Received: 10 February 2022 | Accepted: 27 April 2022

DOI: 10.1111/2041-210X.13945

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Leveraging natural history collections to understand the impacts of global change

First large-scale quantification study of DNA preservation in insects from natural history collections using genome-wide sequencing

Victoria E. Mullin[1,2] | William Stephen[1] | Andres N. Arce[3,4] | Will Nash[5] | Calum Raine[5] | David G. Notton[6] | Ashleigh Whiffin[6] | Vladimir Blagderov[6] | Karim Gharbi[5] | James Hogan[7] | Tony Hunter[8] | Naomi Irish[5] | Simon Jackson[9,10] | Steve Judd[8] | Chris Watkins[5] | Wilfried Haerty[5] | Jeff Ollerton[11] | Selina Brace[1] | Richard J. Gill[3] | Ian Barnes[1]

[1]Department of Earth Sciences, The Natural History Museum, London, UK; [2]Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland; [3]Georgina Mace Centre for The Living Planet, Department of Life Sciences, Silwood Park, Imperial College London, Berks, UK; [4]School of Engineering, Arts, Science & Technology, University of Suffolk, Ipswich, UK; [5]The Earlham Institute, Norwich Research Park, Colney Lane, Norwich, UK; [6]National Museum of Scotland, Edinburgh, UK; [7]Oxford University Museum of Natural History, Oxford, UK; [8]World Museum Liverpool, Liverpool, UK; [9]Tullie House Museum and Art Gallery Trust, Carlisle, UK; [10]Ipswich Museum (Colchester and Ipswich Museums), Ipswich, UK and [11]Faculty of Arts, Science and Technology, University of Northampton, Northampton, UK

Correspondence
Victoria E. Mullin
Email: mullinve@tcd.ie

Ian Barnes
Email: i.barnes@nhm.ac.uk

Funding information
Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/CCG1720/1, BB/CSP1720/1, BBS/E/T/000PR9816 and BBS/E/T/000PR9818; Irish Research Council, Grant/Award Number: GOIPD/2020/605; Natural Environment Research Council, Grant/Award Number: NE/P012574/1 and NE/P012914/1; Natural History Museum, Grant/Award Number: Earth Sciences Departmental Infrastructure Fund

Handling Editor: Natalie Cooper

Abstract

1. Insect declines are a global issue with significant ecological and economic ramifications. Yet, we have a poor understanding of the genomic impact these losses can have. Genome-wide data from historical specimens have the potential to provide baselines of population genetic measures to study population change, with natural history collections representing large repositories of such specimens. However, an initial challenge in conducting historical DNA data analyses is to understand how molecular preservation varies between specimens.
2. Here, we highlight how Next-Generation Sequencing methods developed for studying archaeological samples can be applied to determine DNA preservation from only a single leg taken from entomological museum specimens, some of which are more than a century old. An analysis of genome-wide data from a set of 113 red-tailed bumblebee *Bombus lapidarius* specimens, from five British museum collections, was used to quantify DNA preservation over time. Additionally, to improve our analysis and further enable future research, we generated a novel assembly of the red-tailed bumblebee genome.
3. Our approach shows that museum entomological specimens are comprised of short DNA fragments with mean lengths below 100 base pairs (BP), suggesting

Victoria E. Mullin and William Stephen contributed equally to the work.

a rapid and large-scale post-mortem reduction in DNA fragment size. After this initial decline, however, we find a relatively consistent rate of DNA decay in our dataset, and estimate a mean reduction in fragment length of 1.9 bp per decade. The proportion of quality filtered reads mapping to our assembled reference genome was around 50%, and decreased by 1.1% per decade.

4. We demonstrate that historical insects have significant potential to act as sources of DNA to create valuable genetic baselines. The relatively consistent rate of DNA degradation, both across collections and through time, mean that population-level analyses—for example for conservation or evolutionary studies—are entirely feasible, as long as the degraded nature of DNA is accounted for.

KEYWORDS

aDNA, *Bombus*, collection genomics, DNA degradation, entomological collections, historical DNA, museum specimen, pollinators

1 | INTRODUCTION

Determining changes in genome diversity is a central component of biodiversity monitoring (Jensen et al., 2022). Quantifying such change can reveal past and ongoing demographic processes contributing to our understanding of species and population genetic health and resilience (Morin et al., 2021). Furthermore, it can help determine how populations respond to environmental pressures (Pinsky et al., 2021), assess species adaptive potential and identify signatures of selection representing key adaptations (Colgan et al., 2022). However, to accurately determine the extent and rate of changes in genetic diversity, we must first establish historical baselines (Díez-del-Molino et al., 2018), which we lack for many taxonomic groups.

Loss of important functional groups, such as insect pollinators, has significant implications for wildflower health (Brosi & Briggs, 2013; Ollerton et al., 2014), crop pollination and subsequent food security (Eilers et al., 2011; Gill et al., 2016; Potts et al., 2016). Yet, to date we have a poor understanding of how genetic diversity has changed for the vast majority of wild insect pollinator populations over the past century, a period of both dramatic, large-scale land use change that has reduced pollinator diversity in the UK (Ollerton et al., 2014) and estimated global genetic diversity loss for wild species (Leigh et al., 2019). One opportunity to explore changes over this time period is to create genetic baselines from historical collections.

Museum collections are increasingly being used to provide information relevant to the conservation of biodiversity, above and beyond their function in systematic biology (Raxworthy & Smith, 2021). A challenge, however, is the DNA preservation within pinned and open-dried specimens, as post-mortem degradation reduces the quantity and quality of DNA retained in collections of historic specimens (Heintzman et al., 2014). A possible solution is to utilise Next-Generation Sequencing (NGS) approaches developed for retrieving and analysing ancient DNA (aDNA). These techniques have the potential to be an invaluable tool in determining historical genetic baselines for insect species, and subsequently provide an understanding of evolutionary and demographic processes that would not otherwise be possible. However, to best utilise and safeguard these historical collections, we must understand the extent to which age and/or storage conditions impact the rate of DNA degradation.

While there are no absolute standards for the measure of DNA preservation in museum specimens, several commonly used metrics to summarise the length, amount and authenticity of the sequence reads obtained:

1. Endogenous DNA content—the proportion of the sequenced reads that align to the reference genome. Here we refer to the endogenous percentage that aligns to the reference post quality filtering. This is an important factor in determining the required sequencing depth in a collections-based genomics project, and the ability to predict this in advance will inform the overall financial cost of the project. It is important to note that this metric is also impacted by the reference genome used, with increasing evolutionary distance between the target and the reference reducing the likelihood of mapping sequenced reads.
2. DNA fragmentation—mean read length is used as a proxy for the mean length of DNA molecules in the sample and a common method to determine the extent of fragmentation. Genome sequencing runs are typically described in terms of the number of base pairs sequenced, assuming that each read is the maximum length that can be sequenced. Therefore, the shortfall in information content per read is also an important consideration in determining project cost. Additionally, mean read length is an overestimate of the true mean fragment length, as the shortest DNA fragments are less likely to be recovered during DNA extraction and short sequencing reads are not aligned to the reference genome to prevent spurious alignment. However, as DNA fragmentation is a random, time-dependant process following an exponential decay model (Lindahl & Andersson, 1972; Lindahl

& Nyberg, 1972), the distribution of fragment length frequencies in the sample can be fitted as an exponential curve (Adler et al., 2011; Deagle et al., 2006; Schwarz et al., 2009). Here, we also explore whether the 'lambda' parameter (λ), which describes the shape of the exponential distribution, provides a less biased estimate of the mean fragment length.

3. DNA damage—the extent of chemical modification of the DNA molecules in the sample, with the most common form being hydrolytic deamination of cytosine molecules to form uracil, which is subsequently sequenced as thymine (Binladen et al., 2006; Brotherton et al., 2007; Gilbert, Hansen, et al., 2003; Gilbert, Willerslev, et al., 2003; Hansen et al., 2001; Sawyer et al., 2012; Schwarz et al., 2009). This common process occurs primarily towards 5′-ends of fragmented DNA molecules, and the impact on sequence accuracy is minimised (as in this dataset) by treatment with uracil-DNA glycosylase (UDG) to cleave the DNA at uracils (Hofreiter et al., 2001). This process leaves a proportion of deaminated sites at 3′- and 5′-terminal positions (Briggs et al., 2010), which can be measured to infer general patterns, and explore the extent to which this treatment is required.
4. Molecular preservation—previous studies have been able to detect the residual impact of nucleosome packaging on the degradation pattern of DNA extracted from historical and archaeological samples, due to the observation of periodic 'spikes' in the quantity of DNA at specific lengths (Kistler et al., 2017; Pedersen et al., 2014). In this paper, we use the strength of this pattern, the histone periodicity index, as an alternative measure of DNA preservation.
5. Complexity—the proportion of reads within the dataset that represent novel sequence information, and are not simply PCR-amplification duplicates of other reads. Absolute measurement of complexity is made difficult by variation of library construction efficiency, due to batch variation in laboratory procedures and amplification efficiency (Head et al., 2014). However, a relative estimate can be made where a set of samples have been prepared in the same way. Complexity can be quantified either by measuring coverage for a given sequencing effort, to provide an estimate of the number of times unique reads capture the whole genome, or by estimating the required sequencing depth which would exhaust the information content of a library. If the molecular complexity of a given sequencing library is exhausted before the required depth-of-coverage is achieved, new libraries or extracts will be required, again increasing the project cost.

Here, we use these different measures of molecular preservation to explore DNA decay within a time series dataset of museum collection specimens of the red-tailed bumblebee *Bombus lapidarius*, collected in the UK over the last 130 years. This large dataset enables us to overcome some of the issues inherent in previous studies, as this is a large dataset generated in a consistent manner, by a single laboratory, element and species. This dataset, coupled with a novel genome assembly for *B. lapidarius*, allows us to consider differences in age, and storage location (museum collection), while investigating variation in endogenous DNA content (1), DNA damage and preservation (2–4) and sequence complexity (5).

2 | MATERIALS AND METHODS

2.1 | Museum specimen

This study used an individual leg per specimen for 113 pinned *B. lapidarius* drones (haploid males) collected from 1891 to 2004 housed in the collections of five British institutions; Natural History Museum (NHM London), Oxford Museum of Natural History, Tullie House Museum and Art Gallery (Carlisle), World Museum (Liverpool) and National Museums Scotland (Edinburgh). All specimens were treated as degraded DNA specimens with all pre-PCR laboratory work taking place in the aDNA laboratory in the NHM London.

2.2 | Museum specimen DNA extraction and library preparation

A version of the extraction protocol described in Dabney et al. (2013) was performed on each leg separately, with the following modifications. In the lysis stage, 180 µl Qiagen ATL Buffer for tissue lysis and 20 µl Proteinase K were added to each leg and heated at 56°C for 24 hr. DNA purification followed Dabney et al. (2013) with the modification stated in Brace et al. (2019), replacing the Zymo-Spin V column binding apparatus with the extender assembly from the High Pure Viral Nucleic Acid Large Volume Kit (Roche). DNA extracts were treated with USER enzyme; 20 or 30 µl of extract with 5 µl of USER enzyme for 3 hr at 37°C. Double indexed double stranded libraries were built following Meyer and Kircher (2010) with AmpliTaq Gold polymerase for the PCR amplification with PCR cycles varying from 13 to 20. All sequencing was performed on an Illumina NextSeq 500 (75 bp PE) at the NHM London sequencing centre.

2.3 | Reference genome assembly: Bombus_lapidarius_EIv1

High molecular weight (HMW) DNA was extracted from the mesosoma of a single male *B. lapidarius* (Table S1; Figure S1) collected on 23 August 2019 at the Earlham Institute (TG179075; Lat: 52.622282, Long: 1.2190789). The specimen was snap frozen on dry ice following collection and stored at −80°C until DNA extraction. Extractions were conducted using the Qiagen MagAttract HMW DNA kit, with modifications (see Supplementary Information). DNA concentration was measured using the Qubit HS kit and the Nanodrop was used to measure extraction purity. The distribution of HMW DNA fragment sizes was measured using the Agilent FEMTO Pulse instrument.

HiFi library preparation and Pacific Biosciences sequencing were carried out by Genomics Pipelines at the Earlham Institute (see Supplementary Information).

Hifi reads were extracted from the raw Pacific Biosciences output by the Earlham Institute core bioinformatics group using the Pacific Biosciences SMRTlink pipeline (v10.1.0.119588). Prior to assembly, Hifi reads were trimmed for adapter sequences using Cutadapt (v3.2, Martin, 2011). Trimmed read statistics were generated with seqkit (v0.10.0, Shen et al., 2016). Genome assembly was conducted using the Hifiasm assembler (Cheng et al., 2021). As the sequenced individual was haploid, Hifiasm was run without duplication purging (`-l 0`). The mitochondrial genome was identified using Mitofinder (v1.4.1, Allio et al., 2020). Contaminant contigs were identified using Kraken2 (v2.0.7, Wood et al., 2019) and blobtools (v1.1.1, Laetsch & Blaxter, 2017). Assembly completeness was assessed with BUSCO (v5.0.0, Manni et al., 2021) using hymenoptera_odb10.

2.4 | Museum specimen sequence alignment

Sequence quality was inspected via FastQC v0.11.8 (www.bioinformatics.babraham.ac.uk/projects/fastqc/). AdapterRemoval v2.2.4 (Schubert et al., 2016) was implemented to trim adapter sequences, collapse overlapping read pairs with a minimum 11 bp overlap, filter for a minimum read length of 25 bp and trim Ns and low-quality bases. Trimmed collapsed reads were aligned to the Bombus_lapidarius_Elv1 genome, using bwa aln (v0.7.12-r1039, Li et al., 2009; -l 1024 -n 0.01 -o 2 -q 15) and BAM files sorted with SAMtools (v1.12, Li et al., 2009). Picard (v2.18.7) was implemented to mark and remove duplicates and merge BAM files. Mapping quality filtering (q 30) was performed with SAMtools (v1.12). To further explore DNA fragment size only, we ran a paired-end alignment of the adapter trimmed non-collapsed reads using bwa aln (-l 1024 -n 0.01 -o 2 -q 15), filtered for duplicates and mapping quality through the same pipeline as the collapsed reads and further filtered for properly paired reads using SAMtools (-f 2 -F8).

2.5 | Quantification of endogenous DNA content

The endogenous percentage of total sequencing reads was calculated from the collapsed reads alignment after filtering for duplicates and mapping quality. Additionally, we calculated the 'raw' endogenous percentage of total sequencing reads prior to any filtering.

2.6 | DNA fragmentation: Read length distribution

For the original collapsed alignments, the read length distributions were extracted with SAMtools (v1.12), from which the mean and the median read length were calculated. Additionally, the insert size distributions of non-collapsed read alignments were extracted (SAMtools stats) to visualise how many endogenous reads are potentially removed at the collapsing stage during adapter trimming. A cut-off of 200 bp maximum insert size was applied to remove spurious alignment results. Finally, to achieve a more accurate measure of DNA fragmentation, we combined the distributions of the read length (collapsed alignment) and insert size (non-collapsed alignment), referred to here as the combined length. The mean combined length was also calculated using this combined length distribution.

2.7 | DNA fragmentation: λ parameter estimation

Read length distributions can be quite different to the true fragment length distributions of DNA present in the sample, due to limitations in DNA extraction library construction efficiency, and post-sequencing data processing (Kistler et al., 2017). In a standard aDNA pipeline, longer reads that do not overlap cannot be merged and those shorter than 25 bp are not aligned, resulting in artefactually low frequencies in the distribution. To determine whether these processes impact all samples equally, and to better estimate the mean fragment size in the libraries we applied a method developed by Kistler et al. (2017). This method attempts to mitigate these artefacts by estimating the λ from the portion of the distribution that best fits an exponential distribution, via maximum likelihood. For each sample, the read length density distributions, collapsed alignment only, were used to estimate the λ parameter using the protocol developed by Kistler et al. (2017), using the 'lambdaCalc.pl' and 'readLengthLambda.R' scripts supplied in their Supplementary Information. First, the distributions are checked for peaks at the highest read lengths, which would indicate reads longer than the maximum readable. Then the λ parameter is estimated for every size range going backwards from the longest read, selecting the value with the highest likelihood. The λ parameter was then used to estimate the expected 'true' mean fragment length (μ) for each distribution, using the formula $\mu = 1/\lambda$.

2.8 | DNA damage quantification

The DNA extracts in this study were enzymatically treated to reduce the impact of deamination on the resulting data. The proportions of 5' C-T and 3′ G-A transitions in the first base position were estimated using mapDamage 2.0 (Jónsson et al., 2013).

2.9 | Molecular preservation: Histone periodicity index

As a result of the intermittent protection of lengths of DNA from fragmentation by histones, the resulting read length density distribution would show multiple local peaks as some fragment lengths would be more common than others. To visualise this, read length and combined length distributions were plotted. The protocol developed by Kistler et al. (2017), using the 'histoneCalc.R' script supplied in their Supplementary Information, was used to estimate the intensity of any identified local peaks in the distribution of read lengths from the collapsed alignment only, the histone periodicity index.

2.10 | Library complexity

The preseq (v3.1.2, Daley & Smith, 2013) function 'lc_extrap' (-B) was used to (1) estimate the maximum number of unique reads available in the library by extrapolating the number of unique aligned reads at greater sequencing depths and (2) estimate the expected number of unique reads per 10 million aligned reads. For this analysis, a reduced dataset of 74 individuals was curated to reduce confounding factors such as number of libraries sequenced per sample, quantity of DNA input into library and number of PCR cycles (Table S3). Each individual is represented by one library built with 20 μl of DNA extract, though PCR cycles vary (Table S3). BAM files were filtered for mapping quality (q 30), with no filtering for duplicates, and to mitigate variation in genome coverage, subsampled for 1 million reads using SAMtools (v1.12). Preseq failed to run for five samples due to lack of sufficient variation within the subsampled BAM. The proportion of the total expected available unique reads that were sequenced from the sample was then estimated by dividing the number of unique endogenous reads by the (preseq) estimated maximum number of unique reads in the library.

2.11 | Linear models

All linear regressions were performed in RStudio (v2021.09.1 +372; RStudio Team, 2021) with R v4.1.2 (R Core Team, 2021). In the R STATS package, linear regression was conducted using the lm function and analysis of variance (ANOVA) extracted using the anova function. Model comparison was achieved via likelihood ratio tests using the lrtest function in the LMTEST package (Kuznetsova et al., 2017). Parameters explored by multiple regression analysis include: endogenous DNA %; DNA fragmentation via the mean combined length and λ parameter of the collapsed alignment; DNA damage via the proportion of 5′ first base C-T transitions; molecular preservation via the histone periodicity index; and library complexity via the preseq-estimated maximum number of unique reads in the library. Multiple linear regression models used year of specimen collection and museum collection as explanatory variables, allowing the model intercept to vary by museum to detect variation between museum collections while accounting for variation in time since specimen death. An interaction between museum and year of collection was not investigated to test for a temporal pattern across specimens. The Natural History Museum London (NHM) was used as the reference group for intercept comparisons as they constituted the greatest sample size ($n = 40$) and specimen age range (1892–2002). Temporal analyses assumed linear relationships with sample age as DNA deamination and fragmentation by depurination occur at constant rates (Briggs et al., 2007; Lindahl & Nyberg, 1974; Lindahl & Andersson, 1972; Lindahl & Nyberg, 1972). As environmental variables are hypothesised to affect the rate of damage and fragmentation in ways which are poorly understood, further models were constructed with an exponential relationship by log-transforming each parameter (Table S4).

3 | RESULTS

3.1 | Contemporary

3.1.1 | Reference genome assembly

Approximately 15× coverage of Hifi reads were generated (Table S2). Following the assessment of the primary assembly with Kraken2, blobtools and MitoFinder, 157 contaminant and mitochondrial contigs (5.7 Mb) were removed. The resulting assembly spans 1,473 contigs, has a contig N50 of 2.3 Mb and a longest contig of 14 Mb. The assembly represents 97.1% of the hymenoptera_odb10 BUSCO set complete and single copy, with 0.4% duplicated, 0.4% fragmented and 2.1% missing.

3.2 | Historical

3.2.1 | Endogenous DNA content

Endogenous DNA recovery varied across the 113 *B. lapidarius* specimens (Figure 1). However, the majority of specimens contained

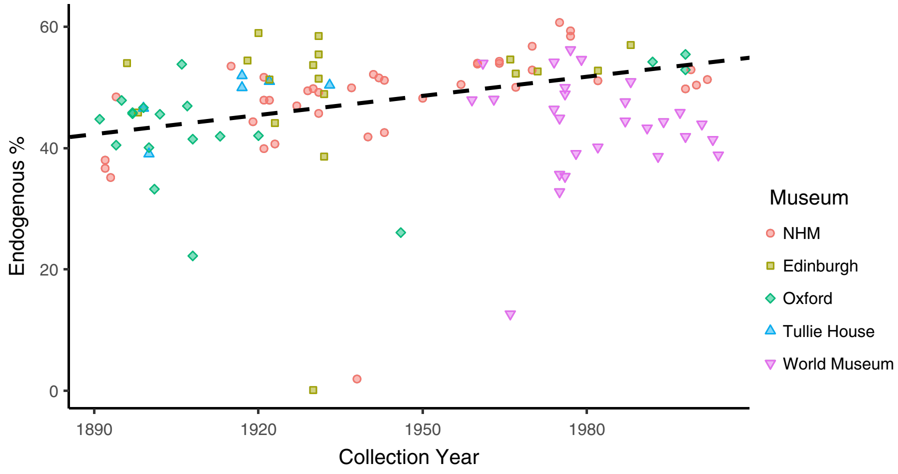

FIGURE 1 Endogenous content decreases with time since death. Endogenous percent (post-filtering) of 110 *Bombus lapidarius* specimens in relation to the year of collection. Best fit line calculated from multiple regression of endogenous percent (Table S4) as a dependent of collection year and museum of origin with the equation $y \sim 0.11x - 156.63$. Model intercept calculated for the reference group (NHM).

relatively high levels of endogenous DNA post filtering, with a dataset median of 48.20% and a range of 0.12–60.68% (Table S3). Only two specimens contained <5% endogenous DNA. As the sequencing effort for three samples was relatively low, the 110 samples with >1 million read pairs sequenced were included in further analyses. Endogenous percent positively correlates with year of collection (adj. $R^2 = 0.10$; $p < 0.01$). For every year a specimen aged, the percentage of endogenous DNA decreased by 0.11%, equating to a 12.43% reduction over the 113 years the study spans (Figure S4). The model intercept was consistent across museums ($p > 0.40$) except for the World Museum, which had an intercept significantly lower than the NHM London reference by 8.17% ($p < 0.01$). Assuming this relationship remains constant, the model formula $y \sim 0.11x - 156.63$ suggests that it may be possible to sequence endogenous DNA from pinned insects collected in any year after 1488, which includes all known collections.

3.3 | DNA fragmentation

To examine variation in DNA fragmentation, we first derived the combined density distribution of read length (collapsed alignment; Figure S3) and insert (non-collapsed alignment) lengths for each sample (Figure 2a). We then compared the variation in the mean combined length (range 44–87 bp; Table 1) through time and across different collections (Figure 2b). Mean combined length positively correlates with year of collection (adj. $R^2 = 0.84$; $p < 0.01$). Every year since collection, mean combined length decreased by 0.19 bp, equating to a decrease of 21 bp over the 113 years. The model intercept was consistent across museums ($p > 0.05$) except for the World Museum, which had a 9.41 bp lower intercept than the reference, NHM ($p < 0.01$; Table S4).

We derived the λ parameter of the read length distribution (collapsed alignment only, Figure S3) and used this to calculate the mean value for the fitted distribution, equivalent to the estimated true mean fragment length. Estimated mean fragment length positively correlated with mean combined length (adj. $R^2 = 0.85$; $p < 0.01$; Figure 3a; Table S4). λ was negatively correlated with collection year (Adj. $R^2 = 0.61$; $p < 0.01$), indicating an increase in fragmentation with time since collection (Figure 3b). λ increased by 0.00046 per year after collection, with an expected increase of 0.052 over the 113 years the study spans, equivalent to a 19.28 bp decrease in estimated mean fragment size.

3.4 | DNA deamination

The proportion of 5′ C-T deaminated sites in the first base position was estimated for 110 samples and plotted against specimen age (Figure 4), which showed no relationship ($p = 0.961$). However, the model did demonstrate significant variation between museums after accounting for year of collection ($F[4,104] = 5.41$; $p < 0.01$; adj. $R^2 = 0.13$; Table S4).

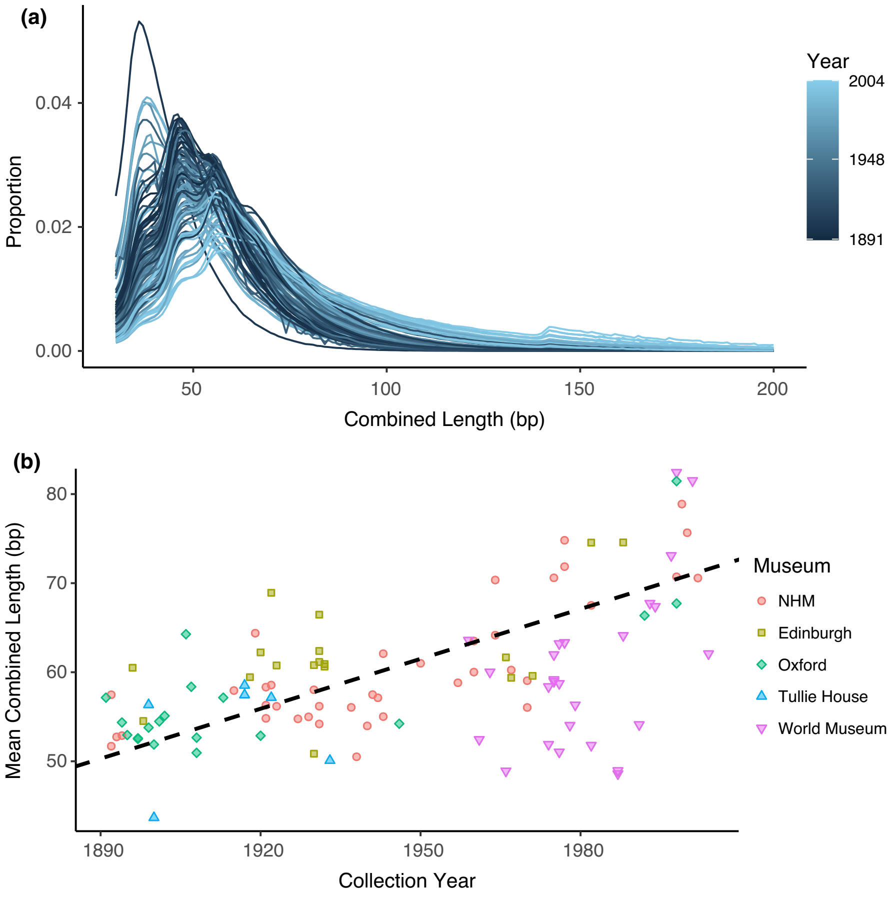

FIGURE 2 Mean combined length decreases with time after death. (a) Combined length (read length + insert size) density distributions for all 110 *Bombus lapidarius* samples. Length is given in base pairs (bp). Colour represents the year of collection. (b) Mean combined length against year of specimen collection. Best fit line calculated from multiple regression of mean combined length (Table S4) as a dependent of collection year and museum with the equation $y \sim 0.19x - 302.58$. Model intercept calculated for the reference group (NHM).

TABLE 1 The range and median values across all samples across the dataset for the three fragmentation parameters calculated for each specimen. Mean fragment length estimate = $1/\lambda$. All values given in base pairs (bp) and rounded to the nearest integer

Estimate	Minimum (bp)	Maximum (bp)	Median across all samples (bp)
Mean combined length	44	87	59
Median combined length	41	76	56
Mean fragment length estimate	9	40	15

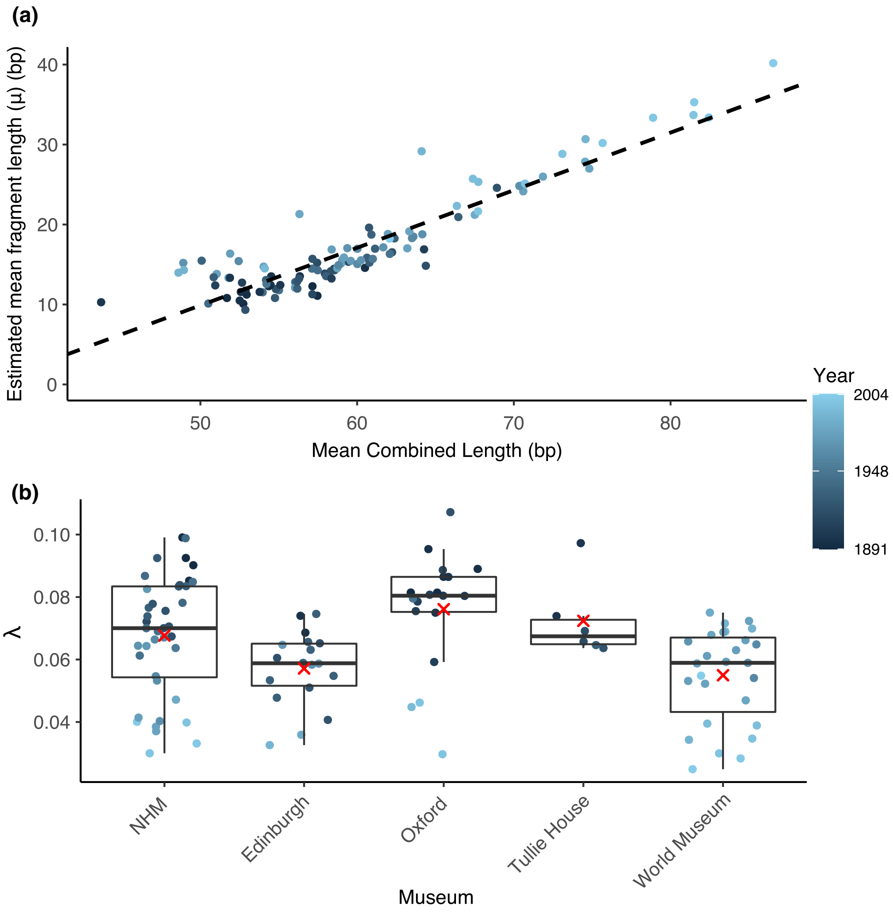

FIGURE 3 λ decreases with time since death, and the corresponding estimated mean fragment length is highly correlated with mean read length. (a) Estimated mean fragment length against mean combined length. Best fit line calculated from linear regression (Table S4) of estimated mean fragment length (μ) as a dependent of mean combined length, with the equation $y \sim 0.72x - 26.12$. Lengths are in base pairs. (b) Variation in the λ parameter of the read length distribution (Figure S3) within and between different museums. Boxes span the first and third quartiles; inner lines = median values; red crosses = mean values.

3.5 | Molecular preservation

We measured the extent of periodicity in our dataset (Figure 5), detecting histone fragmentation bias in the majority of samples (102/110). Histone periodicity index was not significantly influenced by collection year (p = 0.07). The model did, however, demonstrate significant variation between museums after accounting for year of collection (F[4,104] = 4.566; $p < 0.01$; adj. $R^2 = 0.18$), with the NHM demonstrating a significantly higher intercept than the other museums (Table S4).

3.6 | Library complexity

The maximum number of unique reads in the library was successfully estimated using preseq from a one million read subsample in 69 specimens (Figure 6). The maximum number of unique reads is positively correlated with collection year (estimate = 619,700; $p = 6.84 \times 10^{-3}$), although there is a high degree of unexplained variability in the datasets (adj. $R^2 = 0.06$).

4 | DISCUSSION

Here we highlight an aDNA methodological approach to enable sufficient DNA extraction to conduct NGS shotgun sequencing from just a single insect [bumblebee] leg for specimens as old as 113 years. We demonstrate the potential to use museum specimens to investigate genome-level changes over time, and our findings suggest even older specimens could be studied. By studying a large number of individuals, and using a species-specific reference genome, we were able to quantify the rate and extent of change in several important parameters, including decreases in endogenous DNA content, DNA fragmentation and the estimated total unique reads as a proxy for complexity. We also fitted an exponential function to the read length distribution to estimate the mean fragment length, and

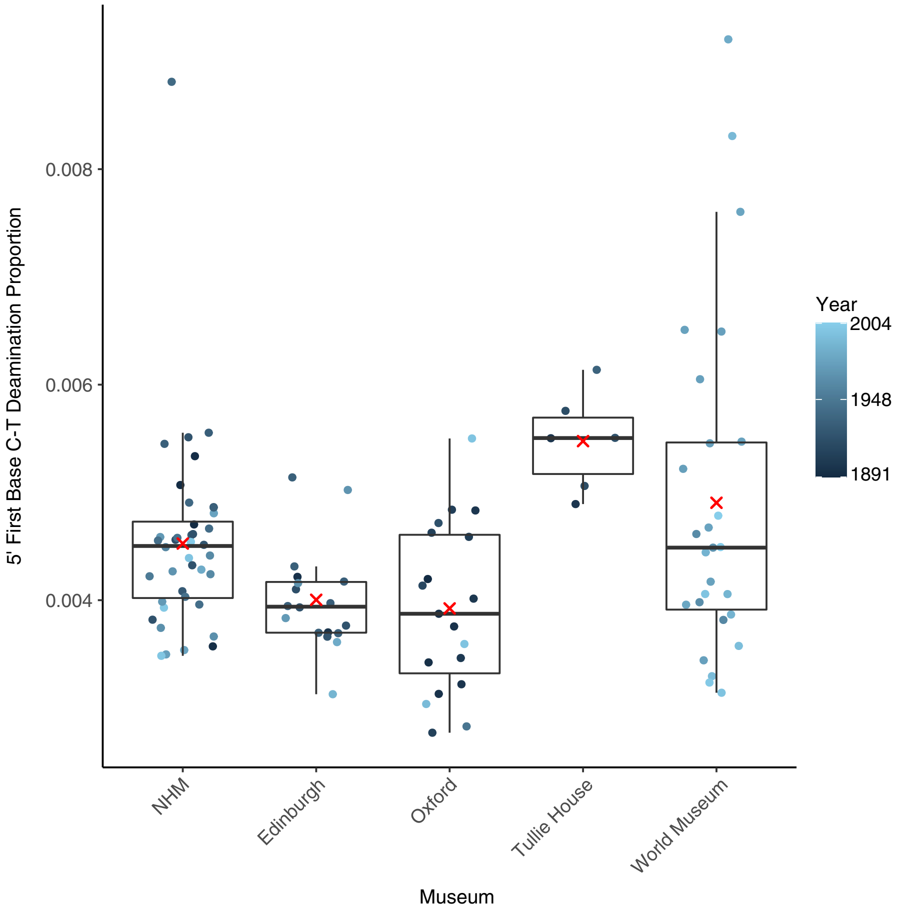

FIGURE 4 Deamination varies between different museums but not with time since death. Data show variation in the proportion of first base position sites with 5′ C to T deamination within and between different museums. Boxes span the first and third quartiles; inner lines = median values; red crosses = mean values.

found this co-varied with the mean of the combined length (read length and insert size). The metrics used in this study can be used to generate baseline information to inform sample numbers, sequencing depth and other aspects of project design. Crucially, our results demonstrate the ability to study historic terrestrial arthropod collections in unprecedented genomic detail, allowing insights into the past to improve our understanding of eco-evolutionary responses to past and future pressures.

Most of the DNA fragments we observe in this study are below 100 bp, even in the more recent samples from 2004. This suggests that DNA fragmentation must occur rapidly after death in insect specimens that are part of museum collections (which tend to be dry-pinned). DNA fragmentation after this initial rapid reduction, however, appears to occur at a much more gradual rate, and this was broadly consistent across different collections. This highlights that specimen age remains an important factor in the extent of DNA degradation, and by extension the availability of information in museum specimens. Similar fragmentation patterns have been demonstrated for museum beetle specimens (Heintzman et al., 2014) and explain why PCR amplification approaches with long insert sizes between primers have struggled to amplify DNA from historical insect specimens (van Houdt et al., 2010; Ugelvig et al., 2011; Andersen & Mills, 2012). We find that the estimated true DNA fragment length and mean combined length are highly correlated, demonstrating that deriving λ provides an accurate estimate for the true mean fragment length even when only collapsed reads are used, as standard in aDNA bioinformatic pipelines.

Base deamination, especially C-T transitions at 5′ fragment ends, accumulates post mortem and is a common diagnostic of aDNA (Briggs et al., 2007; Sawyer et al., 2012). As a result, UDG is commonly used in aDNA studies to aid mapping of damaged reads (Briggs et al., 2010; Hofreiter et al., 2001) and was applied to samples in this study. This removes uracils from the sample leaving only a small proportion (Briggs et al., 2010), which explains the limited evidence of post mortem deaminated sites in this study. We note that it is possible that estimations of fragmentation were impacted by the action of UDG cleaving deaminated sites, affecting read and insert length distributions. However, our specimens are relatively young to have accumulated base deamination, and deamination is more common at strand ends (Sawyer et al., 2012), where UDG would have a negligible impact on read length. It is therefore unlikely that this would have a significant effect on estimated fragmentation rates, and so the trends identified here are likely due to post mortem fragmentation.

Overall, the results demonstrate consistency and predictability of DNA preservation across the five museums. This indicates current curatorial practises are favourable for degraded DNA survival, with most variation in endogenous DNA and fragmentation explained by differences in specimen age. Variation between individuals and museum collections may be accounted for by differences

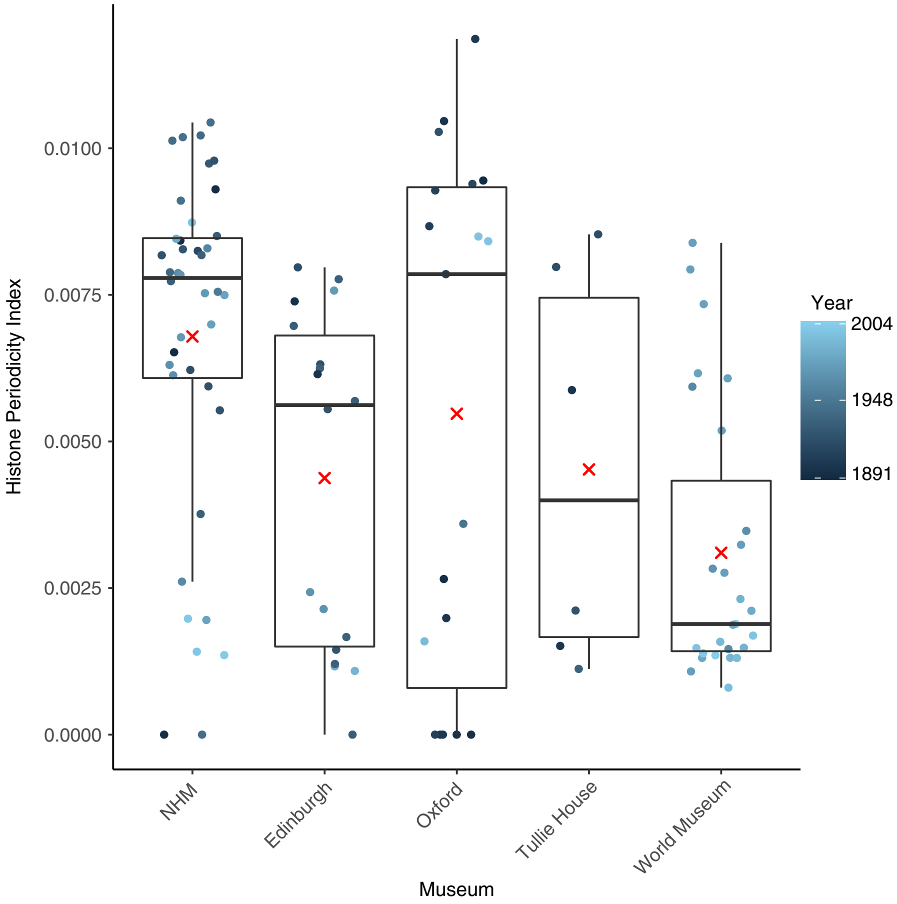

FIGURE 5 Periodicity varies between different museums but not with specimen age. Histone periodicity index is a measure of the intensity of local peaks in the read length distribution (Figure S3) due to periodic protection of DNA from fragmentation due to association with histones (Kistler et al., 2017). Boxes span the first and third quartiles; inner lines = median values; red crosses = mean values.

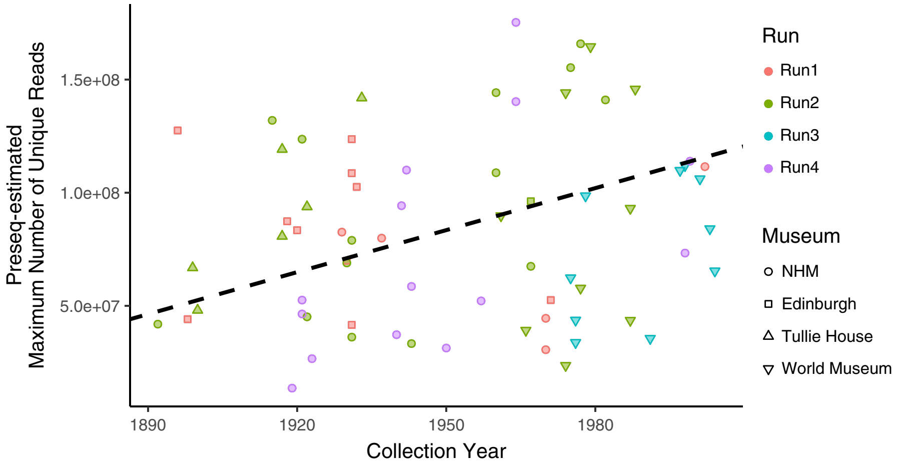

FIGURE 6 Estimated total number of unique reads decreases with time after death. (a) Preseq-estimated total number of unique reads from 69 *Bombus lapidarius* samples against year of collection. Best fit line calculated from multiple regression of estimated total number of unique reads as a dependent of collection year and museum with the equation $y \sim 6.20 \times 10^5 \times - 1.13 \times 10^9$. Model intercept calculated for the reference group (NHM). Data generated from specific sequencing runs (1–4) are indicated; samples sequenced on specific runs were processed identically (Table S3).

in curatorial techniques both during the specimens' time within the collection (museum-specific signatures) and prior to museum collection deposition. For example, variation in histone signal intensity and deamination was not significant over time, but between collections, suggesting that storage conditions may impact the extent of histone-associated fragmentation bias. Kistler et al. (2017) suggested that the histone periodicity index is temperature dependent.

The presence of sample outliers in all parameters demonstrates the unique conditions experienced by individual specimens impacts DNA degradation with variable effects on damage parameters. However, the consistency between museums in the retrieval of endogenous DNA demonstrates the feasibility of using specimens from multiple collections to allow large-scale genomic studies utilising NGS sequencing. Furthermore, of a practical note, these data provide a useful guide to museum curators when assessing the likely success of projects requiring destructive sampling of specimens.

If we are to conserve insect abundance and diversity, and the ecosystem services that they provide, in the face of ongoing environmental changes, we need to understand past processes to predict (and hopefully mitigate) future declines. The erosion of genetic diversity within populations is especially worrying given that this diversity will be fundamental to the ability of insects to adapt to climate change in particular. State-of-the-art aDNA sequencing methods have the potential to reveal insights into past genetic changes that were inconceivable just two decades ago and we hope that in the future they become a standard part of the toolkit for conservation biology.

AUTHORS' CONTRIBUTIONS

I.B., R.J.G. and S.B. conceived the overall project; I.B., R.J.G., S.B., A.N.A., V.E.M. and W.S. designed the study; V.E.M. designed the sampling strategy, with contributions from S.B., R.J.G., A.N.A., J.O. and I.B.; A.N.A., D.G.N. and A.W. sampled museum specimens for DNA analysis; Curatorial support was further provided by J.H., T.H., S.J., V.B. and S.J.; V.E.M. performed the aDNA laboratory work; W.N. collected the specimen used for genome assembly and generated HMW DNA extractions, K.G., C.W. and N.I. conducted the sequencing; C.R. conducted genome assembly and QC with input from W.H.; V.E.M. and W.S. analysed the historical data; I.B., S.B., V.E.M. and W.S. contributed to the interpretation of results. W.S., V.E.M. and I.B. wrote the paper, with contributions from all authors.

ACKNOWLEDGEMENTS

We thank the NHM digitisation team for the transcription of NHM specimen metadata and the NHM HPC team for their support. Will Nash, Calum Raine and Wilfried Haerty thank Dr Gareth Linsmith for input on genome assembly and QC strategy and acknowledge the work delivered via the Earlham Institute Scientific Computing group, as well as support for the physical HPC infrastructure and data centre delivered via the NBI Computing infrastructure for Science (Research Computing) group.

FUNDING INFORMATION

The work was funded by NERC grants NE/P012574/1 and NE/P012914/1 awarded to I.B and R.J.G. respectively, which supported A.N.A., S.B., V.E.M. and J.O. V.E.M. was additionally supported by The Government of Ireland Postdoctoral Fellowship GOIPD/2020/605. W.S. was supported by the NHM Earth Sciences Departmental Infrastructure Fund. Genome analysis was supported by the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation, through the Core Capability Grant BB/CCG1720/1 at the Earlham Institute including the National Capability BBS/E/T/000PR9816. W.H. and W.N. are supported by the BBSRC Core Strategic Programme Grant (Genomes to Food Security) BB/CSP1720/1 and its constituent work package BBS/E/T/000PR9818.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13945.

DATA AVAILABILITY STATEMENT

The Tables S3 and S4 include all the sequencing metrics and model results, and these have been deposited in the Dryad repository https://doi.org/10.5061/dryad.5mkkwh787 (Mullin et al., 2022). The primary Bombus_lapidarius_Elv1 assembly is available through the European Nucleotide Archive (ENA) under the accession number PRJEB51891 https://www.ebi.ac.uk/ena/browser/view/PRJEB51891. The raw sequencing data for the historic samples will be available after data embargo via the European Nucleotide Archive (ENA) under the accession number PRJEB52125 https://www.ebi.ac.uk/ena/browser/view/PRJEB52125.

ORCID

Victoria E. Mullin https://orcid.org/0000-0002-2604-2976
William Stephen https://orcid.org/0000-0002-3807-7391
Andres N. Arce https://orcid.org/0000-0002-3577-2110
Will Nash https://orcid.org/0000-0002-6790-1167
Calum Raine https://orcid.org/0000-0002-9609-4739
David G. Notton https://orcid.org/0000-0002-8933-7915
Ashleigh Whiffin https://orcid.org/0000-0002-2143-2246
Karim Gharbi https://orcid.org/0000-0003-1092-4488
Wilfried Haerty https://orcid.org/0000-0003-0111-191X
Jeff Ollerton https://orcid.org/0000-0002-0887-8235
Selina Brace https://orcid.org/0000-0003-2126-6732
Richard J. Gill https://orcid.org/0000-0001-9389-1284
Ian Barnes https://orcid.org/0000-0001-8322-6918

REFERENCES

Adler, C. J., Haak, W., Donlon, D., Cooper, A., & Genographic Consortium. (2011). Survival and recovery of DNA from ancient teeth and bones. *Journal of Archaeological Science*, *38*(5), 956–964. https://doi.org/10.1016/j.jas.2010.11.010

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources*, *20*(4), 892–905. https://doi.org/10.1111/1755-0998.13160

Andersen, J. C., & Mills, N. J. (2012). DNA extraction from museum specimens of parasitic Hymenoptera. *PLoS ONE*, *7*(10), e45549. https://doi.org/10.1371/journal.pone.0045549

Binladen, J., Wiuf, C., Gilbert, M. T. P., Bunce, M., Barnett, R., Larson, G., Greenwood, A. D., Haile, J., Ho, S. Y., & Hansen, A. J. (2006). Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. *Genetics*, *172*(2), 733–741. https://doi.org/10.1534/genetics.105.049718

Brace, S., Diekmann, Y., Booth, T. J., van Dorp, L., Faltyskova, Z., Rohland, N., Mallick, S., Olalde, I., Ferry, M., Michel, M., & Oppenheimer, J. (2019). Ancient genomes indicate population replacement in early Neolithic Britain. *Nature Ecology and Evolution*, *3*(5), 765–771. https://doi.org/10.1038/s41559-019-0871-9

Briggs, A. W., Stenzel, U., Johnson, P. L., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., & Lachmann, M. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(37), 14616–14621. https://doi.org/10.1073/pnas.0704665104

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), 87. https://doi.org/10.1093/nar/gkp1163

Brosi, B. J., & Briggs, H. M. (2013). Single pollinator species losses reduce floral fidelity and plant reproductive function. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(32), 13044–13048. https://doi.org/10.1073/pnas.1307438110

Brotherton, P., Endicott, P., Sanchez, J. J., Beaumont, M., Barnett, R., Austin, J., & Cooper, A. (2007). Novel high-resolution characterization of ancient DNA reveals C> U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Research*, *35*(17), 5717–5728. https://doi.org/10.1093/nar/gkm588

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Colgan, T. J., Arce, N. A., Gill, R. J., Rodrigues, A. R., Kanteh, A., Duncan, E. J., Li, L., Chittka, L., & Wurm, Y. (2022). Genomic signatures of recent adaptation in a wild bumblebee. *Molecular Biology & Evolution*, *39*(2), msab366. https://doi.org/10.1093/molbev/msab366

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, 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 of the United States of America*, *110*(39), 15758–15763. https://doi.org/10.1073/pnas.1314445110

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods*, *10*(4), 325–327. https://doi.org/10.1038/nmeth.2375

Deagle, B. E., Eveson, J. P., & Jarman, S. N. (2006). Quantification of damage in DNA recovered from highly degraded samples – A case study on DNA in faeces. *Frontiers in Zoology*, *3*(1), 1–10. https://doi.org/10.1186/1742-9994-3-11

Díez-del-Molino, D., Sánchez-Barreiro, F., Barnes, I., Gilbert, M. T. P., & Dalén, L. (2018). Quantifying temporal genomic erosion in endangered species. *Trends in Ecology & Evolution*, *33*(3), 176–185. https://doi.org/10.1016/j.tree.2017.12.002

Eilers, E. J., Kremen, C., Smith Greenleaf, S., Garber, A. K., & Klein, A. M. (2011). Contribution of pollinator-mediated crops to nutrients in the human food supply. *PLoS ONE*, *6*(6), e21363. https://doi.org/10.1371/journal.pone.0021363

Gilbert, M. T. P., Hansen, A. J., Willerslev, E., Rudbeck, L., Barnes, I., Lynnerup, N., & Cooper, A. (2003). Characterization of genetic miscoding lesions caused by postmortem damage. *The American Journal of Human Genetics*, *72*(1), 48–61. https://doi.org/10.1086/345379

Gilbert, M. T. P., Willerslev, E., Hansen, A. J., Barnes, I., Rudbeck, L., Lynnerup, N., & Cooper, A. (2003). Distribution patterns of postmortem damage in human mitochondrial DNA. *The American Journal of Human Genetics*, *72*(1), 32–47. https://doi.org/10.1086/345378

Gill, R. J., Baldock, K. C. R., Brown, M. J. F., Cresswell, J. E., Dicks, L. V., Fountain, M. T., Garratt, M. P. D., Gough, L. A., Heard, M. S., Holland, J. M., Ollerton, J., Stone, G. N., Tang, C. Q., Vanbergen, A. J., Vogler, A. P., Woodward, G., Arce, A. N., Boatman, N. D., Brand-Hardy, R., … Potts, S. G. (2016). Protecting an ecosystem service: Approaches to understanding and mitigating threats to wild insect pollinators. *Advances in Ecological Research*, *54*, 135–206. https://doi.org/10.1016/bs.aecr.2015.10.007

Hansen, A. J., Willerslev, E., Wiuf, C., Mourier, T., & Arctander, P. (2001). Statistical evidence for miscoding lesions in ancient DNA templates. *Molecular Biology and Evolution*, *18*(2), 262–265. https://doi.org/10.1093/oxfordjournals.molbev.a003800

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, *56*(2), 61–77. https://doi.org/10.2144/000114133

Heintzman, P. D., Elias, S. A., Moore, K., Paszkiewicz, K., & Barnes, I. (2014). Characterizing DNA preservation in degraded specimens of *Amara alpina* (Carabidae: Coleoptera). *Molecular Ecology Resources*, *14*(3), 606–615. https://doi.org/10.1111/1755-0998.12205

Hofreiter, M., Jaenicke, V., Serre, D., Haeseler, A. V., & Pääbo, S. (2001). DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research*, *29*(23), 4793–4799. https://doi.org/10.1093/nar/29.23.4793

Jensen, E. L., Díez-del-Molino, D., Gilbert, M. T. P., Bertola, L. D., Borges, F., Cubric-Curik, V., de Navascués, M., Frandsen, P., Heuertz, M., Hvilsom, C., Jiménez-Mena, B., Miettinen, A., Moest, M., Pečnerová, P., Barnes, I., & Vernesi, C. (2022). Ancient and historical DNA in conservation policy. *Trends in Ecology & Evolution*, *37*, 420–429. https://doi.org/10.1016/j.tree.2021.12.010

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., & Orlando, L. (2013). mapDamage2. 0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, *29*(13), 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Kistler, L., Ware, R., Smith, O., Collins, M., & Allaby, R. G. (2017). A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Research*, *45*(11), 6310–6320. https://doi.org/10.1093/nar/gkx361

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

Laetsch, D. R., & Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research*, *6*(1287), 1287. https://doi.org/10.12688/f1000research.12232.1

Leigh, D. M., Hendry, A. P., Vázquez-Domínguez, E., & Friesen, V. L. (2019). Estimated six per cent loss of genetic variation in wild populations since the industrial revolution. *Evolutionary Applications*, *12*(8), 1505–1512. https://doi.org/10.1111/eva.12810

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lindahl, T., & Andersson, A. (1972). Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. *Biochemistry*, *11*(19), 3618–3623. https://doi.org/10.1021/bi00769a019

Lindahl, T., & Nyberg, B. (1972). Rate of depurination of native deoxyribonucleic acid. *Biochemistry*, *11*(19), 3610–3618. https://doi.org/10.1021/bi00769a018

Lindahl, T., & Nyberg, B. (1974). Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry*, *13*(16), 3405–3410. https://doi.org/10.1021/bi00713a035

Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A., & Zdobnov, E. M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

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

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb-prot5448. https://doi.org/10.1101/pdb.prot5448

Morin, P. A., Archer, F. I., Avila, C. D., Balacco, J. R., Bukhman, Y. V., Chow, W., Fedrigo, O., Formenti, G., Fronczek, J. A., Fungtammasan, A., Gulland, F. M. D., Haase, B., Heide-Jorgensen, M. P., Houck, M. L., Howe, K., Misuraca, A. C., Mountcastle, J., Musser, W., Paez, S., … Jarvis, E. D. (2021). Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Molecular Ecology Resources*, *21*(4), 1008–1020. https://doi.org/10.1111/1755-0998.13284

Mullin, V. E., Stephen, W., Arce, A. N., Nash, W., Raine, C., Notton, D. G., Whiffin, A., Blagderov, V., Gharbi, K., Hogan, J., Hunter, T., Irish, N., Jackson, S., Judd, S., Watkins, C, Haerty, W., Ollerton, J., Brace, S., Gill R. J., & Barnes, I. (2022). First large-scale quantification study of DNA preservation in insects from natural history collections using genome-wide sequencing. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.5mkkwh787

Ollerton, J., Erenler, H., Edwards, M., & Crockett, R. (2014). Extinctions of aculeate pollinators in Britain and the role of large-scale agricultural changes. *Science*, *346*(6215), 1360–1362. https://doi.org/10.1126/science.1257259

Pedersen, J. S., Valen, E., Velazquez, A. M. V., Parker, B. J., Rasmussen, M., Lindgreen, S., Lilje, B., Tobin, D. J., Kelly, T. K., & Vang, S. (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

Pinsky, M. L., Eikeset, A. M., Helmerson, C., Bradbury, I. R., Bentzen, P., Morris, C., Gondek-Wyrozemska, A. T., Baalsrud, H. T., Brieuc, M. S. O., & Kjesbu, O. S. (2021). Genomic stability through time despite decades of exploitation in cod on both sides of the Atlantic. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(15), e2925453118. https://doi.org/10.1073/pnas.2025453118

Potts, S. G., Imperatriz-Fonseca, V., Ngo, H. T., Aizen, M. A., Biesmeijer, J. C., Breeze, T. D., Dicks, L. V., Garibaldi, L. A., Hill, R., Settele, J., & Vanbergen, A. J. (2016). Safeguarding pollinators and their values to human well-being. *Nature*, *540*(7632), 220–229. https://doi.org/10.1038/nature20588

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raxworthy, C. J., & Smith, B. T. (2021). Mining museums for historical DNA: Advances and challenges in museomics. *Trends in Ecology & Evolution*, *36*(11), 1049–1060. https://doi.org/10.1016/j.tree.2021.07.009

RStudio Team. (2021). *RStudio: Integrated development for R*. RStudio. Retrieved from http://www.rstudio.com/

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

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*(1), 1–7. https://doi.org/10.1186/s13104-016-1900-2

Schwarz, C., Debruyne, R., Kuch, M., McNally, E., Schwarcz, H., Aubrey, A. D., Bada, J., & Poinar, H. (2009). New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. *Nucleic Acids Research*, *37*(10), 3215–3229. https://doi.org/10.1093/nar/gkp159

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE*, *11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Ugelvig, L. V., Nielsen, P. S., Boomsma, J. J., & Nash, D. R. (2011). Reconstructing eight decades of genetic variation in an isolated Danish population of the large blue butterfly *Maculinea arion*. *BMC Evolutionary Biology*, *11*(1), 1–11. https://doi.org/10.1186/1471-2148-11-201

van Houdt, J. K. J., Breman, F. C., Virgilio, M., & De Meyer, M. (2010). Recovering full DNA barcodes from natural history collections of *Tephritid fruitflies* (Tephritidae, Diptera) using mini barcodes. *Molecular Ecology Resources*, *10*(3), 459–465. https://doi.org/10.1111/j.1755-0998.2009.02800.x

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 1–13. https://doi.org/10.1186/s13059-019-1891-0

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mullin, V. E., Stephen, W., Arce, A. N., Nash, W., Raine, C., Notton, D. G., Whiffin, A., Blagderov, V., Gharbi, K., Hogan, J., Hunter, T., Irish, N., Jackson, S., Judd, S., Watkins, C., Haerty, W., Ollerton, J., Brace, S., Gill, R. J., & Barnes, I. (2022). First large-scale quantification study of DNA preservation in insects from natural history collections using genome-wide sequencing. *Methods in Ecology and Evolution*, *00*, 1–12. https://doi.org/10.1111/2041-210X.13945