of Physical Anthropology, 105(4), 539-543.

Zacho, C. M., Bager, M. A., Margaryan, A., Gravlund, P., Galatius, A., Rasmussen, A. R., & Allentoft, M. E. (2021). Uncovering the genomic and metagenomic research potential in old ethanol-preserved snakes. PLoS One, 16(8), e0256353. https://doi.org/10.1371/journal.pone.0256353

Zimmermann, J., Hajibabaei, M., Blackburn, D. C., Hanken, J., Cantin, E., Posfai, J., & Evans, T. C. Jr (2008). DNA damage in preserved specimens and tissue samples: A molecular assessment. Frontiers in Zoology, 5, 18. https://doi.org/10.1186/1742-9994-5-18

---

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Hahn, E. E., Alexander, M. R., Grealy, A., Stiller, J., Gardiner, D. M., & Holleley, C. E. (2022). Unlocking inaccessible historical genomes preserved in formalin. Molecular Ecology Resources, 22, 2130–2147. https://doi.org/10.1111/1755-0998.13505