UNIVERSITY of York

This is a repository copy of *A new model for ancient DNA decay based on paleogenomic meta-analysis*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/116501/

Version: Published Version

Article:
Kistler, Logan, Ware, Roselyn, Smith, Oliver et al. (2 more authors) (2017) A new model for ancient DNA decay based on paleogenomic meta-analysis. Nucleic Acids Research. ISSN 0305-1048

https://doi.org/10.1093/nar/gkx361

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial (CC BY-NC) licence. This licence allows you to remix, tweak, and build upon this work non-commercially, and any new works must also acknowledge the authors and be non-commercial. You don't have to license any derivative works on the same terms. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx361

A new model for ancient DNA decay based on paleogenomic meta-analysis

Logan Kistler[1,2,*], Roselyn Ware[1], Oliver Smith[1,3], Matthew Collins[3,4] and Robin G. Allaby[1,*]

[1]School of Life Sciences, University of Warwick, Coventry CV4 7AL, UK, [2]Department of Anthropology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA, [3]Section for Evolutionary Genomics, Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1307 Copenhagen K, Denmark and [4]Department of Archaeology, University of York, PO Box 373, York, UK

Received March 21, 2017; Revised April 15, 2017; Editorial Decision April 17, 2017; Accepted April 20, 2017

ABSTRACT

The persistence of DNA over archaeological and paleontological timescales in diverse environments has led to a revolutionary body of paleogenomic research, yet the dynamics of DNA degradation are still poorly understood. We analyzed 185 paleogenomic datasets and compared DNA survival with environmental variables and sample ages. We find cytosine deamination follows a conventional thermal age model, but we find no correlation between DNA fragmentation and sample age over the timespans analyzed, even when controlling for environmental variables. We propose a model for ancient DNA decay wherein fragmentation rapidly reaches a threshold, then subsequently slows. The observed loss of DNA over time may be due to a bulk diffusion process in many cases, highlighting the importance of tissues and environments creating effectively closed systems for DNA preservation. This model of DNA degradation is largely based on mammal bone samples due to published genomic dataset availability. Continued refinement to the model to reflect diverse biological systems and tissue types will further improve our understanding of ancient DNA breakdown dynamics.

INTRODUCTION

The genomic era of massively parallel DNA sequencing has driven a revolutionary body of research using ancient DNA-based genomics (1,2). Paleogenomics has led to the re-writing of recent hominin evolutionary history (3), nuanced understandings of historical human movements and interactions around the globe (4,5), breakthroughs in Quaternary paleontology (6–8), evolutionary ecology, the biology of extinct species (9), impacts of humans on ancient ecosystems and biodiversity (10) and the evolution and movements of domestic plants and animals (11–14). The successful probing of ancient epigenomes, microbiomes and metagenomes further illustrates the flexibility and information value of ancient DNA-based research in the genomic age (15–17). In sum, time-series genomic datasets have proven extremely valuable in diverse research avenues.

In addition to the scale and sensitivity of analysis afforded by genomic methods in ancient DNA research, genomic datasets allow for a revised understanding of the patterns and expectations of DNA survival over millennia. This is beneficial in two key ways: (i) Criteria of ancient DNA authenticity warrant updating for the genomic era and formalized expectations of DNA degradation are necessary for this process; and (ii) Better predictive models of DNA degradation may help researchers target specimens likely to yield high information value where destructive analysis is unavoidable. Generally, ancient DNA is expected to be highly fragmented (18–20) and to carry an abundance of characteristic misincorporations—deaminated cytosine residues appearing as C-to-T transitions in single-stranded fragment overhangs (21). Further, DNA fragmentation is biased by biomolecular context. For example, a short-range (∼10 bp) periodicity observed in the distribution of fragment lengths is attributed to the period of a complete turn of the DNA double-helix around a histone (22), which is thought to offer some protection against breakage at histone-adjacent sites. Finally, base compositional biases have been observed in DNA preservation, especially enrichment of GC-content in ancient DNA (23).

The relationships between these characteristic patterns of DNA degradation and the preservational environment and age of tissues are poorly understood. We carried out a meta-

*To whom correspondence should be addressed. Tel: +1 202 633 1908; Fax: +1 202 357 2208; Email: kistlerl@si.edu
Correspondence may also be addressed to Robin G. Allaby. Tel: +44 24 7657 5059; Fax: +44 7657 4500; Email: r.g.allaby@warwick.ac.uk
Present addresses:
Logan Kistler, Department of Anthropology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, United States.
Oliver Smith, Section for Evolutionary Genomics, Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen K, Denmark.

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

analysis of 185 ancient genomic datasets—dating from the Middle Pleistocene to the nineteenth century from 21 published studies (Figure 1)—to test for relationships between sample age, environmental variables and DNA diagenesis. We used mapDamage 2.0 (24) to quantify deamination and we developed tests for assessing fragmentation, histone periodicity and energetic base-compositional biases in ancient genomic data. We analyzed these damage statistics in relation to sample age, annual mean temperature, temperature fluctuation and precipitation—treated as a proxy for humidity—using simple multivariate linear models. Ultimately, we aimed to establish the key determinants of DNA survival and the specific patterns of DNA breakdown expected under variable conditions.

MATERIALS AND METHODS

Dataset compilation and initial processing

We obtained unmapped (fastq) or mapped (bam) sequence reads from each of 185 publicly available ancient DNA datasets generated by shotgun sequencing without uracil removal, comprising anatomically modern humans ($n = 156$ (26–38)), herbarium plant samples ($n = 15$, aligned to the host plant rather than the pathogen examined in ref (39), with environment data estimated at the herbaria of long-term storage rather than the site of collection), Colombian and woolly mammoths ($n = 4$ (7,40)), neandertals ($n = 3$ (41)), horses ($n = 5$ (13,42–43)) and polar bear ($n = 1$ (44)). We avoided data generated through target capture experiments to avoid possible hybridization biases introduced by misincorporated residues or read length variation. For unmapped samples, we used Flexbar (45) to trim adapter sequences in single-end read data, and PEAR (46) to perform adapter trimming and read merging in paired-end datasets. We used the bwa-backtrack algorithm within the Burrows–Wheeler Aligner (47) to map read data to the corresponding reference genome and collapsed duplicates using the 'rmdup' function in SAMtools (48). We filtered all bam files for a minimum mapping quality of 20 using the SAMtools 'view' function and filtered for minimum read length of 20 using Unix tools. We separated nuclear and organellar reads (mtDNA in mammals, plastid DNA in plants) into separate bam files. For the mammoth samples mapped to the African elephant genome ($n = 4$), we removed mitochondrial reads from the bam file and re-mapped the complete raw datasets to a woolly mammoth mitochondrial sequence (NC_007596.2). Sample latitude and longitude were used to estimate annual mean, minimum and maximum temperature estimates, plus annual precipitation for each of the samples. These were taken from the WorldClim (49) current condition database using the R 'raster' package (50) at a resolution of 2.5 arc-min. In cases where specific site location was not available at the longitude/latitude level, Google Earth was used to estimate longitude and latitude from details or site maps provided in the relevant publications. Location details and temperature estimates are given in Supplemental Dataset S1. Climate estimates reflect modern climate conditions rather than a complete climate legacy over the time span of each sample. We then estimated DNA preservation and breakdown parameters as described below.

Deamination

We used mapDamage 2.0 (24) to estimate cytosine deamination in single-stranded overhangs, δ_s, invoking default settings with the following exceptions: we subsampled large bam files to correspond with 1 Gigabyte input file (~10–20 million reads with typical dataset complexity and a human genome) using the mapDamage '*-n*' option. We analyzed the mapDamage Markov chain Monte Carlo (MCMC) output from each sample using the 'coda' R package (51) to estimate an effective sample size (ESS) for each of the six variables estimated by the mapDamage simulation. ESS values are reported in Supplemental Dataset S1. We enforced a minimum ESS of 200 in all variables to ensure MCMC simulation convergence, excluding nine datasets for deamination analysis. For libraries with highly asymmetrical 3′ and 5′ C-to-T mismatch observed visually in misincorporation plots, indicating the likely use of a non-proofreading DNA polymerase for library amplification—incapable of recovering uracils in template DNA—we re-ran mapDamage with the '–reverse' option to estimate damage from the 3′ end only. We noted extremely high deamination and overhang termination values in the output from Mammoth M4 (7), which suggested a much higher rate of deamination than even much older permafrost samples. However, that library is dominated by very short fragments (ref (7); summarized in the fragment length plot on Dryad), which we hypothesized could influence the mapDamage MCMC to over-estimate both parameters. We re-analyzed that sample considering only reads ≥40nt, yielding the damage parameter values reported in Dataset S1. Data from the Saqqaq Palaeo-Eskimo genome (32) were mostly generated using a non-proofreading enzyme, but a small proportion of read files were reported to have been generated using a proofreading Platinum High Fidelity Taq polymerase (22). We mapped all Saqqaq read files from the Sequence Read Archive ($n = 218$) to a human mitochondrial genome (EU256375.1), used PMDtools (52) to rapidly generate misincorporation plots and visually inspected each for elevated 5′ C-to-T mismatch. This approach yielded two libraries apparently produced using a proofreading enzyme, one of which (SRR030983) was carried through for analysis. All mapDamage output files (run logs, plots, MCMC trace files and summary statistics) from the 185 final runs are available in the Dryad Digital Repository (see Availability below). Finally, we summarized a deamination rate for each sample according to the equation:

$$\text{rate} = \ln\left(\frac{1}{1-\delta_s}\right)\left(\frac{1}{\text{age}}\right)$$

Fragment length distribution

Fragment lengths are expected to form an exponential distribution under random breakdown. The distribution of DNA fragment sizes can therefore be summarized as λ (53), the single parameter of the exponential distribution. To estimate λ, we first summarized a frequency distribution table of fragment lengths. If a frequency spike was observed at the maximum fragment length—indicating fragments greater than the read length and an artifactual peak among reads with no adapter trimming—we re-estimated

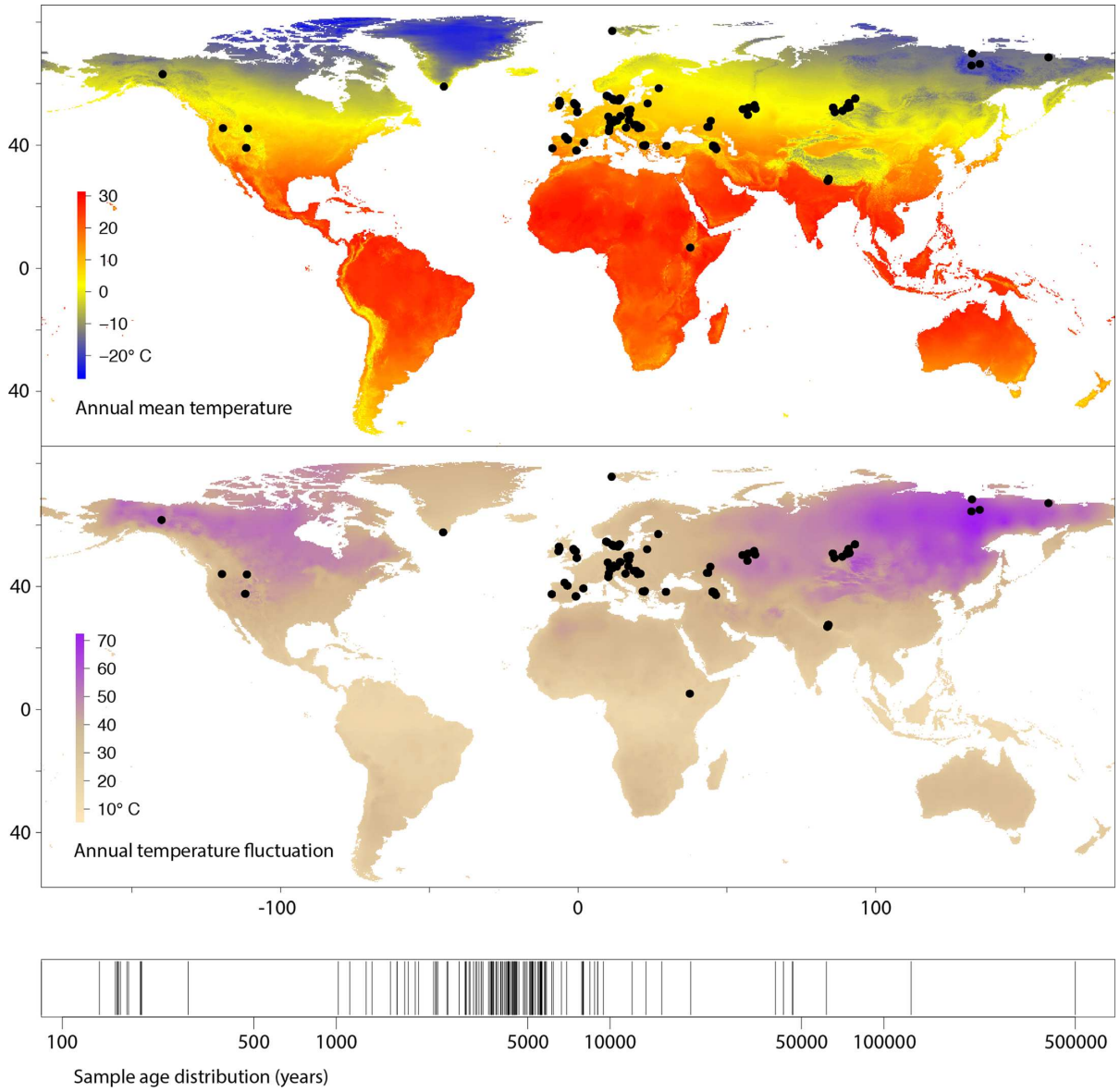

Figure 1. Locations of 185 samples ($n = 94$ unique sites) used in paleogenomic meta-analysis, global variation in mean temperature and temperature fluctuation, and timeline of sample ages. Note the absence of sites with annual mean temperature >20°C, reflecting known preservation bias toward cooler climates (25).

the maximum reliable fragment length as follows: beginning with the longest fragment, we pruned the table back to the point at which the next shorter fragment was observed more frequently, eliminating up to 6 length values (mean = 3). We iterated over all ranges of at least 20 consecutive length values in the table, attempting to fit an exponential formula using the R function: $nls(y \sim N*exp(-k*x))$, with starting values of $k = 0.05$, $N = 0.1$ and λ represented by the inferred value of k. We retained the top 5% of best fits on the basis of P-values obtained by summarizing the formula output in R and estimated the value of λ from the table segment producing the best overall fit. We visualized the results in the top 5% of best fits to confirm a reasonable λ estimate by eye (e.g. Supplementary Figure S1). We observed fragment length heterogeneity in some cases, likely created by mapping biases and occasional anomalous spikes in length frequencies that disrupted automated estimation of λ. Therefore, during visual inspection, we sometimes opted to override the inferred λ value by (i) manually defining a range of fragment lengths over which to re-calculate λ and/or (ii) clipping artifactual frequency spikes by imposing a single frequency threshold value (e.g. Supplementary Figure S1). All summary statistics and plots for λ estimation are available on Dryad, including run logs detailing manual override decisions. Perl and R code for λ estimation are available on Dryad.

Histone periodicity estimation

To estimate the intensity of a preserved histone signal (22), we analyze periodic deviations from a medium-range

smoothing algorithm imposed on the fragment length frequency tables (Supplementary Figure S2). Using a fragment length frequency table, we again eliminated the artifactual peak at the maximum length as above for λ and trimmed the distribution to the innermost length values each representing at least 0.002 of the total fragments. Additionally, for artifactual spikes within the frequency table, we adjusted any single frequency greater than 1.5× the midpoint of its flanking neighbors down to the midpoint. This approach affected only artifacts and not the underlying distribution. We then fit a locally weighted scatterplot smoothing (lowess) curve with the R 'lowess' function, using a smoothing span of 20 to normalize over histone periods expected to be ∼10nt (22). We then removed four additional length values from each end of the distribution to eliminate increased terminal deviations from the lowess curve.

We observed that in samples with a histone signal, the deviation of the observed values from the lowess curve is best approximated by a series of local exponential functions with the midpoint of a complete histone period set to $x = 0$ (Supplementary Figure S2). That is, frequency values across a single histone period form a parabolic curve when normalized for overall fragmentation described by λ. Therefore, we tested for this pattern in all subranges of 8–12 consecutive length values in the table, setting the midpoint of each subrange to $x = 0$ and using the observed value divided by the lowess values for the y-axis. We used the R function *nls*($y \sim k+|N|*x\wedge 2$)), with starting values of $k = 1$ and $N = 0.1$, so that k should deviate minimally from 1 to absorb noise, and positive values of N provide a metric of signal intensity. That is, N increases linearly with the degree of observed frequency deviation from the lowess curve at local maxima. We retained the starting position, range lengths and N values of all significant fits on the basis of P-value. We then used an optimized one-dimensional k means clustering algorithm in the R 'Ckmeans.1d.dp' package (54) to localize strongly significant starting locations of histone periods. For all adjacent (i, j) pairs of cluster positions representing the putative best locations for histone peaks, a periodicity coefficient was calculated: If $j - i \geq 8$ and $j - i \leq 12$, the coefficient increases by 1/(n clusters −1). Otherwise, if some value v in 2 through (range of values)/10 satisfies $(j - i)/\mathrm{v} \geq 8 \cup (j - i)/\mathrm{v} \leq 12$, coefficient increases by (1/v)/(n clusters −1). Thus, given a minimum of three clusters and a minimum periodicity coefficient of 0.3, cluster position values satisfying the requirement could include 10,20,40 (coefficient = 0.75); 10,20,50 (0.67), 10,30,50 (0.5) or 10,40,70 (0.33), but not 10,50,80 (0.29).

To optimize histone signal estimation, we used the organellar datasets (which lack histones *in vivo*) to calibrate model parameters against false positive results (Supplementary Figure S3). We permuted the minimum number of significant clusters detected (2,3,4), a minimum observed proportion of all plausible histone periods given the range of values analyzed (0, 0.1, 0.2 and 0.4), the minimum periodicity coefficient (0.2, 0.3, 0.4 and 0.5) and a minimum P-value threshold for significant exponential fit (0.05, 0.01 and 0.001). We summarized the number of nuclear and organellar datasets satisfying the requirements of 144 separate model permutations (n = 53 280 total iterations). P-value, minimum cluster count and minimum periodicity coefficient proved the best predictors of false positive rates, accounting for 79% of the variance in false positive rate under a simple linear model (Supplementary Figure S3), while proportional number of clusters did not add predictive power. Under a range of parameter values, we were able to estimate nuclear histone signal intensity with no organellar false positives in up to 112 of the 185 samples for a given model. Given a 5% allowable false positive rate in the single model with the highest ratio of nuclear to organellar estimates, we were able to recover 138 nuclear estimates. We recorded the median value of the N intensity parameter for all samples under strict conditions with no organellar false positives (n = 112) for analysis (Dataset S1 'Histone_Intensity' for estimates; see the Dryad repository for pdf and run files). Notably, the Thistle Creek horse genome (43) displays a clear short-range periodicity on visual inspection, but at about half the normal length (∼5 bp). The reason for this behavior is unclear, but this distribution violates the model assumption of ∼10 bp period and therefore this sample only ever presented as a likely false positive during model calibration.

Base composition

To assess biases in base composition related to preservation, we first summarized 8-mer frequencies in each reference genome, excluding soft- and hard-masked repeat regions. We then summarized 8-mer frequencies in the bam files from sequence reads in mapped orientation to match the reference. For each 8-mer, a simple enrichment factor was calculated as (frequency in mapped reads)/(genomic frequency). The enrichment and depletion of ancient DNA motifs is affected by a complex range of conditions, as suggested by clear multimodality in the distribution of 8-mer enrichment factors (Supplementary Figure S4). Additionally, *in vitro* variables further bias the datasets through penalizing GC-extreme reads, for example (55), and chromatin modeling and nucleosome occupancy is expected to have differential effects on the protection and survival of coding versus non-coding DNA. To isolate these effects, we calculated a simple GC proportion for each 8-mer and based on known systematic sequence complexity biases among genomic element types (56), we calculated 8-mer Shannon entropy (H) using the following formula after ref (57):

$$H = \sum\nolimits_{i = \mathrm{A,C,G,T}} -1 * \left(\frac{n_i}{n_{\mathrm{ACGT}}}\right) * \log_2\left(\frac{n_i}{n_{\mathrm{ACGT}}}\right)$$

Finally, we calculated a simple kmer enthalpy as the sum of all dimer base-stacking energy values (kcal/mol/dimer) reported in table IX of ref (58), using the values from the 'corrected optimized potential' method. We first visualized kmer enrichment in relation to all three sequence variables—GC content, entropy and enthalpy—and noted extensive variation among samples as expected (e.g. Supplementary Figure S4, kmer summary files for all datasets, code for 3D scatterplots and code for enthalpy-biased kmer frequency estimation are available on Dryad). For example, GC content and entropy are parabolically related by definition—equal base frequencies are required for maximum entropy. As such, disentangling enrichment of high-

entropy kmers from *in vitro* penalization of GC-extreme kmers is intractable, and kmer enrichment patterns varied extensively in terms of entropy and GC-content. However, we noted a strong relationship between enrichment and enthalpy in several samples (Supplementary Figure S4), and therefore we opted to isolate enthalpy from GC content and entropy for analysis as follows:

In total, there are 52 unique GC-content—Shannon entropy combinations for DNA 8-mers. Each of the 52 configurations of invariant GC and entropy values represents between 2 and 6720 kmers ($n = 65\,536$ total kmers) representing a distribution of enthalpy values and enrichment factors. From within each of the 52 GC–entropy configuration bins, we reiteratively selected random pairs of kmers up to the total number of kmers in the bin (i.e. given a bin containing 32 kmers, 32 random pairs were selected and a total of 65 536 pairs are drawn from the dataset). We then compared enrichment factors and enthalpy values between the two kmers. Given a 'success' if the higher-enthalpy kmer was also the more enriched kmer, as hypothesized, we incremented a counter and decremented the counter for failures. Following iterations, we recorded the value of the counter divided by the number of trials and used this value as the test statistic to compare with environmental variables, where positive values indicate overall enrichment of higher-enthalpy motifs (Dataset S1, 'Enthalpy_bias').

Linear models

We carried out multiple linear regression analyzes to test for relationships between preservation parameters (above) and environmental variables. Specifically, we tested in turn for significant predictors of each damage parameter using a linear model analysis with four independent variables: annual mean temperature, annual temperature fluctuation, annual mean precipitation and the natural log of sample age using the R 'lm' function (59). Results presented are from analysis of the nuclear datasets after excluding the organellar data.

Deamination prediction from temperature

We calculated a density distribution of deamination rates using default parameters and bandwidth in the R 'density' function and creating a weighting vector where each point's weight value was calculated as 1-(local density/maximum density). We then fit a weighted linear regression between temperature and deamination rate using the R 'lm' function with the 'weight' option invoked using the above weighting vector. We used the R 'predict' function to predict a rate and 95% confidence intervals at temperature values of −20°C, 0°C and 20°C. The R code to replicate the analysis and recreate Figure 3 from Dataset S1 is available on Dryad in the 'expedient code' file.

Simulating DNA fragmentation and deamination

We hypothesize below (Results and Discussion) that a time-dependent fragmentation process may be incongruent with the observation of total cytosine deamination in single-stranded overhangs ($\delta_s = 1$). We therefore carried out a set of simulations to test the effects of varying the fragmentation rate on the accumulation of deaminated residues. Simulations were executed using a custom perl script. Beginning with a λ value of fragmentation (e.g. 0.013–0.157 range from our meta-analysis), we infer the number of random fragmentation events necessary to yield the lambda value as λ x (total length of all fragments). We then randomly select a simulated number of imposed total fragmentation events from a Poisson distribution, using the exact number of fragmentation events as the Poisson lambda parameter. We pre-allocate the selected number of breakage events—without replacement, as breakage is impossible twice at the same location—to locations in a population of starting molecules. Breakage occurs at zero-width sites between simulated residues in our simulation, such that molecules can be reduced to 1 nt but not lost completely. We then allocate fragmentation events to timing bins by sampling from the probability density function of a beta distribution with the α parameter held at 1 and the β parameter varying ≥1 to introduce a changing rate profile: $\beta = 1$ describes a constant rate of fragmentation and higher values of β describe increasingly skewed scenarios where fragmentation occurs in the early cycles of the simulation. For example, when $\alpha = 1$ and $\beta = 10$, roughly 90% of fragmentation has occurred when 20% of time has passed. Alternatively, $\alpha = 1$ and $\beta = 1$ describes a uniform distribution of fragmentation where the rate parameter $k = \lambda/\text{time}$, as per ref (18). To complete setup, we impose a δ_s value and calculate a deamination rate as described above in terms of probability of deamination per site per year, and impose a value to describe fragment overhangs per mapDamage 2.0 (0.3 in our simulation).

We then carry out a forward simulation through cycles of drawing from the randomized predetermined breakage sites according to the timing bin allocations and introduce new single-stranded overhangs at newly broken sites by sampling randomly from a Poisson distribution described by the overhang λ value. Following each breakage cycle, each overhang is subjected to a round of deamination according to the rate calculated from δ_s, where each site is given the opportunity to undergo deamination if a pseudo-random number ($0 \leq x \leq 1$) falls below the rate value. Finally, at the end of each deamination cycle, we summarize the current fragmentation λ value as 1/(mean current fragment length) and the current δ_s value as the (number of deamination sites)/(all overhang sites). When $\alpha = 1$ and $\beta = 1$, λ increases linearly to approach the imposed λ value, while with higher β values, alternative patterns occur (Supplementary Figure S5).

RESULTS AND DISCUSSION

We found that cytosine deamination is strongly influenced by both sample age and site mean temperature (multiple $r^2 = 0.264$; age $P = 1.9 \times 10^{-9}$; temperature $P = 1.52 \times 10^{-5}$, model $P = 2.54 \times 10^{-10}$; Figure 2). Previous studies have identified age as the key critical predictor of deamination (60), but our finding is in line with predictions of a time-dependent hydrolytic process where activation energy is achieved more often at higher ambient temperatures. A rate of deamination can be calculated for any sample with a known age and partial conversion of exposed cytosines (Figure 3). The resulting rates vary widely and

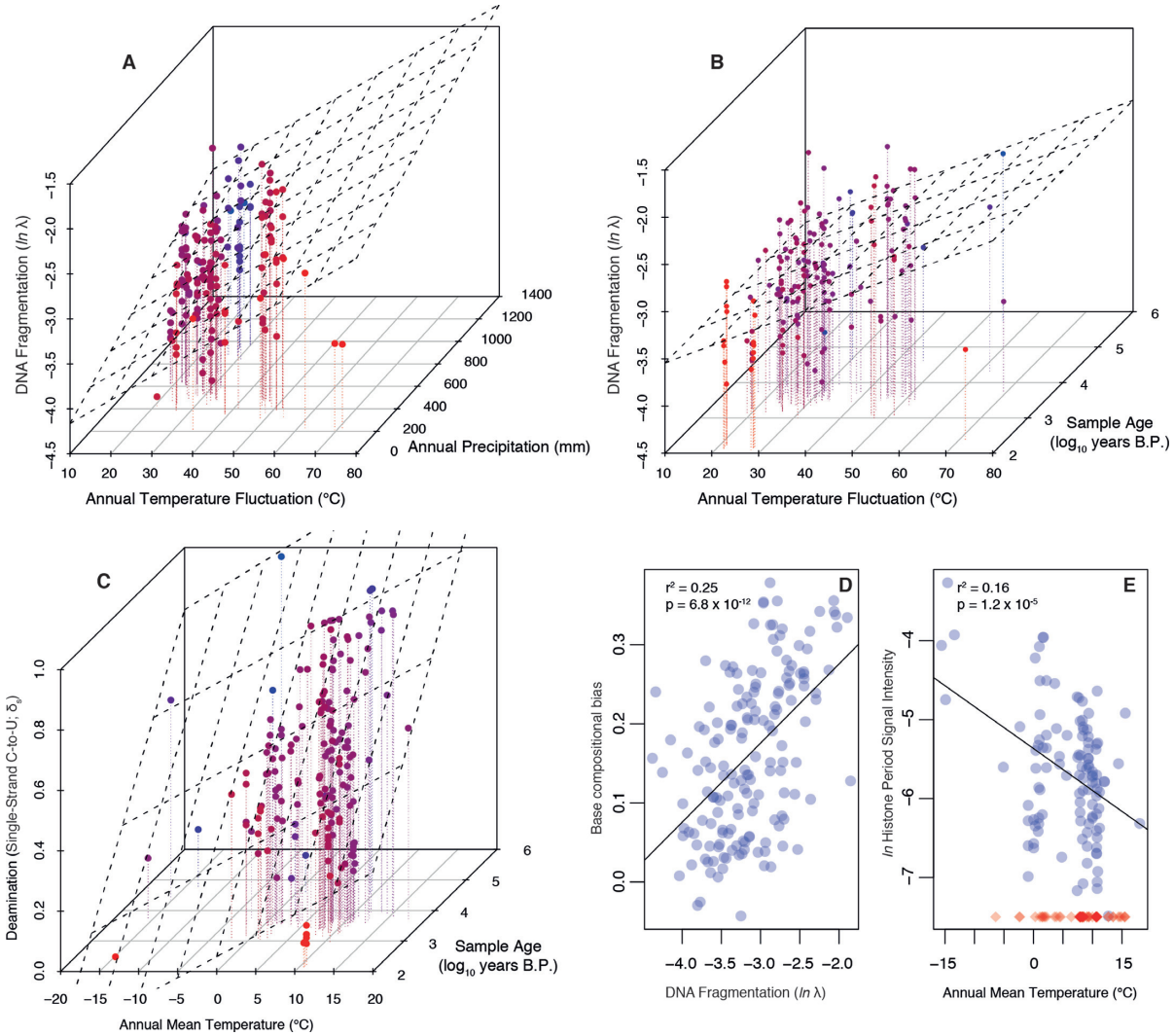

Figure 2. Relationships between DNA degradation parameters and environmental variables. (**A**) DNA fragmentation is correlated with thermal fluctuation and precipitation. (**B**) DNA fragmentation is correlated with thermal fluctuation but is not influenced by sample age. (**C**) Deamination is a thermal age parameter, strongly associated with both age and temperature. Coloring in (A–C) is used to enhance the *z*-axis variation: red points are the nearest and blue are the most distant. (**D**) DNA fragmentation is highly predictive of base compositional biases, with fragmented datasets depleted of motifs with low base-stacking energy. (**E**) Histone periodicity in fragment length distribution is most pronounced in samples from cold environments. Blue circles represent samples where a histone periodicity estimate was possible ($n = 112$; see 'Materials and Methods' section for calibration against false positive results), red diamonds are samples where no periodicity was observed, visualized at −7.5 on the *y*-axis to reflect the observation of no detectable bias.

show a strong correlation with temperature ($r^2 = 0.279$; $P = 1.23 \times 10^{-12}$). In sum, deamination is a time-dependent process heavily modulated by temperature. When analyzing DNA fragmentation, however, we found that precipitation and thermal fluctuation were strongly significant predictors (multiple $r^2 = 0.202$; precipitation $P = 0.0025$; temperature fluctuation $P = 6.18 \times 10^{-8}$) but that age was not significantly correlated with the degree of fragmentation ($P = 0.77$), even when controlling for environmental conditions. We also find that in addition to the humidity and thermal fluctuation pattern, the degree of DNA fragmentation correlates strongly with base compositional biases. Specifically, datasets dominated by short fragments are significantly depleted of weakly-bonded nucleotide motifs ($P = 6.79 \times 10^{-12}$, $r^2 = 0.253$; Figure 2), indicating that DNA breakdown follows predictable patterns with regard to microenvironment and nucleic acid biochemistry. Relatedly, we detected a histone-associated fragmentation bias (22) in the majority of our samples ($n = 112$), and we find that annual mean temperature is strongly associated with the intensity of this pattern ($P = 1.2 \times 10^{-5}$, $r^2 = 0.16$; Figure 2). Specifically, DNA breakdown in colder environments appears to more faithfully reflect cellular architecture and the *in vivo* genome context, whereas breakdown in warmer conditions is much less discriminant.

At present, the ancient DNA literature lacks clear consensus on the fundamental predictors of DNA fragmentation: one recent study identified a strong age dependency in DNA recovery through quantitative polymerase chain reaction (qPCR) analysis of a regionally controlled time series of bone samples (18). The authors proposed, on the basis of this result, that DNA degradation in ancient bone is

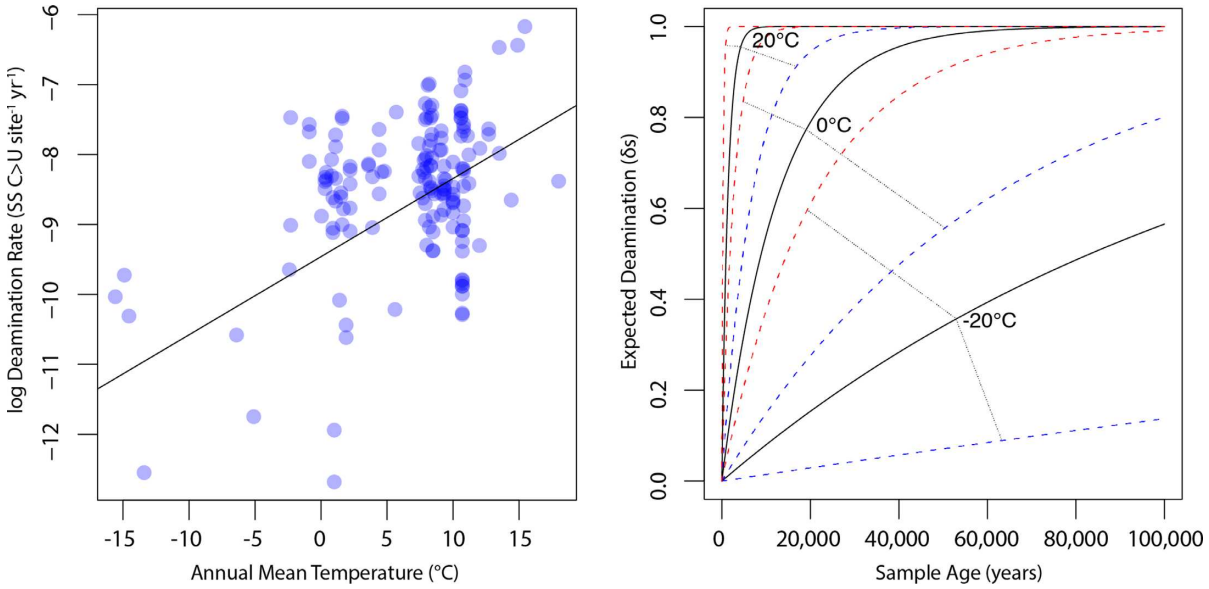

Figure 3. Expectations of deamination over time for variable temperatures. Left: density-weighted linear regression of temperature and log deamination rate calculated using the formula $rate = ln\,(1/(1-\delta_x))$ * $(1/age)$. Right: using rate estimates from the weighted regression, we calculated the expected δ_s values over a 100 000 year time span using the formula $\delta_s = 1-(1/e^{rate*age})$ re-arranged from the above, as well as δ_s values for rates estimated from 95% confidence intervals of the regression. We visualized the expected deamination levels for samples from −20, 0 and 20°C contexts (solid lines), along with upper (red) and lower (blue) confidence bounds. This predictive model is necessarily based primarily on mammalian bone tissue, and we expect refinements to these expectations based on sample type, for example, as more datasets become available.

mainly driven by thermal age-dependent hydrolytic depurination driving rate-constant fragmentation over time. Under this model, fragmentation is gradual and continues under a rate described by a decay constant until the bonds between all DNA fragments are destroyed and DNA is completely lost. Alternatively, a separate analysis (60) found no significant link between sample age and the degree of fragmentation. Consistent with this finding, early ancient DNA research pointed to very rapid initial DNA decay followed by subsequent stabilization (61), rather than fragmentation as an exponential random decay process. Additionally, controlled experiments using qPCR with recently deceased tissues demonstrate a precipitous immediate decline of endogenous DNA content and/or quality, followed by stabilization hypothesized to be linked to the mineral environment of bone (62). This model likewise contradicts the idea that DNA decay can be thought of strictly in terms of exponential breakdown under a decay constant. In total, evidence has been presented for both a rate-constant decay model and a more age-independent scenario. Here, our meta-analysis points to the statistical decoupling of age and fragmentation. We aimed to further test this finding with three strategies:

First, we recognize that there are numerous sources of variance that cannot be controlled in our meta-analysis across several studies—including sample excavation and storage, wet lab and computational methods, and species and tissue types—and that these sources of variance have the potential to obscure subtle relationships. If major sources of inter-study confounding variance in DNA fragmentation were present, the result would be the dampening of any statistical relationship between natural variables and fragmentation as the *in situ* signal for fragmentation is lost. If age was a significant predictor of DNA decay along with thermal fluctuation and humidity, it is difficult to imagine that only the age relationship would be lost due to post-excavation handling and inter-study variation. Therefore, we suggest that confounding variance is not a parsimonious explanation for the lack of a clear age-fragmentation relationship in the presence of a robust environmental association. However, to test this possibility more directly, we restricted our analysis from 185 datasets across 21 studies to 97 Bronze Age human genomes generated from a single study (26). We thereby control for species, tissue type, and biases in sample preparation and we consider a narrower timeframe and more constrained set of preservational conditions, eliminating several potential sources of confounding variance. Additionally, we recognize that by using modern climate estimates, we might obscure some subtle variation driven by climate fluctuations over long time periods. Restricting our analysis to just this Bronze Age dataset substantially constrains the temporal and geographic ranges, so that the sample set is expected to have been exposed to a less variable range of climate fluctuations compared with, for example, much older samples exposed to Pleistocene shifts. Under the same linear model as above (see 'Materials and Methods' section), we find that exactly as in the broader dataset, thermal fluctuation and precipitation were significant predictors of fragmentation (respectively, $P = 0.014$ and $P = 4.2 \times 10^{-4}$; multiple $r^2 = 0.25$), but age was still not a significant predictor of overall fragmentation ($P = 0.420$).

Second, we tested the fundamental assumption that from a single site, DNA from older samples is expected to be more fragmented than from younger ones. While we initially analyzed data from 94 different sites, the meta-dataset includes 114 pairs of samples from the same site separated by

at least 100 years. Thus for these 114 pairs where we can eliminate inter-site variation, the older sample is predicted to be the more fragmented sample a significant majority of the time under the fundamental assumption that fragmentation increases with age in a single environment. Given 114 pairs of samples, only 55 (0.48) satisfy this assumption ('successes'). The null hypothesis of 47–67 successes ($P = 0.05$ calculated using the β-distribution) cannot be rejected and indeed fewer than half of cases satisfy the basic assumption. By increasing the minimum age difference to 1000 years, we retain 55 valid pairwise comparisons and still observe no relationship between age and fragmentation, with only 27 (0.49) satisfying the basic assumption (null hypothesis at $P = 0.05$: 23–32 successes). We validated this approach by replicating the procedure with deamination, a known age-linked variable (26 and above). With deamination, we reject the null hypothesis and find a significant age effect (131 comparisons possible, 80 successes (0.61); null hypothesis at $P = 0.05$: 55–76 successes).

Finally, we routinely observe complete deamination of all exposed cytosine residues. This saturation of measurable deamination has been described in several samples previously (24) and is observed in 14 out of the 185 (7.6%) datasets analyzed here, spanning 2–500 kya (Figure 4). However, complete deamination in single-stranded overhangs is incongruent with a rate-constant fragmentation model: if fragmentation followed a simple rate-constant process that would yield a robust association between thermal age and fragmentation, new overhangs would continually be exposed with the expectation of intact cytosine, suppressing the proportion of deaminated residues and preempting complete deamination. Even by simulating deamination rates 10 fold faster than the most extreme of those estimated in our meta-analysis, deamination fails to converge to saturation under a rate-constant fragmentation model (see 'Materials and Methods' section; Supplementary Figure S5). In total, observing complete deamination under a rate-constant fragmentation model would require that the deamination rate outpaces the fragmentation rate so that new overhangs are rapidly saturated with deamination—all exposed cytosine residues are rapidly converted to uracil. Under such extreme deamination rates, however, it is implausible that deamination would show such a robust correlation with age across samples as observed here and elsewhere (60).

We find strong validation that age does not predict DNA fragmentation in our meta-dataset. However, we recognize that DNA breakdown by hydrolytic depurination is a well-validated and immutable chemical mechanism by which DNA decays exponentially according to first-order kinetics, producing a measurable half-life signal of molecular depletion (63). The mismatch between this predicted behavior and our findings indicates that the preservation state of ancient DNA is determined by multiple processes and cannot be attributed to a simple fragmentation rate as suggested in a rate-constant fragmentation model. Instead, we propose a multi-stage DNA fragmentation model: first, physical and biotic stressors cause rapid breakdown of nucleic acids shortly after organism death. While microbes and cellular processes (e.g. autolysis and nuclease activity) rapidly degrade a large fraction of endogenous DNA—depending on tissue type and depositional environment—fragmentation appears to reach an initial threshold and then stabilize somewhat in contexts where DNA has the potential for long-term preservation.

The strong association of humidity and thermal fluctuation with DNA fragmentation suggests that processes like the loss of bioapatite surface area caused by diagenetic recrystallization and physical shearing effects of hydraulic fluctuations in bone, for example, play a role in the initial breakdown process. Further, DNA may reach a size in bony contexts—the majority of our re-analyzed datasets—where it can penetrate the protective internal porosity of bone and gain some additional protection. The counterintuitive result that DNA is sometimes better preserved in cooked than uncooked medieval bone may offer support for this scenario ((64), although see (65) for for further analysis of cremated bone). In our analysis, 15 plant samples from herbaria (39) fit with the overall fragmentation model—comparing linear model residuals reveals no significant difference between plant samples and non-plant samples (Welch's t-test, $P = 0.44$). However, they make up a very small fraction of the variation here and because of the possible role of the mineral makeup of bone in DNA preservation across samples (62), we suggest that re-analysis of plant data across a much greater age range will be important in understanding any possible differences in preservation between plant and animal tissues. Over a short time span, age-dependency in fragmentation has been suggested in plant tissues (66), but a paucity of paleogenomic plant data currently precludes a comprehensive analysis spanning thousands of years. In total, our meta-analysis and model are necessarily focused on mammalian hard tissue ($n = 169$ out of 185 datasets) given dataset availability. As more datasets are generated from diverse systems and tissue types, we expect further refinement of these general findings to reflect a more nuanced understanding behind the specific drivers of DNA diagenesis and factors underlying preservation. For example, DNA is integrated into hair during programmed cell death and keratinization leading to some amount of immediate shearing which might affect downstream processes (67). Thus ancient DNA in hair might warrant a modified set of expectations for preservation relative to bony tissue given a certain background environment. Recent experimentation comparing tooth cementum and petrous bone DNA characteristics reinforces the necessity of integrating sample type information in assessing DNA degradation in the future (65). Finally, the relationships between biomolecule preservation and the microenvironment of sample deposition (e.g. soil pH, mineral content, sample depth and aerobic activity) bear significant further investigation as more datasets become available.

Previous research identified a strong age dependency in DNA recovery—assayed by qPCR—in a controlled time-series of bone samples from a regional set of depositional sequences and interpreted the result as evidence for an exponential decay process due to time-dependent DNA fragmentation (18). However, bulk diffusion of DNA—rather than rate-constant fragmentation—provides an equally parsimonious scenario for the observed signal. Specifically, the previous study estimated a 521-year half-life for a target fragment of 242 bp under the conditions of the site (18).

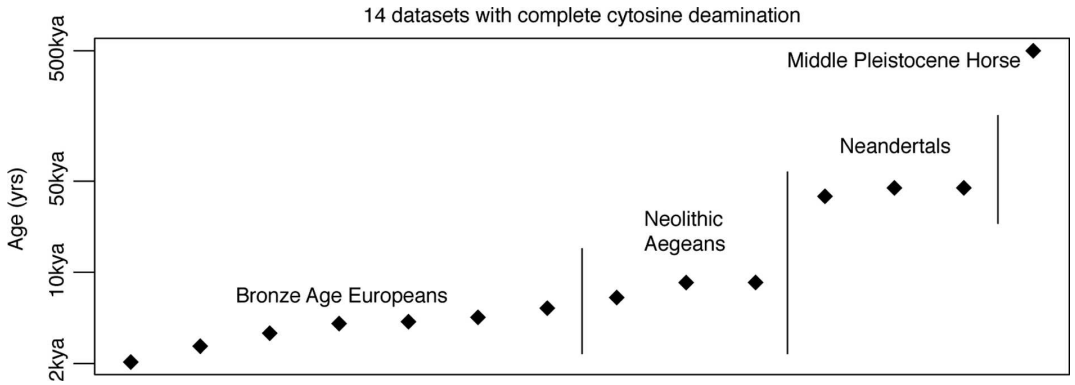

Figure 4. Fourteen samples included in the meta-analysis with saturated cytosine-to-uracil deamination in single-stranded overhangs.

In the example from that study (18), researchers inferred a nucleotide fragmentation rate (k) of 5.5×10^{-6} damage events per nucleotide per year on the basis of a strong relationship between age and 242 bp target molarity across a large set of mass-standardized bone samples. Under a rate-constant fragmentation model where $k = 5.5 \times 10^{-6}$, the probability of retaining any fragment of length L per year is the probability that no random breakage event occurs at any of its sites, or $(1-k)^L$ and the probability of fragment loss (k_L) is $1-(1-k)^L$. For a 242 bp target, therefore, $k_{242} = 1.33 \times 10^{-3}$—each year 0.0013 of remaining 242 bp templates are severed on average. However, this model assumes no time-dependent loss in DNA by mechanisms other than fragmentation; if fragmentation stabilizes and bulk depletion of endogenous DNA continues, a similar pattern would result. Specifically, if each fragment has a 0.0013 probability of being lost to bulk DNA movement rather than fragmentation, the same qPCR signal of decreasing target molarity over time would result.

Our results do not conflict with the previous experiment identifying a time-dependent decay behavior in relative copy number of a given fragment size, but decay behavior in the molarity of a target DNA fragment over time in a qPCR assay could be attributed to either rate-constant fragmentation (18) or bulk loss of endogenous DNA. We propose that bulk DNA loss is congruent with both the previous qPCR signal (18) and our meta-analysis, whereas exponential decay by fragmentation is not supported as the primary mechanism of DNA loss in our study. Therefore, we propose that much of the time-dependent nature of ancient DNA recovery may be due to bulk loss of DNA from tissue. Recent research focusing on the dense, non-vascularized petrous part of the temporal bone as a source of high endogenous DNA content (27,65) demonstrates that targeting 'semi-closed' systems with little opportunity for chemical exchange may be the best strategy to continue pushing the boundaries of DNA preservation by combating this diffusion process. This idea has also been illustrated in studies dealing with DNA preserved in hair, which is thought to confer a protective microenvironment that impedes biological degradation, leaching and possibly hydrolytic damage, and therefore often constitutes a good source of relatively high-quality endogenous DNA (32,67).

We suggest that rate-constant fragmentation through hydrolytic depurination is seldom the limiting factor to long-term DNA preservation, but we offer some caveats: fragmentation through depurination is a well-characterized process (63) and we do not propose that it is irrelevant for long-term DNA degradation. We suggest, rather, that the rate of this process is significantly slower than previously estimated in many ancient tissues (18) and the signal over the time span re-analyzed here is overprinted by other factors in a multi-faceted breakdown process. Thus when estimating the value of 'lambda' for a dataset—the parameter describing fragment length distribution—we are analyzing the outcome of multiple processes rather than inferring a simple decay rate. Further, importantly, any paleogenomic meta-analysis is fundamentally limited to those scenarios in which DNA actually survives over Quaternary timescales and so hydrolytic fragmentation as previously described might be a central mechanism for the total postmortem depletion of DNA in many tissues and conditions. That is to say, we can only analyze DNA that has survived, which may represent an abnormal mode of diagenesis. Our model for ancient DNA decay therefore necessarily speaks only to the special case in which conditions exist for long-term DNA survival. The immutable depurination process likely still imposes practical limits on DNA recovery in deep time and recovering Mesozoic DNA, for example, remains extremely unlikely. However, semi-closed chemical exchange systems like the petrous bone, though rare, offer excellent potential for the long-term retention of DNA in tissues and extraordinary preservational microenvironments created by chemical interactions have proven valuable for deep-time protein preservation (68). Breaching the current Middle Pleistocene age boundary of genomics seems entirely plausible.

AVAILABILITY

Analyzes were based on publicly available datasets. Summary data are available for reanalysis as Supplemental Dataset S1. Additionally, all metadata and results from analyzes, custom scripts and run logs from analyzes have been uploaded to the Dryad Digital Repository (doi:10.5061/dryad.5r10j).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank Ludovic Orlando, Beth Shapiro and Kes Schroer for comments on an early version of the manuscript.

FUNDING

Natural Environment Research Council Independent Research Fellowship [NE/L012030/1 to L.K.]. Funding for open access charge: Natural Environment Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Shapiro,B. and Hofreiter,M. (2014) A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science*, **343**, 1236573–1236573.
2. Leonardi,M., Librado,P., Der Sarkissian,C., Schubert,M., Alfarhan,A.H., Alquraishi,S.A., Al-Rasheid,K.A.S., Gamba,C., Willerslev,E. and Orlando,L. (2017) Evolutionary Patterns and Processes: Lessons from Ancient DNA. *Syst. Biol.*, **66**, e1–e29.
3. Haber,M., Mezzavilla,M., Xue,Y. and Tyler-Smith,C. (2016) Ancient DNA and the rewriting of human history: be sparing with Occam's razor. *Genome Biol.*, **17**, 1.
4. Skoglund,P., Mallick,S., Bortolini,M.C., Chennagiri,N., Hünemeier,T., Petzl-Erler,M.L., Salzano,F.M., Patterson,N. and Reich,D. (2015) Genetic evidence for two founding populations of the Americas. *Nature*, **525**, 104–110.
5. Haak,W., Lazaridis,I., Patterson,N., Rohland,N., Mallick,S., Llamas,B., Brandt,G., Nordenfelt,S., Harney,E., Stewardson,K. *et al.* (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, **522**, 207–211.
6. Palkopoulou,E., Mallick,S., Skoglund,P., Enk,J., Rohland,N., Li,H., Omrak,A., Vartanyan,S., Poinar,H., Götherström,A. *et al.* (2015) Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr. Biol.*, **25**, 1395–1400.
7. Lynch,V.J., Bedoya-Reina,O.C., Ratan,A., Sulak,M., Drautz-Moses,D.I., Perry,G.H., Miller,W. and Schuster,S.C. (2015) Elephantid Genomes Reveal the Molecular Bases of Woolly Mammoth Adaptations to the Arctic. *Cell Rep.*, **12**, 217–228.
8. Park,S.D.E., Magee,D.A., McGettigan,P.A., Teasdale,M.D., Edwards,C.J., Lohan,A.J., Murphy,A., Braud,M., Donoghue,M.T., Liu,Y. *et al.* (2015) Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. *Genome Biol.*, **16**, 234.
9. Kistler,L., Ratan,A., Godfrey,L.R., Crowley,B.E., Hughes,C.E., Lei,R., Cui,Y., Wood,M.L., Muldoon,K.M., Andriamialison,H. *et al.* (2015) Comparative and population mitogenomic analyzes of Madagascar's extinct, giant 'subfossil' lemurs. *J. Hum. Evol.*, **79**, 45–54.
10. Hofman,C.A., Rick,T.C., Fleischer,R.C. and Maldonado,J.E. (2015) Conservation archaeogenomics: Ancient DNA and biodiversity in the Anthropocene. *Trends Ecol. Evol.*, **30**, 540–549.
11. Palmer,S.A., Clapham,A.J., Rose,P., Freitas,F.O., Owen,B.D., Beresford-Jones,D., Moore,J.D., Kitchen,J.L. and Allaby,R.G. (2012) Archaeogenomic evidence of punctuated genome evolution in gossypium. *Mol. Biol. Evol.*, **29**, 2031–2038.
12. Kistler,L., Newsom,L.A., Ryan,T.M., Clarke,A.C., Smith,B.D. and Perry,G.H. (2015) Gourds and squashes (Cucurbita spp.) adapted to megafaunal extinction and ecological anachronism through domestication. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 15107–15112.
13. Schubert,M., Jónsson,H., Chang,D., Der Sarkissian,C., Ermini,L., Ginolhac,A., Albrechtsen,A., Dupanloup,I., Foucal,A., Petersen,B. *et al.* (2014) Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E5661–E5669.
14. Frantz,L.A.F., Mullin,V.E., Pionnier-Capitan,M., Lebrasseur,O., Ollivier,M., Perri,A., Linderholm,A., Mattiangeli,V., Teasdale,M.D., Dimopoulos,E.A. *et al.* (2016) Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, **352**, 1228–1231.
15. Orlando,L., Gilbert,M.T.P. and Willerslev,E. (2015) Reconstructing ancient genomes and epigenomes. *Nat. Rev. Genet.*, **16**, 395–408.
16. Warinner,C., Speller,C., Collins,M.J. and Lewis,C.M. (2015) Ancient human microbiomes. *J. Hum. Evol.*, **79**, 125–136.
17. Smith,O., Momber,G., Bates,R., Garwood,P., Fitch,S., Pallen,M., Gaffney,V. and Allaby,R.G. (2015) Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago. *Science*, **347**, 998–1001.
18. Allentoft,M.E., Collins,M., Harker,D., Haile,J., Oskam,C.L., Hale,M.L., Campos,P.F., Samaniego,J.A., Gilbert,M.T.P., Willerslev,E. *et al.* (2012) The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B*, **279**, 4724–4733.
19. Paabo,S., Higuchi,R.G. and Wilson,A.C. (1989) Ancient DNA and the polymerase chain reaction. The emerging field of molecular archaeology. *J. Biol. Chem.*, **264**, 9709–9712.
20. Poinar,H.N., Schwarz,C., Qi,J., Shapiro,B., MacPhee,R.D.E., Buigues,B., Tikhonov,A., Huson,D.H., Tomsho,L.P., Auch,A. *et al.* (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, **311**, 392–394.
21. Briggs,A.W., Stenzel,U., Johnson,P.L.F., Green,R.E., Kelso,J., Prufer,K., Meyer,M., Krause,J., Ronan,M.T., Lachmann,M. *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 14616–14621.
22. 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. *et al.* (2014) Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.*, **24**, 454–466.
23. Wales,N., Carøe,C., Soval-Velasco,M., Gamba,C., Barnett,R., Samaniego,J.A., Madrigal,J.R., Orlando,L. and Gilbert,M.T.P. (2015) New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA. *Biotechniques*, **59**, 368–371.
24. Jónsson,H., Ginolhac,A., Schubert,M., Johnson,P.L.F. and Orlando,L. (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, **29**, 1682–1684.
25. Wade,L. (2015) Breaking a tropical taboo. *Science*, **349**, 370–371.
26. Allentoft,M.E., Sikora,M., Sjögren,K.-G., Rasmussen,S., Rasmussen,M., Stenderup,J., Damgaard,P.B., Schroeder,H., Ahlström,T., Vinner,L. *et al.* (2015) Population genomics of bronze age Eurasia. *Nature*, **522**, 167–172.
27. Gamba,C., Jones,E.R., Teasdale,M.D., McLaughlin,R.L., Gonzalez-Fortes,G., Mattiangeli,V., Domboróczki,L., Kővári,I., Pap,I., Anders,A. *et al.* (2014) Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.*, **5**, 5257.
28. Gallego Llorente,M., Jones,E.R., Eriksson,A., Siska,V., Arthur,K.W., Arthur,J.W., Curtis,M.C., Stock,J.T., Coltorti,M., Pieruccini,P. *et al.* (2015) Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science*, **6**, 2647–2653.
29. Olalde,I., Allentoft,M.E., Sánchez-Quinto,F., Santpere,G., Chiang,C.W.K., DeGiorgio,M., Prado-Martinez,J., Rodríguez,J.A., Rasmussen,S., Quilez,J. *et al.* (2014) Derived immune and ancestral pigmentation alleles in a 7,000-year-old mesolithic European. *Nature*, **507**, 225–228.
30. Olalde,I., Schroeder,H., Sandoval-Velasco,M., Vinner,L., Lobón,I., Ramirez,O., Civit,S., García Borja,P., Salazar-García,D.C., Talamo,S. *et al.* (2015) A common genetic origin for early farmers from mediterranean cardial and central European LBK cultures. *Mol. Biol. Evol.*, **32**, 3132–3142.
31. Rasmussen,M., Anzick,S.L., Waters,M.R., Skoglund,P., DeGiorgio,M., Stafford,T.W., Rasmussen,S., Moltke,I., Albrechtsen,A., Doyle,S.M. *et al.* (2014) The genome of a late Pleistocene human from a Clovis burial site in western Montana. *Nature*, **506**, 225–229.
32. Rasmussen,M., Li,Y., Lindgreen,S., Pedersen,J.S., Albrechtsen,A., Moltke,I., Metspalu,M., Metspalu,E., Kivisild,T., Gupta,R. *et al.* (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, **463**, 757–762.
33. Günther,T., Valdiosera,C., Malmström,H., Ureña,I., Rodriguez-Varela,R., Sverrisdóttir,Ó.O., Daskalaki,E.A., Skoglund,P., Naidoo,T., Svensson,E.M. *et al.* (2015) Ancient

genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 11917–11922.
34. Keller,A., Graefen,A., Ball,M., Matzas,M., Boisguerin,V., Maixner,F., Leidinger,P., Backes,C., Khairat,R., Forster,M. *et al.* (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.*, **3**, 698.
35. Cassidy,L.M., Martiniano,R., Murphy,E.M., Teasdale,M.D., Mallory,J., Hartwell,B. and Bradley,D.G. (2016) Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 368–373.
36. Martiniano,R., Caffell,A., Holst,M., Hunter-Mann,K., Montgomery,J., Müldner,G., McLaughlin,R.L., Teasdale,M.D., van Rheenen,W., Veldink,J.H. *et al.* (2016) Genomic signals of migration and continuity in Britain before the Anglo-Saxons. *Nat. Commun.*, **7**, 10326.
37. Hofmanová,Z., Kreutzer,S., Hellenthal,G., Sell,C., Diekmann,Y., Díez-del-Molino,D., van Dorp,L., López,S., Kousathanas,A., Link,V. *et al.* (2016) Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 6886–6891.
38. Jeong,C., Ozga,A.T., Witonsky,D.B., Malmström,H., Edlund,H., Hofman,C.A., Hagan,R.W., Jakobsson,M., Lewis,C.M., Aldenderfer,M.S. *et al.* (2016) Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 7485–7490.
39. Yoshida,K., Schuenemann,V.J., Cano,L.M., Pais,M., Mishra,B., Sharma,R., Lanz,C., Martin,F.N., Kamoun,S., Krause,J. *et al.* (2013) The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *Elife*, **2**, e00731.
40. Enk,J., Devault,A., Debruyne,R., King,C.E., Treangen,T., O'Rourke,D., Salzberg,S.L., Fisher,D., MacPhee,R. and Poinar,H. (2011) Complete Columbian mammoth mitogenome suggests interbreeding with woolly mammoths. *Genome Biol.*, **12**, R51.
41. Green,R.E., Krause,J., Briggs,A.W., Maricic,T., Stenzel,U., Kircher,M., Patterson,N., Li,H., Zhai,W., Fritz,M.H.-Y. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–22.
42. Librado,P., Der Sarkissian,C., Ermini,L., Schubert,M., Jónsson,H., Albrechtsen,A., Fumagalli,M., Yang,M.A., Gamba,C., Seguin-Orlando,A. *et al.* (2015) Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E6889–E6897.
43. Orlando,L., Ginolhac,A., Zhang,G., Froese,D., Albrechtsen,A., Stiller,M., Schubert,M., Cappellini,E., Petersen,B., Moltke,I. *et al.* (2013) Recalibrating Equus evolution using the genome sequence of an early middle Pleistocene horse. *Nature*, **499**, 74–78.
44. Miller,W., Schuster,S.C., Welch,A.J., Ratan,A., Bedoya-Reina,O.C., Zhao,F., Kim,H.L., Burhans,R.C., Drautz,D.I., Wittekindt,N.E. *et al.* (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2382–E2390.
45. Dodt,M., Roehr,J.T., Ahmed,R. and Dieterich,C. (2012) FLEXBAR-flexible barcode and adapter processing for next-generation sequencing platforms. *Biology (Basel)*, **1**, 895–905.
46. Zhang,J., Kobert,K., Flouri,T. and Stamatakis,A. (2014) PEAR: a fast and accurate Illumina paired-end read mergeR. *Bioinformatics*, **30**, 614–620.
47. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
48. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
49. Hijmans,R.J., Cameron,S.E., Parra,J.L., Jones,P.G. and Jarvis,A. (2005) Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.*, **25**, 1965–1978.
50. Hijmans,R.R. (2015) raster: geographic data analysis and modeling. R Packag. version 2.4-18.
51. Plummer,M., Best,N., Cowles,K. and Vines,K. (2006) CODA: convergence diagnosis and output analysis for MCMC. *R. News*, **6**, 7–11.
52. Skoglund,P., Northoff,B.H., Shunkov,M. V, Derevianko,A.P., Pääbo,S., Krause,J. and Jakobsson,M. (2014) Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2229–34.
53. Deagle,B.E., Eveson,J.P. and Jarman,S.N. (2006) Quantification of damage in DNA recovered from highly degraded samples–a case study on DNA in faeces. *Front. Zool.*, **3**, 11–20.
54. Wang,H. and Song,M. (2011) Ckmeans.1d.dp: optimal k-means clustering in one dimension by dynamic programming. *R. J.*, **3**, 29–33.
55. Dabney,J. and Meyer,M. (2012) Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques*, **52**, 87–94.
56. Koslicki,D. (2011) Topological entropy of DNA sequences. *Bioinformatics*, **27**, 1061–1067.
57. Schmitt,a O. and Herzel,H. (1997) Estimating the entropy of DNA sequences. *J. Theor. Biol.*, **188**, 369–377.
58. Ornstein,R.L., Rein,R., Breen,D.L. and Macelroy,R.D. (1978) An optimized potential function for the calculation of nucleic acid interaction energies I. Base stacking. *Biopolymers*, **17**, 2341–2360.
59. R Development Core Team (2016) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, http://www.R-project.org/.
60. Sawyer,S., Krause,J., Guschanski,K., Savolainen,V. and Pääbo,S. (2012) Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. *PLoS One*, **7**, e34131.
61. Pääbo,S. (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 1939–1943.
62. Campos,P.F., Craig,O.E., Turner-Walker,G., Peacock,E., Willerslev,E. and Gilbert,M.T.P. (2011) DNA in ancient bone–where is it located and how should we extract it? *Ann. Anat. - Anat. Anzeiger*, **194**, 7–16.
63. Grass,R.N., Heckel,R., Puddu,M., Paunescu,D. and Stark,W.J. (2015) Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angew. Chem. Int. Ed.*, **54**, 2552–2555.
64. Ottoni,C., Koon,H.E.C., Collins,M.J., Penkman,K.E.H., Rickards,O. and Craig,O.E. (2009) Preservation of ancient DNA in thermally damaged archaeological bone. *Naturwissenschaften*, **96**, 267–278.
65. Hansen,H.B., Damgaard,P.B., Margaryan,A., Stenderup,J., Lynnerup,N., Willerslev,E. and Allentoft,M.E. (2017) Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum. *PLoS One*, **12**, e0170940.
66. Weiß,C.L., Schuenemann,V.J., Devos,J., Shirsekar,G., Reiter,E., Gould,B.A., Stinchcombe,J.R., Krause,J. and Burbano,H.A. (2016) Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *R. Soc. Open Sci.*, **3**, 160239.
67. Gilbert,M.T.P., Tomsho,L.P., Rendulic,S., Packard,M., Drautz,D.I., Sher,A., Tikhonov,A., Dalén,L., Kuznetsova,T., Kosintsev,P. *et al.* (2007) Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science*, **317**, 1927–30.
68. Demarchi,B., Hall,S., Roncal-Herrero,T., Freeman,C.L., Woolley,J., Crisp,M.K., Wilson,J., Fotakis,A., Fischer,R., Kessler,B.M. *et al.* (2016) Protein sequences bound to mineral surfaces persist into deep time. *Elife*, **5**, e17092.