Nucleic Acids Research Advance Access published August 18, 2006

Nucleic Acids Research, 2006, Vol. 00, No. 00 ***1–8***
doi:10.1093/nar/gkl483

Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis

M. Thomas P. Gilbert[1,*], Jonas Binladen[1], Webb Miller[2], Carsten Wiuf[3,4], Eske Willerslev[1], Hendrik Poinar[5], John E. Carlson[6], James H. Leebens-Mack[7] and Stephan C. Schuster[2]

[1]Center for Ancient Genetics, Niels Bohr Institute and Biological Institutes, The University of Copenhagen, Juliane Maries vej 30, DK-2100 Copenhagen Ø, Denmark, [2]Center for Comparative Genomics and Bioinformatics, Penn State University, 310 Wartik Lab, University Park, PA 16802, USA, [3]Bioinformatics Research Center, University of Aarhus, Aarhus DK-8000, Denmark, [4]Molecular Diagnostic Laboratory, Aarhus University Hospital, DK-8200, Denmark, [5]McMaster Ancient DNA Center, Department of Anthropology and Pathology & Molecular Medicine, McMaster University, Hamilton, Ontario L82 4L9, Canada, [6]School of Forest Resources, Penn State University, 323 Forest Resources Building, University Park, PA16802, USA and [7]Department of Plant Biology, University of Georgia, Athens, Georgia 30602-7271

Received May 25, 2006; Revised June 21, 2006; Accepted June 23, 2006

ABSTRACT

Although ancient DNA (aDNA) miscoding lesions have been studied since the earliest days of the field, their nature remains a source of debate. A variety of conflicting hypotheses exist about which miscoding lesions constitute true aDNA damage as opposed to PCR polymerase amplification error. Furthermore, considerable disagreement and speculation exists on which specific damage events underlie observed miscoding lesions. The root of the problem is that it has previously been difficult to assemble sufficient data to test the hypotheses, and near-impossible to accurately determine the specific strand of origin of observed damage events. With the advent emulsion-based clonal amplification (emPCR) and the sequencing-by-synthesis technology this has changed. In this paper we demonstrate how data produced on the Roche GS20 genome sequencer can determine miscoding lesion strands of origin, and subsequently interpreted to enable characterization of the aDNA damage behind the observed phenotypes. Through comparative analyses on 390 965 bp modern chloroplast and 131 474 bp ancient woolly mammoth GS20 sequence data we conclusively demonstrate that Type 2 (C→T/G→A) miscoding lesions represent the overwhelming majority of damage derived miscoding lesions. In addition, we show that an as yet unidentified G→A modification, not the conventionally argued cytosine→uracil deamination, underpins the majority of Type 2 damage.

INTRODUCTION

The study of post mortem DNA damage is critically important to help ensure the generation of accurate data from ancient or degraded sources of DNA. DNA damage not only rapidly reduces the length and number of PCR amplifiable starting template molecules within a biological sample, but can also lead to the generation of erroneous sequence. The better characterization of aDNA damage will help the development of new damage strategies to both extend the range of samples from which useful DNA can be recovered, and help monitor and account for potentially erroneous data, which can have disastrous consequences on any study that requires the recovery of accurate sequence, e.g. phylogenetic and population genetic studies (1), genomic data analyses (2) and environmental reconstructions (3).

Miscoding lesions are a defining characteristic of ancient DNA (aDNA) studies. Although usually observed as variations on individual sequences among datasets of cloned PCR products (4), they are sometimes noticeable within directly sequenced PCR products, conferring the impression of sequence heteroplasmy (5). Two mechanisms have been suggested as the underlying cause behind the observed miscoding lesions. The first is the result of regular PCR polymerase amplification errors. In situations where starting PCR template molecules are low (as with many aDNA studies), such errors can result in the modification of template molecules during early stages of the PCR, and thus will produce a significant (i.e. observable through cloning or direct sequencing) proportion of descendents with the modification. The alternative, probably complementary, mechanism is the generation of errors owing to post mortem biochemical damage of the original starting template molecules. The chemical structure of these damage derived miscoding lesions is such that they can be read by the PCR enzyme, although

*To whom correspondence should be addressed. Tel: +45 35 32 05 87; Fax: +45 35 36 53 57; Email: mtpgilbert@gmail.com

erroneously (4,6,7). Probably the most commonly accepted example of the latter is the hydrolytic deamination of cytosine to uracil or its analogues, which during subsequent enzymatic replication leads to the generation of cytosine to thymine miscoding lesions (4).

The existence of damage-derived miscoding lesions in DNA from fossil remains has previously been proven using several methods. Through statistical comparisons, it has been demonstrated that aDNA sequences are characterized by relatively high occurrences of transitions compared with sequences from contemporary specimens. Therefore it has been argued that enzymatic error alone cannot explain the aDNA observations (7,8). The nature of these transitions themselves has however been a source of debate, both as to which types are truly associated with damage, and what the underlying cause of the miscoding lesions might be. Partly to blame for this lack of concordance is that the study of such damage is not trivial. Conclusions from most previous studies have been based on sequences generated from cloned PCR products. However, the nature of PCR causes any single original damage event to be viewed as two possible manifestations, dependent on whether a descendent of the damaged DNA strand, or of a complement sequence generated from the damaged strand, is sequenced. For example, consider an original cytosine to thymine transition (C→T) on a mitochondrial Light (L) strand molecule. Sequencing of descendent L strand molecules post PCR will lead to the observation of a C→T miscoding lesion. However, if the DNA is sequenced in the complementary Heavy (H) strand sequence, the transition will be read as a guanine to adenine (G→A) transition. Similarly, an original A→G damage event can be observed as either an A→G miscoding lesion, or the complementary T→C (7). This observation, describing the underlying difficulty of attributing a source to observed miscoding lesions owing to the lack of ability to identify the strand of origin of the event, has lead to the suggestion by Hansen and colleagues that miscoding lesions be grouped into the six complementary, effectively indistinguishable pairs (A→C/T→G), (A→G/T→C), (A→T/T→A), (C→A/G→T), (C→G/G→C) and (C→T/G→A) (7). Furthermore, the dominance of transitions in aDNA damage datasets has lead to the same authors to suggest that the two pairs of transitions be referred to as Type 1 (A→G/T→C) and Type 2 (C→T/G→A), respectively.

Although both types of transitions have been observed and commented on among aDNA datasets [e.g. (4,7,8–12)] controversy has raged as to whether both types truly represent damage [as argued by (7,8,10,12)], or whether Type 1 damage simply represents polymerase enzyme misincorporation errors at early stages of the PCR process (9,13). Furthermore, the debate does not stop there; the underlying causes of the damage are also under question. Though the few studies that attempt to examine miscoding lesions in detail have concurred that, as *in vivo*, Type 2 transitions arise from the deamination of cytosine to uracil (4,8,9), those studies that argue for the existence of Type 1 damage also argue that the deamination of adenine to hypoxanthine, an analogue of guanine, is also important to aDNA (8).

The limitation of such arguments is that they are not to a large extent based on observations of the actual raw data, but rather on theoretical arguments drawn from what is known about *in vivo* damage systems, thus about what damage may exist, and how the polymerase enzymes therefore may react to them. A small number of studies have attempted to investigate damage directly using various biochemical experiments, e.g. the treatment of aDNA extracts using uracil-N-glycosylase prior to PCR amplification, in order to investigate how the distribution of miscoding lesions varies as a result [e.g. (4,8,9,14,15)]. However such studies are subject to the limitation that they can only provide information about damage types that are specifically targeted, thus leaving the existence of other modifications unknown.

The recent development of the sequencing-by-synthesis technology (Genome sequencer GS20, Roche Applied Science) (16) offers a solution to these previously intractable problems. Specifically, the nature of the data-generation process is such that DNA sequence data can be assigned to individual, original single-stranded molecules. In brief, during the initial stage of the data preparation DNA molecules are first fragmented, then denatured and single-stranded molecules are emulsified with amplification reagents in a water-in-oil immersion, within which subsequent PCR occurs. During the subsequent emulsion PCR (emPCR), individual PCR occur in large-scale in parallel, and the descendant molecules of each individual reaction that are in the same orientation as the original single-stranded molecule are bound by the capture bead. During the final stages of the data generation process, the bound molecules on each individual capture bead are pyrosequenced as a single unit, in parallel with to up to 0.8 million other beads from the same emPCR reaction. Each contains PCR products from a different original, single-stranded DNA template molecule, and the data from each are recorded separately. The key benefit, therefore, is that each final sequence reaction is generated from a single single-stranded DNA molecule, and as such, provides a direct window into any damage-derived miscoding lesions that were present on the molecule, thus in an instant providing the critical information that has been lacking from previous aDNA damage studies.

In light of these benefits, we have analysed a dataset of DNA sequences produced using the GS20 DNA sequencing platform to further explore the nature of aDNA damage-driven miscoding lesions. First, through comparative analyses on 390 965 bp chloroplast DNA (cpDNA) generated from fresh (thus not containing damage forms that arise through conventional aDNA degradation processes) yellow-poplar (*Liriodendron tulipfera*) chloroplasts and 131 474 bp ancient woolly mammoth (*Mammuthus primigenius*) mitochondrial DNA (mtDNA), we show that a clear difference exists between the miscoding lesion spectra of modern and ancient DNA. Second, through statistical analysis of the data we conclusively demonstrate that Type 2 (C→T/G→A) miscoding lesions represent the overwhelming majority (88% total miscoding lesions, 94% of transitions) of damage derived miscoding lesions in aDNA from this specimen, in accordance with the hypothesis of Hofreiter *et al*. (9) and in contrast to others, including those postulated by some of the authors of this article (7,8,12). Third, using a simple logical argument based in principle on our observations on the GS20 data generation process, we demonstrate how the strand of origin of the sequences can be identified, and further how the underlying cause of observed damage types on the aDNA data can be identified, thus removing the need to

group miscoding lesions into complementary pairs. Through subsequent subdivision of the aDNA data into the different (L and H) strands of origin using this method, we demonstrate that the rate of occurrence and distribution of damage types is not significantly different between the two strands of the mitochondria. Finally, we explore the biochemical basis of the damage and demonstrate that it is an as yet unidentified derivative of guanine, leading to the generation of G→A miscoding lesions, and not the conventionally argued [e.g. (4,7–9,12)] deamination of cytosine to uracil and its analogues, that in this specimen at least (from the aDNA point of view a fairly standard, permafrost preserved bone) underpin the majority of aDNA damage-driven miscoding lesions.

MATERIALS AND METHODS

aDNA sequence data

As a consequence of the relatively rare occurrence of aDNA damage derived miscoding lesions, comparative studies require large quantities of DNA sequence data in order that statistically supported conclusions can be drawn. Furthermore, genetic regions analysed require multiple sequence coverage, so that true damage can be discriminated from other sources of sequence variation, such as allelic variation or the co-amplification of nuclear-mitochondrial sequences (numts). In short, this explains why to date few studies have been able to investigate damage derived miscoding lesion damage in detail [e.g. (7–9,12)]. With the advent of the GS20 sequencing platform has come the ability to rapidly generate large amounts of aDNA sequence data (with the caveat that samples contain enough quantity of DNA, of a minimum quality, to enable successful analysis). Furthermore, because of their relatively high copy number, and thus overall cellular abundance in comparison to nuDNA, the initial GS20 analyses on aDNA extracts have characteristically produced large amounts of mtDNA [(2), W. Miller, H. N. Poinar, J. Qi, C. Schwartz, L. P. Tomsho, R. D. E. MacPhee and S. C. Schuster, manuscript submitted], enabling the generation of the complete or near complete, ancient mtDNA genomes, with high levels of sequence coverage (Miller *et al*., manuscript submitted). For our analysis, we have used a dataset of woolly mammoth ancient mtDNA sequence, that comprises the sequences published in the first GS20 aDNA study (2), plus further mtDNA sequences from the same individual that have been generated since (Miller *et al*., manuscript submitted), representing a total of 131 474 bp of mammoth mtDNA sequence. Care was taken to avoid nuclear copies of mtDNA (numts), as follows. Analysis of numts in fully-sequenced mammalian genomes showed that at most 3% of the reads aligning to the mitochondrial genome (at our criteria) could be expected to be numts. Requiring that a read be 98% identical to mammoth mtDNA eliminated 15% of the aligning reads, most of which we believe to be low-quality data. Even if as much as 1% of the remaining reads are numts (so recent as to retain 98% identity), none of our broad conclusions would be materially affected. The large amount of sequence data, in contrast to the length of the mammoth mitochondria (16 770 bp for this individual, Miller *et al*., manuscript submitted) results in up to 21× coverage of some parts of the mtDNA genome, with a mean and modal coverage of 7.8 and 7 times respectively.

The individual sequence reads were aligned with the predetermined consensus sequence of the mtDNA genome using the Blastz program (17), and miscoding lesions were extracted from the alignment and assigned to the six complementary pairs of miscoding lesions of Hansen *et al*. (7). See Table 1 for summary data.

Analysis 1: statistical discrimination of damage from PCR enzyme misincorporation error

The GS20 emPCR process incorporates the use of the hifidelity polymerase, Platinum *Taq* Hifidelity (Invitrogen), an enzyme mixture composed of recombinant *Taq* DNA polymerase, *Pyrococcus* spp. GB-D thermostable polymerase, and Platinum *Taq* Antibody. This enzyme is marketed partly on its very low misincorporation rate, 2×10^{-6} (Invitrogen). In this study we find the actual rate of misincorporation to be higher ($\approx 7 \times 10^{-4}$), similar to results from a previous aDNA study that has also specifically examined these properties of this enzyme (8). To discriminate between true aDNA damage and enzyme error or potential damage that may have arisen during the DNA extraction or that may have been present in the DNA prior to extraction, we analysed a further dataset of GS20 sequences, generated from a modern DNA extract, comprising 390 965 bp of *Liriodendron tulipfera* cpDNA. These data are part of the first chloroplast genome sequenced using the GS20 (JEC, JHLM and Daniel G. Peterson, manuscript in preparation) and constitute all the sequence reads

Table 1. Number of miscoding lesions observed within chloroplast and mammoth datasets

	Miscoding lesions originally derived from A and T nucleotides				Miscoding lesions originally derived from C and G nucleotides			
	A→G T→C	A→C T→G	A→T T→A	Total A+T[a]	C→A G→T	C→G G→C	C→T G→A	Total C+G[b]
Chloroplast	78	24	89	244 230	33	9	52	146 735
Mammoth	39	7	9	81 790	16	8	597	49 684
Corrected Mammoth[c]	116	21	27		47	24	1763	
Nucleotide ratio				2.99				2.95

[a]Total number of adenine and thymine nucleotides in dataset.
[b]Total number of cytosine and guanine nucleotides in dataset.
[c]Corrected Mammoth: the number of observed lesions among the mammoth sequence data, scaled to match the total chloroplast nucleotides sequenced. For example, corrected mammoth count for A→G/T→C pair was calculated as [Observed Mammoth A→G/T→C]*[Total Chloroplast A+T]/[Total Mammoth A+T] = 39*244 320/81 790 = 116.

between np 45 000–90 000 of the genome (J. Carlson, J. Leebens-Mack and S. Schuster, unpublished data). The data analysed here have maximal coverage of 36 times, with a mean and modal coverage of 8.7 and 8 times, respectively.

As DNA from this sample was freshly extracted from modern tissue, miscoding lesions observed in the data are unlikely to be due to anything other than PCR or other sequencing error that arises during the GS20 data production process. The miscoding lesion spectra were extracted from the data using straightforward computer programs that we wrote for that purpose. For data summary see Table 1.

A χ^2 test of independence was used to investigate whether the distribution of miscoding lesions was the same in the mammoth and chloroplast sequence data. The data were first summarized into the six complementary damage pairs (Table 1). Subsequently, because nucleotide usage is different between the mammoth and chloroplast data, tests were performed separately on those miscoding lesions that originated from an A or T (A + T), and those that originated from a G or T (G + T).

Analysis 2: determination of which complementary miscoding lesion pairs represent true damage

To identify which of the six complementary pairs of miscoding lesions represent true damage in the mammoth mtDNA data as opposed to enzyme misincorporation errors, the data were modeled using the Poisson distribution with rates derived from the chloroplast data, i.e. taking the assumption that the observed miscoding lesion rates from the chloroplast data represent the true enzyme rates of lesions (Table 2). For example, the number of A→G/T→C miscoding lesions in the mammoth data was assumed to follow a Poisson distribution, with rate [Observed chloroplast A→G/T→C miscoding lesions]*[[Total Mammoth A+T nucleotides]/[Total Chloroplast A+T nucleotides]] = 78*81790/244230 = 26.12. Subsequently, the test-probability $P(X \geqslant \text{Observed, or } X \leqslant \text{Observed} - \text{Expected}) = P(X \geqslant 39 \text{ or } X \leqslant 39 - 26.12)$ was calculated, where X represents the number of lesions. If P is low, then it is likely that another mechanism than enzyme failure is accountable for the observed number of lesions in mammoth. The test-probability was made two-sided, because a priori we do not know the direction of deviation from the chloroplast data.

Under the assumption that the chloroplast data represent the true enzyme error, a basic rate of damage occurrence can be calculated for the six miscoding lesion types with the formula Max(Observed-Expected,0)/[Total source nucleotides].

Analysis 3: investigation for differences in strand damage accumulation rates

It has previously been speculated that the miscoding lesion damage rates on the different mtDNA strands (i.e. the L and H) may vary owing to base composition, secondary structure or other reasons (8,18). To examine for this phenomena, the miscoding lesion distributions were statistically compared using a χ^2 test on the datasets from the two different strands. The data are shown in Table 3, classified according to the type of the original base. A separate χ^2 test was performed for each type of nucleotide to account for the differences in nucleotide compositions between the two strands; because they are complementary and because the two strands are sampled randomly.

Analysis 4: the underlying causes of the complementary damage pairs

As mentioned above, the GS20 data production differs from conventional PCR and sequencing methods, in so far as each individual DNA sequence is derived from a single, single-stranded DNA molecule. Thus it is possible to actually identify the strand of origin of each generated DNA sequence. In this context this is whether the original template molecule for each DNA sequencing reaction was a Heavy (H) or Light (L) strand molecule. Further, once the strand of origin of the sequence is identified, it is possible using a simple logical argument to examine the miscoding lesions within each sequence, and reverse engineer both which original strand the damage occurred on, and what the original cause

Table 2. Number of observed and expected miscoding lesions in mammoth dataset

	A→G T→C	A→C T→G	A→T T→A	C→A G→T	C→G G→C	C→T G→A
Observed[a]	39	7	9	16	8	597
Expected[b]	26.12	8.04	29.81	11.17	3.05	17.61
P-value[c]	0.011	0.86	8×10^{-4}	0.17	0.013	$<1 \times 10^{-5}$
Occurrence per bp sequenced	1.5×10^{-4}	0	0	9.7×10^{-5}	9.9×10^{-5}	0.01

[a]Absolute number of miscoding lesions observed.
[b]Expected number of miscoding lesions, modeled using the Poisson distribution with rates derived from the chloroplast data.
[c]When using a 5% Bonferroni corrected significance level P-values < 5%/6 = 0.0084 are significant, leaving only (A→T/T→A) and (C→T/G→A) significant.

Table 3. Absolute number of damage events underlying observed miscoding lesions, subdivided by Light and Heavy template molecules

	T→N[a]				G→N				C→N				A→N			
	T	G	C	A	T	G	C	A	T	G	C	A	T	G	C	A
Heavy	22 613	1	9	0	2	15 808	1	290	54	2	8727	6	0	4	2	19 116
Light	18 816	0	15	3	0	8417	0	141	112	5	16 111	8	6	11	4	21 190
Total	41 429	1	24	3	2	24 225	1	431	166	7	24 838	14	6	15	6	40 306

[a]Where N refers to four possible derived nucleotide states, as listed in subsequent sub-columns.

of the miscoding lesion was. The arguments are presented in more detail (with explanatory figures) in the Supplementary Data. The DNA sequences were divided into two datasets, those derived from the L and those from the H strand. Each dataset was then aligned to the mtDNA genome and the frequency of each of the 12 possible miscoding lesions (damage per total nucleotide sequence) obtained. For data see Table 4.

To test whether there are any differences in the rates of change in each of the six complementary pairs of miscoding lesions (e.g. whether the rate of A→G occurrences differs from the rate of T→C occurrences within the A→G/T→C complementary pair) we used a χ^2 test of independence. The test relies on the assumption that there are no differences between the rates of occurrence on the two strands (see results of Analysis 3). The data from the two strands were therefore pooled, and classified according to the source type (A+T or G+C) of the original base. If there are no differences in the rates of damage, miscoding lesions would happen at the same rates from A as from T, and likewise at the same rates from C as from G. Therefore two χ^2 tests of independence were performed on the two datasets.

RESULTS

DNA sequence data

The DNA sequence data analysed in this study are available on NCBI Trace Archive. The Trace Identifiers are 153988 and 153989.

Analysis 1: statistical discrimination of damage from PCR enzyme misincorporation error

The χ^2 analysis of the modern cpDNA versus ancient mtDNA datasets (Table 1) provides strong statistical support for the notion that the miscoding lesion spectra are different (A→G/T→C, A→C/T→G and A→T/T→A: P-value $<1.3 \times 10^{-4}$), C→A/G→T, C→G/G→C and C→T/G→A: P-value $\leqslant 2.2 \times 10^{-16}$). As such, a significant part of the lesions within the aDNA dataset are derived from damage.

Analysis 2: determination of which complementary miscoding lesion pairs represent true damage

Once corrected for multiple comparisons, the χ^2 analysis of the observed versus expected mammoth aDNA miscoding lesion distribution (Table 2) provides statistical support that only two of the pairs cannot be attributed to enzymatic error. In particular, Type 2 transitions (C→T/G→A) are exceedingly over-represented, constituting 88% of the observed miscoding lesions. This provides strong support of previous arguments that they form the dominant form of aDNA damage derived miscoding lesions (4,7–9,12). However, in contrast to some of our previous observations, and in agreement with the arguments of Hofreiter *et al*. (9), Type 1 transitions (A→G/T→C), which here constitute <6% of the total miscoding lesions, and just over 6% of the total transitions observed, appear to play little or no role in aDNA damage derived miscoding lesions in this study. The overall Type 1:Type 2 ratio of ∼1:15 is considerably lower than that observed in all the previous studies [≈1:2 (8), ≈1:3, (12) and ≈1:6 in the data used in the study of Hofreiter *et al*. (9) (M. Hofreiter, personal communication)]. Further, we observe that A→T/T→A transversions are unusually under-represented in the mammoth aDNA data. As this clearly cannot be a result of damage, it seems likely that this observation is the result of the small number of A→T/T→A observations overall. In contrast, the much larger number of observations, and much stronger statistical support (much smaller P-value) suggest that this is not the case for the Type 2 transitions.

Analysis 3: investigation for differences in strand damage accumulation rates

A χ^2 analysis of the miscoding lesions within the two datasets representing the L and H strand sequences (Table 3) shows that when corrected for multiple tests (actual P-value required for 5% significance level of $0.05/4 = 0.0125$), there is no significant difference between the distributions (miscoding lesion from T: P-value $= 0.03$, from G: P-value $= 0.53$, from C: P-value $= 0.80$, from A: P-value $= 0.03$). Therefore there is no evidence to support previous hypotheses [e.g. (8)] that the damage distribution may vary significantly by strand.

Analysis 4: the underlying causes of the complementary damage pairs

The statistical analysis of the constituent damage types within the six complementary miscoding lesion pairs (Table 4) shows that although there is no evidence for a bias in contribution by the various damage events for A + T miscoding lesions ($P > 0.08$), there is significant support for a bias within C + G miscoding lesions ($P < 2.2 \times 10^{-16}$), arising due to the over-abundance of G→A over C→T transitions. With an occurrence per G nucleotide sequenced of 0.01779 in comparison to 0.00668 per C nucleotide (2.7 times more common), G→A modifications clearly represent the bulk of aDNA miscoding lesion damage. This finding is in stark contrast to all previously published hypotheses, that have concurred that it is cytosine to uracil deamination, resulting in C→T miscoding lesions that is the predominant, if not sole, cause of Type 2 transitions (4,7–9,12).

Table 4. Contribution of individual damage events to observed miscoding lesion pairs

Original damage[a]	*i j*	A→G T→C	A→C T→G	A→T T→A	C→A G→T	C→G G→C	C→T G→A
Mammoth dataset	*i*	15	6	6	14	7	166
	j	24	1	3	2	1	431
Complementary pair total	*i* + *j*	39	7	9	16	8	597
Percent of total mammoth observations		5.8	1.0	1.3	2.4	1.2	88.3

[a]Constituent damage events within each of the six complementary miscoding lesion pairs, identified as *i* and *j* respectively. Subsequent rows of the table describe observed number of *i* and *j* for each dataset, plus total (*i* + *j*).

As the C^{14} age of the mammoth sample is known [27 740 ± 220 years, (2)], the damage rate (r) can be calculated for the Type 2 transitions, both as the complementary pair and individually, using the following equation:

$$r = \frac{-\ln(1-x)}{t},$$

where t = time (e.g. in years or seconds) and x is the damage occurrence per base sequenced (either adjusted to take into account the assumed error misincorporation rate (rate of occurrence of Type 2 complementary pair), or unadjusted, thus reflecting the total miscoding lesion rate (individual C→T and G→A observations), which as it does not account for PCR enzyme error, represents a slight over estimate of the true rate. The damage rates are shown in Table 5. This damage rate is probably more accurate than rates that could be calculated using previous aDNA data, as they face the limitation of being unable to discriminate whether multiple miscoding lesions observed at a single position within cloned sequences from a single PCR product, are actually independent damage events, or simply descendents of a single damaged molecule [c.f. (8,11)]. Although no other data exist with which we can compare the rates calculated here, as other datasets from dated samples become available these rates can be compared to investigate whether there is any universal aspect to the rates of occurrences.

DISCUSSION

We have demonstrated using DNA sequence data generated using the GS20 emPCR and sequencing platform that, in this particular dataset, and in agreement with the arguments of Hofreiter *et al*. (9), Type 2 transitions are the overwhelmingly dominant cause of post mortem damage derived miscoding lesions. This is in stark contrast to other studies (published by several authors of this study), that report significantly higher levels (than the 6% reported here) of Type 1 miscoding lesions within aDNA datasets (7,8,12). Although it is tempting to explain this discrepancy as laboratory specific phenomena, it is worth noting that the conclusions of the aforementioned studies were based on both new data generated in the respective studies, plus data from a number of previous aDNA studies. Therefore, Type 1 transitions appear to be a true phenomenon of at least some aDNA sequence data, and we are therefore left in the difficult position of how to explain the discrepancy in the findings.

Caveat about conclusions drawn from the modern cpDNA data

The modern cpDNA dataset can be expected to contain some innate levels DNA damage, e.g. DNA that had not been repaired prior to extraction or DNA that was damaged during the extraction. Furthermore, it is not a unreasonable hypothesis that this damage spectra may differ from that found in modern mtDNA, owing to the differences between the structure of the genomes and the organelle biology. Thus, as the miscoding lesions observations on the cpDNA may represent the sum of the enzymatic error plus prior damage, they may represent an overestimate of the enzymatic error. However, the observed increase in the ratio of Type 1 to Type 2 transitions between the modern and aDNA datasets is so great (1:1 in the modern DNA versus 1:15 in the aDNA) that any cpDNA damage is unlikely to significantly affect the conclusions of this study.

Table 5. Type 2 damage rate, nucleotides per unit time

	Per year	Per second
Type 2 damage[a]	4.2×10^{-7}	1.3×10^{-14}
G→A[b]	6.2×10^{-7}	2.0×10^{-14}
C→T[b]	2.3×10^{-7}	7.4×10^{-15}

[a]Adjusted to account for enzyme contribution to miscoding lesions.
[b]Unadjusted for enzyme contribution, therefore overestimate of true rate.

Explanations for the lack of Type 1 transitions

One potential explanation is that the differences in DNA extraction methodologies may play a significant role in the observed results. For example, the dataset of Hofreiter *et al*. (9) was generated from DNA extracted and purified using a silica-based methodology [modified from Boom *et al*. (19)], and thus the nucleic acids were exposed to both acidic conditions and high concentrations of guanidinium chaotropes. In contrast, however, the majority of data from the conflicting studies were generated using a buffered digestion mix at neutral pH, followed by organic purification of the nucleic acids [in general modified from Sambrook *et al*. (20)]. Thus it might be argued that the conditions of the silica method might somehow result in the fragmentation of DNA at positions where the underlying cause of Type 1 transitions have occurred. Unfortunately however, a major problem with this explanation is that the DNA analysed in this study was initially purified using the non-silica method, thus rendering this explanation unlikely.

An alternative explanation that has been proposed previously is that Type 1 transitions derive from innate enzyme error at early stages of the PCR process (9,13), giving rise to what have been described as 'singleton' miscoding lesions, in contrast to 'consistent' miscoding lesions among cloned data (9). In light of our data, this explanation is equally problematic, as unlike previously generated data, our sequences all stem from single, single-stranded template molecules. This places us in an optimal position to observe the true enzymatic misincorporation behaviour, and as the enzyme used in the emPCR (Invitrogen's Platinum *Taq* Hifidelity) is that used in many of the previously studied aDNA datasets, it is difficult to support the argument.

A third explanation is that Type 1 transitions, if they had existed in the original data, may have been removed through strand fragmentation at the site of Type 1 transitions during the multiple preparation steps that DNA is required to go through during the GS20 process. Specifically the DNA is first fragmented through physical shearing. Subsequently the DNA must be polished through blunt ending and phosphorylation using T4 DNA polymerase (exhibits 3′–5′ exonuclease activity), *E.coli* DNA polymerase (Klenow fragment) (fills in recessed 3′ ends, and T4 polynucleotidekinase (phosphorylating 3′ ends). Neither of these treatments offer a good reason for why miscoding lesion damage in the middle of a template molecule may be removed. The next stage however may provide a key as the enzyme treated

DNA is subsequently re-purified using silica spin columns (e.g. Qiagen's QIAquick columns). As this involves the use of further guanidinium containing buffers it could be that Type1 damage is removed at this stage.

The fourth explanation is that the damage observed in this sample is simply the true damage spectra, but owing to as yet undetermined factors the spectra varies by individual ancient sample. This specimen is unique to some extents as it was recovered in a frozen state directly from frozen Siberian permafrost, and subsequently retained at sub-zero (predominantly −15°C) conditions prior to DNA extraction (2). The damage spectra therefore may not directly reflect on the DNA damage within all other specimens, perhaps through an unusually limited access of the DNA to free water molecules.

The dual causes of Type2 transitions

The most intriguing finding of this study is that our data demonstrate that, in contrast to all previous statements on the subject (4,8,9,12,13), the predominant cause of Type 2 transitions is not cytosine to uracil deamination, but the degradation of guanine to a derivative that is misread by the PCR enzyme as an adenine. However, this is not to say that cytosine to uracil deamination does not exist—the resultant manifestation of C→T transitions are clearly observed here at a highly significant rate (in comparison to the enzyme misincorporation rates), and cytosine to uracil deamination has been experimentally identified through previous UNG treatment assays of purified aDNA (4,8,9). It is worth noting here that in our experience UNG treatment of aDNA extracts sometimes leaves some remaining C→T transitions in the resultant PCR amplified and cloned sequences (M. Thomas P. Gilbert, unpublished data). Whereas we previously thought these to simply result from incomplete enzymatic cleavage of all the damaged sequences a more likely explanation is that the remaining C→T lesions were in fact derived from guanine derivatives. Naturally we caution again that owing to the recent advent of the GS20 technology, limited aDNA data are publicly available for study, thus our conclusions are based on data from a single sample. Therefore until further studies are undertaken on additional samples, conclusions as to how widespread this phenomenon is cannot be drawn. That guanine degradation appears to be the dominant cause of Type 2 transitions is interesting, in so far as a previous study has also remarked on the dominance of other guanine modifications among oxidative forms of aDNA damage. Specifically, using Gas Chromatography/Mass Spectrometry (GC/MS), to identify PCR-blocking hydantoins, Höss *et al*. (14) report that guanine modifications dominate in the majority (3/5) of samples from which they can successfully PCR amplify DNA.

Clearly the major outstanding question arising from this study is what exactly the damage to guanine is, which can give rise to G→A miscoding lesions. Although various modifications of guanine are known to cause miscoding lesions, during enzymatic replication these result in the generation of transversions; e.g. a common product of oxidative degradation, 8-oxoguanine, generates G→T transversions (21), while other guanine products such as 8-methyl-2′-deoxyguanosine generate both G→C and G→T transversions (22). We are unaware of any explanation as to what might cause G→A transitions, although the damaged form would of course have to result in the misincorporation of a thymine opposite. However, at this point we note that a problem with previous studies on damage is that they have attempted to draw explanations from what is known about damage in *in vivo* systems in order to explain post mortem observations. There is no good reason why the two systems need be identical, indeed it might be expected that key differences exist between metabolically active, and the deceased environments.

Reevaluation of previous conclusions in light of current findings

Under the assumption that Type 1 transitions do not represent true aDNA damage, and that Type 2 transitions may originate from both C→T and G→A events, several past statements with regards to aDNA need to be evaluated as follows.

Damage hotspots

The existence of particular for post mortem DNA damage hotspots (specific nucleotide positions that appear to undergo damage at a rate significantly above that expected under the hypothesis that damage is equally likely to affect all positions) has been argued based on observations of the distribution patterns of miscoding lesions (predominantly Type 1 and 2 transitions) (8,11). Although the possibility has been raised that the hotspot observations originally made on human DNA sequences may be flawed due to contamination of the samples (13), this seems unlikely as a second study on bison produced similar findings (11). However, if Type 1 damage is not a true phenomena, but simply represents PCR enzyme misincorporation, then while the underlying observation of miscoding lesions observed at specific non-random nucleotide positions does not change, the argument that damage may preferentially occur at these positions does. Data used in such studies warrant reanalysis, to remove Type 1 transitions, then statistical tests require recalculation in order to investigate whether support still exists for damage hotspots. A recent study has reported the existence of DNA sequencing error hotspots (23), thus these may also play some effect in the original damage hotspot observations. However, we stress that one of the original conclusions of the damage hotspot papers, that aDNA sequence authenticity may be challenged by predominance of miscoding lesions at specific, phylogenetically informative nucleotide positions, remains unchanged whether the cause of hotspots is damage or position specific sequencing errors.

Jumping PCR

Based on the hypothesis that single damage events can explain the origin of Type 1 and Type 2 transitions, respectively, (cytosine to uracil and adenine to hypoxanthine deamination, respectively), Gilbert *et al*. (8) have argued that this provides a tool for identifying recombinant aDNA sequences that may have arisen through jumping PCR (24). For example, if, as previously hypothesized, Type 2 transitions could only arise through cytosine to uracil deaminations, then in an absence of jumping PCR the resultant two damage phenotypes (C→T and G→A) transitions should never be observed within the same individual cloned DNA sequence (as the C→T observation must have arisen from an original

cytosine deamination on a template molecule in the same orientation as the final read sequence, while the G→A observation must have arisen on a different template molecule of the complementary orientation). In light of the findings that Type 1 transitions may not represent true damage, and that Type 2 transitions may originate owing from both cytosine and guanine degradation, the theory behind this argument does not hold. Therefore we advise that jumping PCR analyses cannot be performed in the described manner.

CONCLUSIONS

The advent of the GS20 and other high-throughput DNA sequencing techniques will rapidly increase the data available for aDNA damage analyses. As these data become available, new analyses should be able to investigate how general the conclusions of this study are. In combination with improved methods for the efficient recovery of aDNA, and newly developed biochemical assays that have started to overturn conventional damage dogma [e.g. the dominance of DNA cross-linking in some aDNA sources (15)], our understanding about the extent and biochemical basis behind aDNA damage should rapidly increase, enabling future expansion on what samples are available for aDNA analyses, and what can be done with the recovered DNA.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

ACKNOWLEDGEMENTS

The study was conceived by M.T.P.G. and S.C.S. S.C.S. and W.M. generated and assembled the data. M.T.P.G., J.B., W.M., S.C.S., C.W. and E.W. developed and performed the data analysis and wrote the manuscript. J.E.C., J.H.L.M. and H.P. provided the DNA extractions. M.T.P.G. was supported by the Marie Curie FP6 Actions 'FORMAPLEX' grant. J.B. and E.W. were supported by the Wellcome Trust, UK, the Carlsberg Foundation, DK, and the National Science Foundation, DK. W.M. was supported by HIN grant HG002238. This project is funded, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds appropriated by the legislature. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions. Funding to pay the Open Access publication charges for this article was provided by Marie Curies FP6 Actions 'FORMAPLEX' grant.

Conflict of interest statement. None declared.

REFERENCES

1. Shapiro,B., Drummond,A.J., Rambaut,A., Wilson,M., Sher,A., Pybus,O.G., Gilbert,M.T.P., Barnes,I., Binladen,J., Willerslev,E. *et al*. (2004) Rise and fall of the Beringian steppe bison. *Science*, **306**, 1561–1565.
2. 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.
3. Willerslev,E., Hansen,A.J., Binladen,J., Brandt,T.B., Gilbert,M.T.P., Shapiro,B., Bunce,M., Wiuf,C., Gilichinsky,D.A. and Cooper,A. (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, **300**, 791–795.
4. Pääbo,S. (1989) Ancient DNA: extraction, characterization, molecular cloning and enzymatic amplification. *Proc. Natl Acad. Sci. USA*, **86**, 1939–1943.
5. Anderung,C., Bouwman,A., Persson,P., Carretero,J.M., Ortega,A.I., Elburg,R., Smith,C., Arsuaga,J.L., Ellegren,H. and Götherström,A. (2005) Prehistoric contacts over the Straits of Gibraltar indicated by genetic analysis of Iberian Bronze Age cattle. *Proc. Natl Acad. Sci. USA*, **102**, 8431–8435.
6. Lindahl,T. (1993) Instability and decay of the primary structure of DNA. *Nature*, **362**, 709–715.
7. Hansen,A., Willerslev,E., Wiuf,C., Mourier,T. and Arctander,P. (2001) Statistical evidence for miscoding lesions in ancient DNA templates. *Mol. Biol. Evol*., **18**, 262–265.
8. Gilbert,M.T.P., Hansen,A.J., Willerslev,E., Rudbeck,L., Barnes,I., Lynnerup,N. and Cooper,A. (2003) Characterisation of genetic miscoding lesions caused by post mortem damage. *Am. J. Hum. Genet*., **72**, 48–61.
9. Hofreiter,M., Jaenicke,V., Serre,D., von Haeseler,A. and Pääbo,S. (2001) DNA sequences from multiple amplifications reveal artefacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res*., **29**, 4693–4799.
10. Gilbert,M.T.P., Willerslev,E., Hansen,A.J., Rudbeck,L., Barnes,I., Lynnerup,N. and Cooper,A. (2003) Distribution patterns of post mortem damage in human mitochondrial DNA. *Am. J. Hum. Genet*., **72**, 32–47.
11. Gilbert,M.T.P., Shapiro,B., Drummond,A. and Cooper,A. (2005) Post mortem DNA damage hotspots in Bison (*Bison bison* and *B.bonasus*) provide supporting evidence for mutational hotspots in human mitochondria. *J. Archaeol. Sci*., **32**, 1053–1060.
12. Binladen,J., Wiuf,C., Gilbert,M.T.P., Bunce,M., Larson,G., Barnett,R., Hansen,A.J. and Willerslev,E. (2006) Comparing miscoding lesion damage in mitochondrial and nuclear ancient DNA. *Genetics*, **172**, 733–741.
13. Pääbo,S., Poinar,H., Serre,D., Jaenicke-Despres,V., Hebler,J., Rohland,N., Kuch,M., Krause,J., Vigilant,L. and Hofreiter,M. (2004) Genetic analyses from ancient DNA. *Ann. Rev. Genet*., **38**, 645–679.
14. Höss,M., Jaruga,P., Zastawny,T., Dizdaroglu,M. and Pääbo,S. (1996) DNA damage and DNA sequence retrieval from ancient tissue. *Nucleic Acids Res*., **24**, 1304–1307.
15. Hansen,A.J., Mitchell,D.L., Wiuf,C., Paniker,L., Brand,T.B., Binladen,J., Gilichinsky,D.A., Ronn,R. and Willerslev,E. (2006) Crosslinks rather than strand breaks determine access to ancient DNA sequences from frozen sediments. *Genetics*, **173**, 1175–1179.
16. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z.T. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
17. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with Blastz. *Genome Res*., **13**, 103–107.
18. Heyer,E., Zietkiewicz,E., Rochowski,A., Yotova,V., Puymirat,J. and Labuda,D. (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: Study of control region mutations in deep-rooting pedigrees. *Am. J. Hum. Genet*., **69**, 1113–1126.
19. Boom,R., Sol,C.J., Salimans,M.M., Jansen,C.L., Wertheim-van Dillen,P.M. and van der Noordaa,J. (1990) Rapid and simple method for purification of nucleic acids. *J. Clin. Microbiol*., **28**, 495–503.
20. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*, *2nd edn*. Cold Spring Harbor Press, Cold Spring Harbor, NY.
21. Maki,H. and Sekiguchi,M. (1992) MutT protein specifically hydrolyses a potent mutagenic substrate for DNA synthesis. *Nature*, **355**, 273–275.
22. Kohda,K., Tsunomoto,H., Kasamatsu,T., Sawamura,F., Tershima,I. and Shibutani,S. (1997) Synthesis and miscoding specificity of oligodeoxynucleotide containing 8-phenyl-2′-deoxyguanosine. *Chem. Res. Toxicol*., **10**, 1351–1358.
23. Brandstätter,A., Sanger,T., Lutz-Bonengel,S., Parson,W., Beraud-Colomb,E., Wen,B., Kong,Q.P., Bravi,C.M. and Bandelt,H.-J. (2005) Phantom mutation hotspots in human mitochondrial DNA. *Electrophoresis*, **18**, 3414–3429.
24. Pääbo,S., Irwin,D. and Wilson,A. (1990) DNA damage promotes jumping between templates during enzymatic amplification. *J. Biol. Chem*., **265**, 4718–4721.