



Cite this article: Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, Stinchcombe JR, Krause J, Burbano HA. 2016 Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *R. Soc. open sci.* **3**: 160239. <http://dx.doi.org/10.1098/rsos.160239>

Received: 5 April 2016

Accepted: 24 May 2016

Subject Category:

Genetics

Subject Areas:

evolution/genomics/plant science

Keywords:

ancient DNA, DNA damage, herbarium, DNA decay

Author for correspondence:

Hernán A. Burbano

e-mail: hernan.burbano@tuebingen.mpg.de

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rsos.160239> or via <http://rsos.royalsocietypublishing.org>.

Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens

Clemens L. Weiß¹, Verena J. Schuenemann³, Jane Devos², Gautam Shirsekar², Ella Reiter³, Billie A. Gould⁵, John R. Stinchcombe^{5,6}, Johannes Krause^{3,4,7} and Hernán A. Burbano¹

¹Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, and ²Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tuebingen 72076, Germany

³Institute of Archaeological Sciences, and ⁴Departments of Paleoanthropology and Archaeogenetics, Senckenberg Center for Human Evolution and Paleoenvironment, University of Tübingen, Tuebingen 72076, Germany

⁵Department of Ecology and Evolutionary Biology, and ⁶University of Toronto, Toronto, Ontario, Canada M5S

⁷Max Planck Institute for the Science of Human History, Jena 07743, Germany

HAB, 0000-0003-3433-719X

Herbaria archive a record of changes of worldwide plant biodiversity harbouring millions of specimens that contain DNA suitable for genome sequencing. To profit from this resource, it is fundamental to understand in detail the process of DNA degradation in herbarium specimens. We investigated patterns of DNA fragmentation and nucleotide misincorporation by analysing 86 herbarium samples spanning the last 300 years using Illumina shotgun sequencing. We found an exponential decay relationship between DNA fragmentation and time, and estimated a per nucleotide fragmentation rate of 1.66×10^{-4} per year, which is six times faster than the rate estimated for ancient bones. Additionally, we found that strand breaks occur specially before purines, and that depurination-driven DNA breakage occurs constantly through time and can to a great extent explain decreasing fragment length over time. Similar to what has been found analysing ancient DNA from bones, we found a strong correlation between the deamination-driven accumulation of cytosine to thymine substitutions and time, which reinforces the importance of substitution patterns to authenticate the ancient/historical nature of DNA fragments. Accurate estimations of DNA degradation through time

1. Introduction

Under favourable conditions DNA fragments can survive in plant [1] and animal tissues [2] for hundreds of thousands of years providing a molecular record of the past. Therefore, the examination of historical genomes permits the inclusion of temporal data into evolutionary studies, which allows a more accurate inference of rates and timing of key evolutionary events. The vast majority of ancient DNA (aDNA) studies have focused on animal remains, whereas plant remains have received less attention [3] despite the abundance of historic plant specimens.

DNA retrieved from historic specimens comes in small amounts and is a mixture of endogenous and microbial DNA that either was present pre-mortem or colonized the tissue post-mortem [4]. The aDNA comes in small fragment sizes [5] and holds various modifications that distinguish it from DNA extracted from fresh tissue [6]. DNA fragmentation is partially driven by spontaneous depurination and subsequent hydrolysis of the DNA backbone [7,8]. The typical sign of depurination, which is the excess of both adenine (A) and guanine (G) before DNA breaking points, has been detected by high throughput sequencing (HTS) in libraries constructed from aDNA [9]. DNA degradation is additionally marked by an increase of cytosine (C) to thymine (T) substitutions towards the end of aDNA fragments. This pattern results from spontaneous deamination of C residues to uracils (U) that are read as T by the polymerase and occur in higher proportion in single-stranded DNA overhangs [9,10]. A biochemical definition of aDNA includes all above-mentioned characteristics but does not delineate a time boundary between ancient and modern DNA [3].

It is interesting to understand quantitatively how these aDNA-associated patterns change through time, as they could be used to both authenticate DNA fragments retrieved from historic samples of different ages, and to calculate DNA decay rate based on their fragmentation patterns [2]. Using animal remains it has been found that there is a strong positive correlation between the amount of putative deamination (excess of C to T substitutions) and the sample age [11]. Hence, the excess of C to T substitutions has been used as a criterion of authenticity in aDNA studies [12]. The correlation between other aDNA-associated patterns and sample age is weaker [11], which could be a consequence of the different environmental conditions in which fossils were preserved, processed and stored (e.g. freshly excavated ancient bones are best for amplification of aDNA [13]). To reduce the effect of environmental variation on DNA degradation, a more spatially constrained sample of animal remains has been studied [2]. Allentoft *et al.* [2] calculated the long-term DNA decay rate in bone tissue, which could be used to estimate DNA half-life and, consequently, to put a boundary on how far back in the past DNA could be theoretically retrieved. Because most of the bone samples in Allentoft *et al.* [2] were analysed only by quantitative PCR and not by HTS, it was not possible to investigate how the signals left by deamination or depurination correlate with time in a spatially constrained sample.

Since herbaria contain time snapshots of global biodiversity and could be informative to address a broad spectrum of biological questions, it is fundamental to understand how DNA survives in this type of specimens. Investigating the effect of time on DNA degradation is normally difficult, since environmental conditions such as temperature, pH and humidity, among others, affect DNA stability [14]. Therefore, it is highly advantageous that herbaria samples are prepared and stored using standardized procedures, which reduces the effect of environmental variation among herbaria samples compared with ancient bones. Consequently, herbarium samples are ideal to study the temporal patterns of damage and decay kinetics of DNA. In this study, we analysed 86 herbarium samples collected over the last 300 years using library-based methods coupled with HTS, and produce for the first time an in-depth description of aDNA-associated patterns and its dynamics through time. Additionally, we use the power of multiple DNA sequencing libraries to calculate DNA decay rate in plant desiccated tissue.

2. Results

2.1. DNA fragmentation

We used a group of multiple species herbarium samples and also freshly prepared (less than 1 year old) herbarium samples of *Arabidopsis thaliana* dried using a wooden press (table 1; electronic supplementary

Table 1. Type and number of herbarium samples.

type of sample	species	number of samples	collection year (range)	number of infected samples ^a
historic	<i>Arabidopsis thaliana</i>	54	1863–1993	—
	<i>Solanum tuberosum</i>	12	1845–1896	12
	<i>Solanum lycopersicum</i>	5	1737–1876	2
	total	71	1737–1993	14
modern	<i>Arabidopsis thaliana</i>	15	2014	—

^aSamples with lesions are compatible with *Phytophthora infestans* lesions.

material, table S1). From here on, we will refer to these groups as historic and modern herbarium, respectively. A fraction of historic herbarium samples from Solanaceae have lesions compatible with *Phytophthora infestans* infection and have been previously studied [15]. Since we expect that DNA retrieved from historic samples will be highly fragmented, it is likely that a fraction of the molecule will be covered by both the forward and reverse read. After adapter trimming forward and reverse reads were merged, requiring an overlap of 10 base pairs (bp) between them. We were able to merge on average 96% (83–99%) of the reads from historic herbarium samples, whereas on average 21% (18–40%) of reads could be merged from modern herbarium samples, due to the presence of much longer DNA fragments (electronic supplementary material, table S2). In the modern herbarium samples, the mean of the fragment length distribution corresponded to the fragment size intended during sonication (400 bp) and the merged reads were located at the left tail of the fragment length distribution (electronic supplementary material, figure S1). For all further analysis correlating DNA fragmentation with time we used only merged reads from historic herbarium samples.

The distribution of fragment lengths of merged reads is not normally distributed and could be better described by a lognormal distribution (figure 1a). To evaluate the correlation between fragment lengths with the collection year of each sample, we chose the log-mean value of a fitted lognormal distribution. The regression between the log-mean fragment length and the sample collection year was statistically significant ($R^2 = 0.2$; $p = 6.33 \times 10^{-5}$; $N = 71$; figure 1b). To check if the signal was driven only by the oldest eighteenth century *Solanum lycopersicum* samples (figure 1b), we repeated the analysis only for the *A. thaliana* samples and found that the regression was still significant ($R^2 = 0.175$; $p = 1.6 \times 10^{-3}$; $N = 54$), which implies that the signal arises from the whole set of herbarium specimens and is not driven only by the oldest samples. Since DNA was extracted from some herbarium specimens using CTAB (cetyl-trimethyl ammonium bromide) and PTB (*N*-phenacylthiazolium bromide) extraction protocols [16], we evaluate the effect of these methods on the length distribution of DNA reads and found no difference between them ($p = 0.75$; $N = 54$; electronic supplementary material, figure S2).

2.2. DNA break points

It is possible using reads mapped to their respective reference genome to analyse the genomic nucleotide context surrounding the ends of the DNA fragments, and thus look indirectly at DNA break points. We found an excess of purine frequency (both adenine and guanine) in DNA retrieved from historic herbarium samples at position -1 (5' end; electronic supplementary material, figure S3a). This pattern was not found in modern herbarium samples (electronic supplementary material, figure S4a), hence in all further analysis correlating DNA breaking points and time we used only historic herbarium samples. We calculated the relative enrichment in purine frequency of both adenine and guanine at position -1 compared with position -5 . We then correlate these signatures of depurination with the collection year of the sample. Neither adenine (electronic supplementary material, figure S3b) nor guanine (electronic supplementary material, figure S3c) relative enrichment showed a significant correlation with collection year. Additionally, we did not find a difference between the average relative enrichment of adenine when compared with guanine (electronic supplementary material, figure S3b,c). When we analyse independently chloroplast-derived reads, we found purine enrichment at position -1 , and no correlation between the relative enrichment of purines and collection year. There were no significant differences between nuclear- and chloroplast-derived reads ($p(\text{adenine}) = 0.34$; $p(\text{guanine}) = 0.7$; $N = 71$; electronic supplementary material, figure S5).

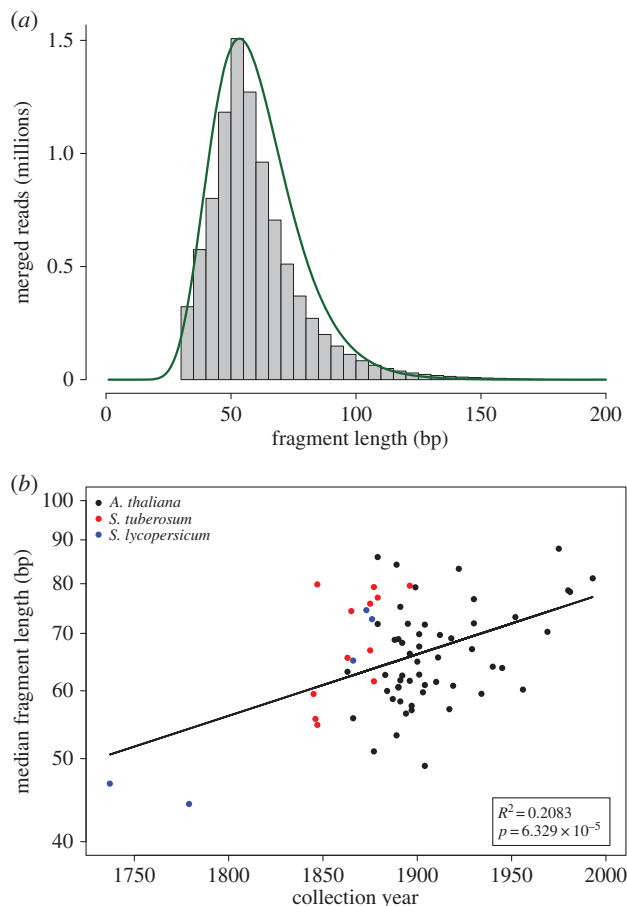


Figure 1. DNA fragmentation. (a) Distribution of fragment lengths of merged reads from *A. thaliana* sample NY1365354. The green line shows the fit between the empirical and the lognormal distribution. (b) Median length of merged reads as a function of collection year ($N = 71$). The line indicates the linear regression. The inset shows the regression statistics between the natural logarithm of median length and collection year. The y -axis is log-scaled and shows, therefore, that the correlation is exponential.

2.3. DNA decay rate

The length distribution in aDNA libraries shows an exponential decline as the result of random fragmentation (figure 2a) [17]. After logarithmic transformation of the fragment length frequencies, the exponential decline can be described by a linear function with slope λ , which corresponds to the damage fraction per site (figure 2b). Damage should be interpreted here as DNA bond breaking. To get the overall decay rate for all herbarium samples, we analysed the relationship between λ and the age of each sample and found a linear relation. The slope corresponds to the overall per nucleotide decay rate $k = 1.66 \times 10^{-4}$ per year ($R^2 = 0.26$; $p = 6.46 \times 10^{-6}$; $N = 71$; figure 2c), which turned out to be six times faster than the rate estimated based on ancient bones, $k = 2.71 \times 10^{-5}$ per site per year [2].

2.4. Nucleotide misincorporation

The most abundant miscoding lesions in aDNA are C to T substitutions, which is caused by deamination of C to U. The U is then read as T by the polymerase during sequencing [9]. The excess of C to T substitutions occurs primarily at the ends of the reads and declines exponentially inwards. We found this pattern present in all historic herbarium samples analysed (figure 3a), but absent in modern herbarium samples (electronic supplementary material, figure S4b). For all further analysis correlating nucleotide misincorporation with time, we used only historic herbarium samples. Since the excess of C to T substitutions is more manifest at first base, we chose, as previously described [11], the percentage of C to T substitutions at first base as a proxy for miscoding lesions and correlate this value with the samples' collection year (figure 3b). We found a very strong linear relationship between these two values ($R^2 = 0.45$; $p = 1.44 \times 10^{-10}$; $N = 71$). As it was previously carried out in the DNA fragmentation

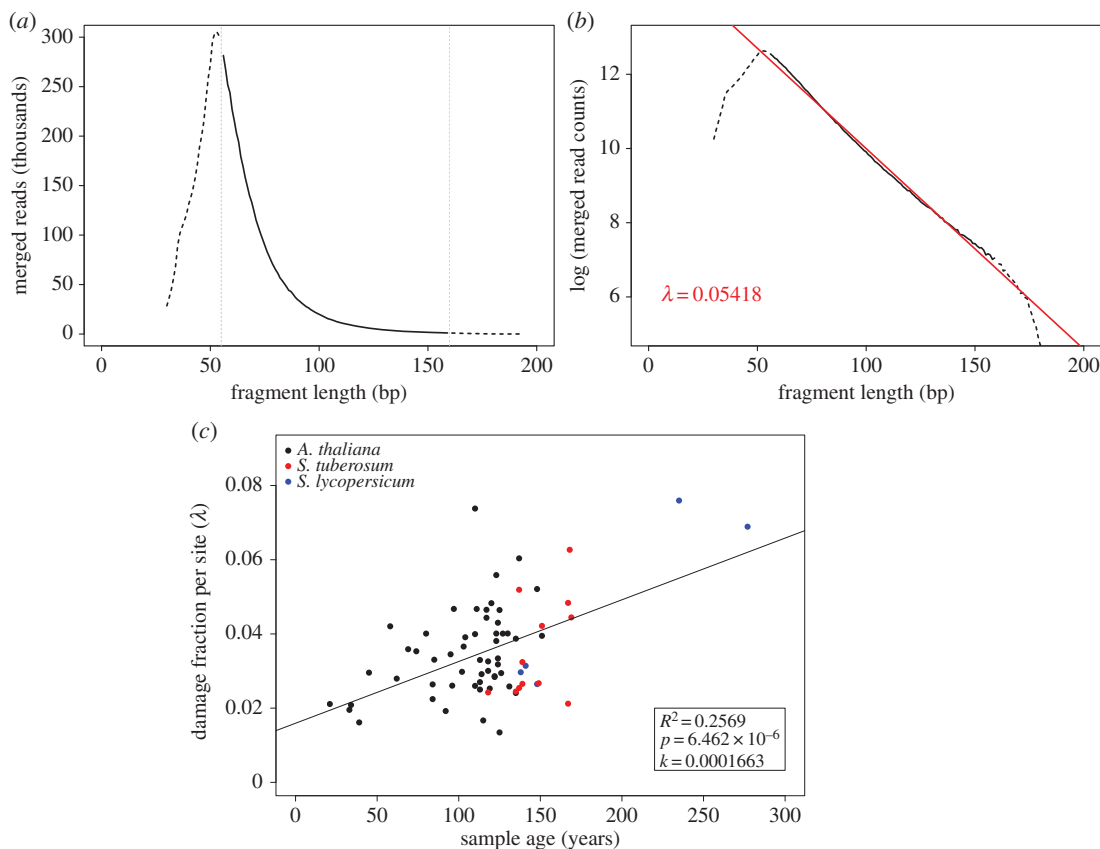


Figure 2. DNA fragmentation and decay rate. (a) Distribution of fragment lengths of merged reads from *A. thaliana* sample NY1365354. The solid line, which is surrounded by horizontal dotted lines, shows the part of the distribution that follows an exponential decline. (b) Distribution of fragments length for the same library using a y-axis with a logarithmic scale. The slope of the exponential part of the distribution (red line) corresponds to the damage fraction per site (λ). (c) Damage fraction per site (λ) as a function of sample age ($N = 71$). The slope of the regression corresponds to the DNA decay rate (k) following the formula: $\lambda = k \times \text{age}$.

part, we sought to investigate the effect of the oldest eighteenth century *S. lycopersicum* samples in the correlation between deamination and time. Therefore, we repeated the analysis using only the *A. thaliana* samples and found that the regression was weaker but still significant ($R^2 = 0.27$; $p = 5.5 \times 10^{-5}$; $N = 54$). Again, we conclude that the signal arises from the whole set of samples and is not driven only by the oldest samples.

For the infected samples, we also calculated the percentage of C to T substitutions in *P. infestans* derived reads at first base and found the same signature, although it was weaker than the signal found in their host plant (electronic supplementary material, figure S6).

2.5. Differences between nuclear- and chloroplast-derived reads

We found that chloroplast-derived reads showed a slightly lower decay rate than the nuclear-derived reads ($k_{\text{chloroplast}} = 1.29 \times 10^{-4}$; electronic supplementary material, figure S7a). To test if the two decay rates were different, we performed an ANOVA that showed significant effects of both sample age and origin of DNA (nuclear- or chloroplast-derived) on the rate of bond breaking (λ) ($Pr(\text{sample age}) = 4.84 \times 10^{-8}$, $Pr(\text{DNA origin}) = 0.012$; $N = 71$). However, the effect of DNA origin was very small and there was no significant interaction between sample age and DNA origin ($Pr(\text{Sample age: DNA origin}) = 0.46$; $N = 71$). This indicates that the slopes of the two regressions, which correspond to the decay rate k , do not differ significantly.

The chloroplast-derived reads show a lower excess of C to T substitutions than the nuclear-derived reads (electronic supplementary material, figure S7b). The ANOVA showed in this case highly significant effects of sample age and DNA origin on the percentage of deamination at first base ($Pr(\text{sample age}) = 1.78 \times 10^{-14}$, $Pr(\text{DNA origin}) = 5.69 \times 10^{-9}$; $N = 71$). However, there was no significant interaction between sample age and DNA origin ($Pr(\text{sample age: DNA origin}) = 0.075$; $N = 71$). This indicates that

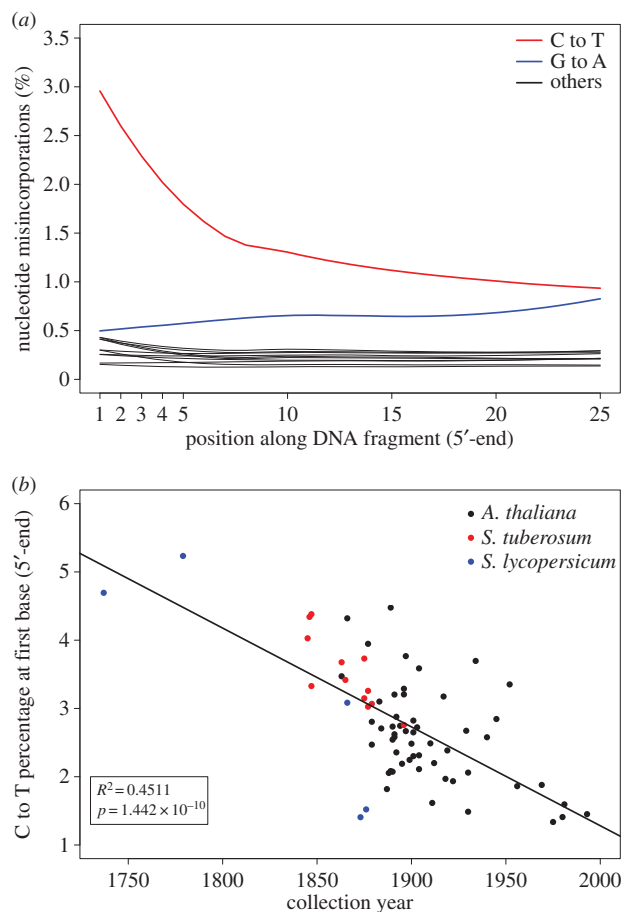


Figure 3. Nucleotide misincorporation. (a) Nucleotide misincorporation profile at 5'-end of the reads of *A. thaliana* sample NY1365354. The red line shows an excess of C to T substitutions at the beginning of the read that declines exponentially inwards. (b) C to T percentage at first base (5'-end) as a function of collection year ($N = 71$). The C to T percentage and the collection year have a linear relationship.

nuclear- and chloroplast-derived sequences differ significantly in the extent of deamination (the intercept of the regressions), but not in its rate (slope of the regressions).

3. Discussion

Herbaria contain millions of dried plant specimens that provide a record of worldwide changes in biodiversity spanning five centuries. Although plants were not originally collected and stored for genetic studies, the value of these collections as source of DNA has been long recognized by plant biologists [18]. There are a larger number of studies that have used PCR-based approaches to survey these collections, but only a handful of endeavours have used library-based methods coupled with HTS [15,19,20]. Since herbaria collections are an invaluable source of genetic information, it is important to investigate in detail both the properties of DNA retrieved from them and the rate at which DNA damage takes place through time.

3.1. DNA fragmentation and decay rate

We confirmed the highly fragmented nature of DNA retrieved from herbarium samples [15,20] (figure 1a). The DNA fragmentation is comparable with the level found in animal remains that are several hundreds or even thousands of years old [11], although our samples are utmost 278 years old. In contrast with animal remains [11], we found a weak but significant exponential relation between fragment length and collection year, where more recent samples have longer DNA fragments (figure 1b). The lower levels of environmental variation experienced by herbarium samples relative to animal remains could have increased the signal-to-noise ratio allowing the detection of the relation between time and DNA fragmentation.

Since depurination can be inferred by examining DNA breaking points in HTS data [9,11], we analysed our sequencing libraries for an excess of purines at genomic positions surrounding sequencing reads. Both A and G were found overrepresented upstream of the 5'-end break points (electronic supplementary material, figure S3a), but no correlation was found between the relative fold enrichment of either A or G and collection year (electronic supplementary material, figure S3b,c), which implies that the contribution of depurination to DNA breakage does not change through time.

DNA decay and degradation can be understood as a two-step process, with a first rapid phase where the damage is caused mainly by nucleases and digestion by microorganisms, and a second phase where the damage is driven by hydrolytic and oxidative reactions that occur at a much lower rate than the first phase [21]. The correlation between fragmentation and time might be the result of a process occurring in the second phase that can be only detected in samples that have experienced very similar environments, as is the case for herbarium samples.

Modern herbarium samples did not show any age-related fragmentation or excess of purines at DNA breaking points. In fact, the distribution of fragment lengths was centred on the intended fragment length during sonication (electronic supplementary material, figure S1). It has been suggested based on PCR-based methods that most DNA fragmentation in herbarium samples occurs during sample desiccation (drying at 60°C for 18 h) before they are fixed on herbarium sheets, and only a small portion of damage could be attributed to long-term storage [22]. We did not find the sample preparation effect in our herbarium samples however, on the contrary to previous studies [22], we did not use heating to dry our herbarium samples, as it is well established that heat increases the rate of depurination and subsequently β elimination leading to DNA strand breaks [8].

We found that the DNA decay rate in herbarium samples is about six times faster than the rate in bones [2]. It is possible that this big difference could be explained by the characteristic nature of each tissue. In bone, DNA is adsorbed to hydroxyapatite, which decreases the rate of depurination compared with free DNA [14]. Additionally hydroxyapatite binds nucleases [23], which further prevents DNA degradation, especially in the first rapid phase of DNA degradation. DNA in plants' desiccated tissue might be less protected and more exposed to enzymatic and chemical damages. Furthermore, the vast majority of herbarium samples are not mounted on acid-free paper. Acidic paper was regularly used, which could have contributed to DNA degradation, as acid pH increases the rate of depurination *in vitro* [8]. We calculated independently nuclear and chloroplast DNA decay rates and found that the chloroplast DNA decay rate is 0.75 times the nuclear rate (electronic supplementary material, figure S7a). In ancient bone, the mitochondrial DNA decay rate is 2–2.5 times slower than the nuclear one [2], in agreement with a study that reported a better preservation of mitochondrial relative to nuclear DNA in permafrost mammoth remains [24]. The slower decay rate in organelle DNA might be a consequence of its circular structure, which makes DNA less accessible to endonucleases [2]. An early report of equal rates of degradation between nuclear and chloroplast DNA in herbarium samples was based on a smaller dataset only interrogated by PCR-based methods, and could be a consequence of lacking experimental resolution [22].

3.2. DNA misincorporation

We observed an increase in the percentage of C to T substitutions at the end of the molecule (figure 3a) and found a strong correlation between deamination and age (figure 3b), as has been found using animal remains [11]. Although chloroplast reads were less deaminated, the correlation between deamination and age held also for them (electronic supplementary material, figure S7b). Notably, modern herbarium samples did not show excess of any misincorporation and resembled DNA extracted from fresh tissue.

Since the signal of C deamination has been found recurrently in aDNA studies and there is a strong positive relationship between deamination and sample age [11], the presence of deamination patterns in aDNA HTS studies has been proposed as an authenticity criterion [12]. It is remarkable that C to T substitutions from both animal remains [11] and our data correlate strongly with time, although at a different rate in the two tissues, which implies that deamination is strongly related to the phase two of slow DNA degradation. An excess of C to T substitution at the end of the molecule has been also found in plant [15] and human pathogens [25–27] DNA. We found here that the deamination in plant-pathogen-derived reads is intermediate between nuclear- and chloroplast-derived reads (electronic supplementary material, figure S6). However, we think that the signal is sufficient to be used as an authenticity criterion. We do not suggest that a sample of given age should match a given level of deamination, but instead propose that the excess of C to T at the DNA ends, independent of its magnitude, should be presented as evidence for authenticity. In the future—given an appropriate depth of coverage—it might be possible to

also use deamination patterns to authenticate metagenomic aDNA derived from plant or animal tissue, or from environmental DNA profiling.

3.3. Practical implications

On the contrary to historic herbarium samples, modern herbarium samples resembled DNA extracted from fresh tissue, which shows that drying by pressing is an ideal method to collect plant samples in long field trips. This also implies that the magnitude of damage that happens in the first phase is highly dependent on the method used to prepare the herbarium specimen.

DNA misincorporations can be confused with natural variation, which will compromise variant calling and increase terminal branches in a phylogenetic context. Both effects are especially prominent in highly deaminated (old) samples that are sequenced at low coverage. Fortunately, it is now possible to almost eliminate this source of error either by removing uracils from DNA molecules during library preparation [28] or by statistically distinguishing true variants from aDNA-associated misincorporations post-sequencing, in reads derived from single-stranded library preparation methods [29].

Since DNA in dried tissue degrades rapidly, the retrieval of DNA from very old samples will require the use of DNA and library extraction preparation methods capable of recovering short length molecules [29,30]. The high DNA fragmentation of historic herbarium samples poses a challenge to genome reduced-representation methods such as RAD (restriction site associated markers)-sequencing [31,32], which has shown low DNA yields and low percentage of reads that could be mapped to the reference genome [33]. Thanks to improvements in library preparation and HTS accuracy, it is possible to sequence and perform mapping-guided assemblies of complete genomes from historic specimens with quality that matches genomes derived from modern specimens and, therefore, exploit the millions of plant remains stored in herbaria worldwide.

4. Material and methods

4.1. Previously published datasets

Sequences derived from *Solanum tuberosum* and *Solanum lycopersicum* infected by *P. infestans* are deposited in the European Nucleotide Archive, with accession number PRJEB1877.

4.2. New datasets

New DNA sequences are deposited in the European Nucleotide Archive, with accession number PRJEB9878

4.3. Herbarium samples

Historic herbarium samples were either directly sampled by us in different herbaria both in North America and Europe, or sampled there by collection curators and sent to us by post (electronic supplementary material, table S1). The amount of tissue used for destructive sampling ranged from 2 to 8 mm².

Modern herbarium samples were derived from a recent collection of *A. thaliana* wild populations in North America by the Max Planck Institute for Developmental Biology. After collection plant tissue was dried by pressing between acid-free papers using a wooden press for 6 weeks and subsequently mounted in herbarium sheets.

4.4. DNA extraction, library preparation and sequencing

4.4.1. DNA extraction from historic herbarium samples

DNA extractions were carried out in clean room facilities in all cases. The majority of the samples were extracted following the PTB extraction protocol [16] as previously described [15]. Samples from the Cornell Bailey Hortorium were extracted using the CTAB extraction protocol [16] (electronic supplementary material, table S2).

4.4.2. DNA extraction of modern herbarium samples

DNA extractions were carried out following the PTB extraction protocol [16].

4.4.3. Library preparation historic samples

Illumina double indexed sequencing libraries [34,35] were prepared from each sample as previously described [15]. The excess of C-to-T substitutions associated with DNA damage and caused by deamination of cytosines [36] was not repaired in order to quantify the amount of damage present in samples of different ages.

4.4.4. Library preparation modern herbarium samples

Indexed libraries were prepared using the Illumina TruSeq Nano DNA sample preparation kit following the manufacturer's instructions.

4.4.5. Sequencing

Libraries were paired-end sequenced on the Illumina HiSeq 2000, HiSeq 2500 or MiSeq instruments (electronic supplementary material, table S2).

4.5. Read processing and mapping

4.5.1. Historic herbarium samples

Reads were assigned to each sample based on their indices. Adapters were trimmed using the program SKEWER (v. 0.1.120) with default settings with the natively implemented Illumina TruSeq adapter sequences [37]. Forward and reverse reads were merged using the program FLASH (v. 1.2.11) with default settings, except for an elevated maximum overlap (100–150 bp depending on read length) to allow a more accurate scoring of highly overlapping read pairs [38]. Merged reads were mapped as single-end reads to their respective reference genomes: *Arabidopsis thaliana* [39,40], *S. tuberosum* [41], *S. lycopersicum* [42], *P. infestans* [43]. The mapping was performed using BWA-MEM (v. 0.7.10) with default settings [44]. PCR-duplicates were identified after mapping based on start and end coordinates and for every cluster of duplicate reads a consensus sequence was generated [45].

4.5.2. Modern herbarium samples

Reads were processed very similarly to the reads that belong to historic samples. The vast majority of reads could not be merged, which indicates that the DNA was not as fragmented as in older herbarium samples. Therefore, we mapped the paired-end reads using BWA-MEM (v. 0.7.10) with default parameters [44] and inferred fragment size based on paired-end mapping.

4.6. Analysis of DNA damage patterns

4.6.1. Fragment length

We analysed the fragment length distributions of merged reads. We fitted a lognormal distribution to the empirical fragment length distributions using the `fitdistr` function from the package MASS using R. Since in a lognormal distribution the logarithm of a variable is normally distributed, we used the mean of this distribution (log-mean) to summarize the fragment length distribution. The regression on the relationship among log-mean of fragment lengths and collection year was carried out using the `lm` function in R. For visualization (figure 1b), we used the fragment length median on a log-scaled y -axis, since the median is more intuitive to understand than the log-mean value. The relationship between log-mean and median follows the formula: $\text{median} = e^{\log\text{-mean}}$.

4.6.2. DNA break points

To analyse the nucleotide genomic context around DNA break points, we used the software MAPDAMAGE 2.0 (v. 2.0.2–12) [46]. MAPDAMAGE calculates the genomic base frequencies around mapped reads and within reads, which allows the inference of the bases most likely to be present before DNA break points. We calculated the relative enrichment of either adenine or guanine at the

5'-end (position -1 compared with position -5). The frequencies of both adenine and guanine were extracted from the output file `dnacomp.txt` produced by MAPDAMAGE. The regression on the relationship among purine relative enrichment (either adenine or guanine) and collection year was carried out using the `lm` function in R. The whole procedure was carried out for plant nuclear and chloroplast reads independently.

4.6.3. Nucleotide misincorporation

All types of nucleotide misincorporations relative to the reference genome were calculated per library using MAPDAMAGE 2.0 (v. 2.0.2–12) [46]. The percentage of C to T substitutions at first base was extracted from the output file `5pCtoT_freq.txt` produced by MAPDAMAGE. The regression on the relationship among the percentage of C to T substitutions at first base (5'-end) and collection year was carried out using the `lm` function in R. For the regression we used the percentage of deamination at first base. The whole procedure was carried out for plant nuclear and chloroplast reads independently, and also for pathogen nuclear reads in the case of samples infected with *P. infestans*.

4.7. Calculation of DNA decay rate

To calculate the decay rate of DNA retrieved from plant desiccated tissue, we used a previously described methodology [2] and adapted it to multiple samples. The random fragmentation of DNA molecules that occurs post-mortem follows a model of exponential decay, i.e. the amount amplifiable template decreases exponentially with increasing length [17]. We used the distributions of fragment length (L) of mapped reads to calculate the DNA decay rate, which is determined by the proportion of damage sites (λ). Thus, the process can be described using an exponential distribution:

$$F(L) = F_0 \times e^{-\lambda L}, \quad (4.1)$$

where L is the fragment length, $F(L)$ the frequency of fragment with length L and F_0 the frequency intersect at length 0.

After logarithmic transformation there is a linear relationship between the logarithms of the fragment frequency and fragment length with a slope $-\lambda$:

$$\log(F(L)) = \log(F_0) - \lambda L. \quad (4.2)$$

In this relationship, λ describes the fraction of bond survival per base in a single sample/library [2,17]. As previously described [2], the DNA decay rate per base per year, k , can then be calculated as:

$$k = \frac{\lambda}{\text{age}}. \quad (4.3)$$

We calculated the decay rate across all analysed samples taking advantage of the negative correlation between fragment length and age of the sample. We plotted the damage fraction per site (λ) as a function of sample age. The slope of the linear regression on the relationship among λ and samples age yields k , the decay rate, according to the linear relationship:

$$\lambda = k \times \text{age}. \quad (4.4)$$

The whole procedure was carried out for plant nuclear and chloroplast reads independently.

4.8. Analysis of covariance

To test if the regressions between chloroplast- and nuclear-derived reads were significantly different, we performed an analysis of covariance. We used the `'aov'` function in R to test models where the sample age was the covariate and the DNA origin (chloroplast- or nuclear-derived) was the factor. In the first step, a model of type `'y ~ covariate × factor'` was used to include a possible interaction between covariate and factor, which would mean that there is a difference in the slope of the regression depending on the factor. If no significant interaction was detected, the `'anova'` command in R was used to test this model against a model of type `'y ~ covariate + factor'`. This last model does not include the interaction, therefore we can test whether the removal of the interaction has an effect on the fit of the model. If not, the second model was accepted with the conclusion that the regressions do not differ in slope, but possibly in their intersects (if there is a significant effect of the factor on the dependent variable y).

To test whether the linear regressions of read lengths and collection year between samples extracted with CTAB and PTB methods were different, we used the same approach as for chloroplast- and nuclear-derived reads. In this comparison, we used extraction method as the factor in the linear model.

Data accessibility. New DNA sequences are deposited in the European Nucleotide Archive, with accession no. PRJEB9878. Authors' contributions. C.L.W. and H.A.B. conceived and designed the experiments. V.J.S., E.R., G.S., J.D. and B.A.G. performed DNA extractions; V.J.S., E.R. and J.D. performed libraries; J.R.S. and J.K. coordinated laboratory experiments; C.L.W. and H.A.B. analysed the data; C.L.W. and H.A.B. wrote the paper with help from all authors. Competing interests. We have no competing interests.

Funding. This work was supported by the Presidential Innovation Fund of the Max Planck Society.

Acknowledgements. We thank Moisés Espósito-Alonso, Rafal Gutaker and Rebecca Schwab for helpful comments on the manuscript; Matthias Meyer and Petra Korlević for input on data analysis and guidance through the specialized literature; Michael Dannemann for statistical advice; Magdalena Grenda for input on herbaria conservation. We are grateful to all researchers and curators for providing the historic specimens: Marco Thines (Goethe Universität, Frankfurt, Germany), Sandra Knapp (Natural History Museum, London, UK), Bryn Dentinger (Royal Botanical Gardens, Kew, UK), Dagmar Triebel (Botanische Staatssammlung, Munich, Germany), Anna Stalter (Cornell Bailey Hortorium, Ithaca, USA), John Freudenstein (The Ohio State University, Columbus, USA), Christine Niezgodá (The Field Museum, Chicago, USA), Robert Capers (University of Connecticut, Storrs, USA), Carol Ann McCormick (University of North Carolina, Chapel Hill, USA), Michael S. Dossman (Arnold Arboretum, Boston, USA), Cathy M. Herring (Agricultural Research Station, North Carolina State University, Clayton, USA) and John Peter (New York Botanical Garden, New York, USA). We thank Oliver Haddrath for laboratory space for DNA extraction in the Royal Ontario Museum.

References

- Palmer SA, Smith O, Allaby RG. 2012 The blossoming of plant archaeogenetics. *Ann. Anat.* **194**, 146–156. (doi:10.1016/j.aanat.2011.03.012)
- Allentoft ME *et al.* 2012 The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B* **279**, 4724–4733. (doi:10.1098/rspb.2012.1745)
- Shapiro B, Hofreiter M. 2014 A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science* **343**, 1236573. (doi:10.1126/science.1236573)
- Poinar HN *et al.* 2006 Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* **311**, 392–394. (doi:10.1126/science.1123360)
- Paabo S *et al.* 2004 Genetic analyses from ancient DNA. *Annu. Rev. Genet.* **38**, 645–679. (doi:10.1146/annurev.genet.37.110801.143214)
- Dabney J, Meyer M, Paabo S. 2013 Ancient DNA damage. *Cold Spring Harb. Perspect. Biol.* **5**, a012567. (doi:10.1101/cshperspect.a012567)
- Lindahl T, Andersson A. 1972 Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. *Biochemistry* **11**, 3618–3623. (doi:10.1021/bi00769a019)
- Lindahl T, Nyberg B. 1972 Rate of depurination of native deoxyribonucleic acid. *Biochemistry* **11**, 3610–3618. (doi:10.1021/bi00769a018)
- Briggs AW *et al.* 2007 Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA* **104**, 14 616–14 621. (doi:10.1073/pnas.0704665104)
- Brotherton P, Endicott P, Sanchez JJ, 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 Res.* **35**, 5717–5728. (doi:10.1093/nar/gkm588)
- Sawyer S, Krause J, Guschanski K, Savolainen V, Paabo S. 2012 Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE* **7**, e34131. (doi:10.1371/journal.pone.0034131)
- Krause J, Briggs AW, Kircher M, Maricic T, Zwyns N, Derevianko A, Paabo S. 2010 A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr. Biol.* **20**, 231–236. (doi:10.1016/j.cub.2009.11.068)
- Pruvost M, Schwarz R, Correia VB, Champlot S, Bragaier S, Morel N, Fernandez-Jalvo Y, Grange T, Geigl EM. 2007 Freshly excavated fossil bones are best for amplification of ancient DNA. *Proc. Natl Acad. Sci. USA* **104**, 739–744. (doi:10.1073/pnas.0610257104)
- Lindahl T. 1993 Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715. (doi:10.1038/362709a0)
- Yoshida K *et al.* 2013 The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *eLife* **2**, e00731. (doi:10.7554/eLife.00731)
- Kistler L. 2012 Ancient DNA extraction from plants. *Methods Mol. Biol.* **840**, 71–79. (doi:10.1007/978-1-61779-516-9_10)
- Deagle BE, Eveson JP, Jarman SN. 2006 Quantification of damage in DNA recovered from highly degraded samples—a case study on DNA in faeces. *Front. Zool.* **3**, 11. (doi:10.1186/1742-9994-3-11)
- Rogers SO, Bendich AJ. 1985 Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol. Biol.* **5**, 69–76. (doi:10.1007/BF00020088)
- Martin MD *et al.* 2013 Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. *Nat. Commun.* **4**, 2172.
- Staats M, Erkens RH, van de Vossen B, Wieringa JJ, Kraaijeveld K, Stielow B, Geml J, Richardson JE, Bakker FT. 2013 Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS ONE* **8**, e69189. (doi:10.1371/journal.pone.0069189)
- Molak M, Ho SY. 2011 Evaluating the impact of post-mortem damage in ancient DNA: a theoretical approach. *J. Mol. Evol.* **73**, 244–255. (doi:10.1007/s00239-011-9474-z)
- Staats M, Cuenca A, Richardson JE, Vrieling-van Ginkel R, Petersen G, Seberg O, Bakker FT. 2011 DNA damage in plant herbarium tissue. *PLoS ONE* **6**, e28448. (doi:10.1371/journal.pone.0028448)
- Brundin M, Figdor D, Sundqvist G, Sjogren U. 2013 DNA binding to hydroxyapatite: a potential mechanism for preservation of microbial DNA. *J. Endodont.* **39**, 211–216. (doi:10.1016/j.joen.2012.09.013)
- Schwarz C, Debruyne R, Kuch M, McNally E, Schwarcz H, Aubrey AD, Bada J, Poinar H. 2009 New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. *Nucleic Acids Res.* **37**, 3215–3229. (doi:10.1093/nar/gkp159)
- Bos KI *et al.* 2011 A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* **478**, 506–510. (doi:10.1038/nature10549)
- Schuenemann VJ *et al.* 2013 Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science* **341**, 179–183. (doi:10.1126/science.1238286)
- Bos KI *et al.* 2014 Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494–497. (doi:10.1038/nature13591)
- Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Paabo S. 2010 Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87. (doi:10.1093/nar/gkp1163)
- Gansauge MT, Meyer M. 2013 Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* **8**, 737–748. (doi:10.1038/nprot.2013.038)
- Dabney J *et al.* 2013 Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110**, 15 758–15 763. (doi:10.1073/pnas.1314445110)
- Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007 Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* **17**, 240–248. (doi:10.1101/gr.5681207)
- Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008 Rapid SNP discovery and genetic mapping using

- sequenced RAD markers. *PLoS ONE* **3**, e3376. (doi:10.1371/journal.pone.0003376)
33. Tin MM, Economo EP, Mikheyev AS. 2014 Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. *PLoS ONE* **9**, e96793. (doi:10.1371/journal.pone.0096793)
 34. Kircher M, Sawyer S, Meyer M. 2012 Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3. (doi:10.1093/nar/gkr771)
 35. Meyer M *et al.* 2012 A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226. (doi:10.1126/science.1224344)
 36. Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Paabo S. 2001 DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29**, 4793–4799. (doi:10.1093/nar/29.23.4793)
 37. Jiang H, Lei R, Ding SW, Zhu S. 2014 Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinform.* **15**, 182. (doi:10.1186/1471-2105-15-182)
 38. Magoc T, Salzberg SL. 2011 FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963. (doi:10.1093/bioinformatics/btr507)
 39. The Arabidopsis Genome Initiative. 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815. (doi:10.1038/35048692)
 40. Swarbreck D *et al.* 2008 The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36** (Database issue), D1009–D1014.
 41. The Potato Genome Sequencing Consortium. 2011 Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195. (doi:10.1038/nature10158)
 42. The Tomato Genome Consortium. 2012 The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641. (doi:10.1038/nature11119)
 43. Haas BJ *et al.* 2009 Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461**, 393–398. (doi:10.1038/nature08358)
 44. Li H. 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (<http://arxiv.org/abs/1303.3997>).
 45. Kircher M. 2012 Analysis of high-throughput ancient DNA sequencing data. *Methods Mol. Biol.* **840**, 197–228. (doi:10.1007/978-1-61779-516-9_23)
 46. Jonsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. 2013 mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684. (doi:10.1093/bioinformatics/btt193)