Letter to the Editor

Statistical Evidence for Miscoding Lesions in Ancient DNA Templates

Anders J. Hansen, Eske Willerslev,* Carsten Wiuf,† Tobias Mourier,* and Peter Arctander**

*Department of Evolutionary Biology, Zoological Institute, University of Copenhagen Denmark, Copenhagen, Denmark; and †Department of Statistics, Oxford University, Oxford, England

It is generally believed that sequence heterogeneity in PCR products from fossil remains are due to regular DNA polymerase errors as well as miscoding lesions compounded by damage in the template DNA (Pääbo 1990; Handt et al. 1994*b*, 1996; Höss et al. 1996; Krings et al. 1997). However, it has been difficult to test the frequency with which this assumption holds. First, DNA extractions from fossil remains rarely produce a yield large enough for pre-PCR analysis of postmortem modifications (Höss et al. 1996). Second, in most cases, it is not possible to determine whether nucleotide misincorporations by the DNA polymerase enzyme during amplification are caused by regular DNA polymerase errors or miscoding lesions in the template DNA sequences (Greenwood et al. 1999). Finally, the error rates of the DNA polymerase enzymes for PCR have proved to be highly unpredictable, making it difficult to account for regular DNA polymerase errors in amplified DNA sequences (Eckert and Kunkel 1991).

Here, we present a statistical model for analyzing PCR-mediated base-misincorporations, catalyzed by the commonly used *Thermus aquaticus* (*Taq*) polymerase enzyme, in amplification products from fossil remains.

The error rate of the *Taq* polymerase enzyme may vary more than 10-fold ($\sim 2 \times 10^{-4}$ to $< 1 \times 10^{-5}$ per nucleotide per cycle) according to the precise DNA sequence and the in vitro conditions of DNA synthesis (Eckert and Kunkel 1991). Therefore, the tests of the model rely solely on the relative distribution of the distinct *Taq* polymerase errors, which, in contrast to the highly variable error rate, is nearly constant and independent of the starting template material and the conditions for the PCR, as shown in table 1. Hence, the tests are not affected by variations in PCR efficiencies and accuracy. The model compares the distribution of the regular *Taq* polymerase errors with the observed substitutions in amplification products from fossil remains under the hypothesis that any significant differences between the distributions are due to miscoding lesions in the template DNA sequences used for PCR. The model was applied to published multiple clone sequences of the mitochondrial (mt) DNA control region from three differently preserved specimens of Homo representing different ages: a ~600-year-old Hokokam Indian (VC15A) found in a cave in Arizona, southwestern United States (Handt et al. 1996), the ~5,000-year-old ice man recovered from a glacier in the Tyrolean Alps (Handt et al. 1994*b*), and the >30,000-year-old Neanderthal-type specimen found in a limestone quarry near Düsseldorf, Germany (Krings et al. 1997).

Contamination by contemporary DNA poses a serious threat to studies of ancient DNA, especially from human remains (Pääbo, Higuchi, and Wilson 1989). Therefore, the data sets applied to the statistical analysis were carefully chosen from the literature to ensure that all recommended criteria and controls were fulfilled, in order to verify the authenticity of the sequence material (Lindahl 1993*a*; Handt et al. 1994*a*; Austin et al. 1997). The estimated ages of all three specimens fall within the theoretical limit of 50,000–100,000 years for amplifiable ancient DNA sequences (Pääbo and Wilson 1991; Lindahl 1997, 2000). All DNA extractions and PCR setups were physically separated from running, cloning, and sequencing through the use of fully equipped pre-PCR laboratories solely dedicated to ancient DNA work. Appropriate controls were used to detect possible contamination. For each of the specimens, unambiguous and reproducible results were obtained from independent DNA extracts by different laboratories. Finally, the sequences were congruent with what can reasonably be expected from known mitochondrial sequence variation in present populations of Homo and Pan.

Clone sequences whose ancient origins were considered uncertain by the authors were omitted from the analysis. Furthermore, the sequence materials used in the model were all obtained using different primer pairs that enabled partially overlapping sequences to be amplified in order to prevent amplification of nuclear insertions (Handt et al. 1994*b*, 1996; Krings et al. 1997). Therefore, contaminant DNA, as well as nuclear insertions, were unlikely to be present in the clone sequences used for analysis.

For each of the specimens, a sequence was constructed that contained all of the observed substitutions in the multiple-clone data set. This sequence was then compared with the proposed consensus sequence of the specimen, and the number of substitutions was calculated (table 2). Identical substitutions in a given position present in more than one clone sequence were treated as single events. Ambiguous residues (0.3% in the Tyrolean ice man, 1.0% in the Neanderthal), indels (0.3% in the Hokokam Indian, 1.1% in the Tyrolean ice man, 1.1% in the Neanderthal), and positions with two or more nonidentical substitutions (0.5% in the Neanderthal) were omitted from the analysis. All columns in the alignment of the consensus sequence and the sequence incorporating substitutions were considered as independent observations arising from a common distribution. As a consequence, if p is the probability of a pre-PCR derived substitution and q is the probability of a regular

Key words: ancient DNA, miscoding lesions, *Taq* polymerase errors.

Address for correspondence and reprints: Eske Willerslev, Department of Evolutionary Biology, Zoological Institute, University of Copenhagen Denmark, Universitetsparken 15, DK-2100, Copenhagen Ø, Denmark. E-mail: ewillerslev@zi.ku.dk.

Mol. Biol. Evol. 18(2):262–265. 2001

Table 1
***Taq* Polymerase Errors Obtained from the Literature**

		Study						
		I	II	III	IV	V	Total	d_{TSj}[a]
AT→GC. . .	TS1	10	11	19	17	19	76	0.84
AT→TA. . .	TV1	2	3	3	0	3	13	
AT→CG. . .		0	1	0	0	1		
CG→TA. . .	TS2	3	4	1	3	4	15	0.16
CG→GC . .	TV2	0	0	1	2	3	10	
CG→AT. . .		2	0	0	0	2		
Total		17	19	24	22	32	114	

Note.—Studies I (Saiki et al. 1988), II (Dunning, Talmud, and Humphries 1988), III, and IV (Eckert and Kunkel 1990) are based on PCR and vary considerably with regard to number of cycles (20–35) and concentrations of dNTP (0.25–1.32 mM), $MgCl_2$ (1–10.0 mM), and template material (reaction temperature 70°C). Study V (Tindall and Kunkel 1988) is based on a forward mutational assay: one round of DNA synthesis, 0.04 mM dNTP, 10.0 mM $MgCl_2$, and a 55–70°C reaction temperature. TS = transition(s); TV = transversion(s).

[a] $d_{TS,j}$ is the ratio of j given TS, e.g., d_{TS1} = no. of TS1/(no. of TS1 + no. of TS2) = 76/(76 + 15) = 0.84 is the probability that a *Taq* polymerase error is of type TS1, given that a TS error occurs.

Table 2
Data and Notation

		Sequence with Substitutions		
		Identity	Transition	Transversion
Consensus sequence	A, T	ID1	TS1	TV1
	C, G	ID2	TS2	TV2
Neanderthal bone	A, T	183	16	4
	C, G	140	25	2
Tyrolean ice man	A, T	166	12	7
	C, G	149	13	2
Hokokam Indian	A, T	187	2	4
	C, G	156	3	1

Note.—A double index, ij, is used to describe the categorization of data and the model: i = ID (identities), TS (transitions), or TV (transversions) refers to the state of the sequence incorporating substitutions (see main text), and j = 1 or 2 refers to the state of the consensus sequence. The observed counts of the different types are denoted n_{ij}; i = ID, TS, or TV, and j = 1 or 2. If the consensus sequence has either A or T in a position, j = 1; otherwise, j = 2. TS1 covers the transitions A→G and T→C, grouped together in AT→GC, as these substitutions are indistinguishable due to the complementarity of the sequence material; likewise, TS2 covers the CG→TA transitions. TV1 covers the AT→CG and AT→TA transversions which, for convenience, were pooled into AT→(CG)(TA), and, similarly, TV2 covers the CG→(AT)(GC) transversions.

Taq polymerase error, the additive probability of a substitution is $p + q$. The index notation used to describe the data and the model is explained in table 2.

Using chi-square statistics, the clone data sets were tested under the following hypothesis (H_1): Can all of the observed substitutions be ascribed to regular *Taq* polymerase errors? The test of H_1 is shown in table 3.

We find the distribution of substitutions in the amplification products for all three specimens to be significantly different ($P < 0.05$) from the distribution expected solely from regular *Taq* polymerase errors (table 3). Therefore, regular *Taq* errors cannot account for all of the heterogeneity observed in the multiple-clone sequences. Mitochondrial heteroplasmy can possibly account for some of the observed substitutions in the clone sequences. However, single-site heteroplasmy in the human mitochondrial control region has been encountered at no more than one or two sites in only 1%–3% of all individuals investigated (Gocke, Benko, and Rogan 1998). Therefore, possible sequence variation caused by heteroplasmy is of insignificant importance to this investigation.

When contamination, nucleic insertions, and mitochondrial heteroplasmy are excluded as significant contributors to the observed sequence heterogeneity, we find the only plausible reason for the discrepancy between the expected and observed distribution of base-misincorporations in the clone sequences to be miscoding lesions in the template DNA sequences. As all three specimens differ in age and preservation conditions, the result suggests that miscoding lesions are common in DNA from fossil remains, across the ages of specimens and their preservation conditions.

To investigate for significant differences in the distribution of pre-PCR derived transitions, the clone data sets were tested under the following hypothesis (H_2): Do AT→GC (TS1) and GC→AT (TS2) substitutions occur at the same rate? The test of H_2 is shown in table 3.

We found that only the clone sequences from the Neanderthal specimen contain significantly larger amounts of CG→TA changes than TA→CG changes ($P < 0.05$) (table 3). As this is the oldest of the specimens, the results suggest that distinct miscoding lesions occur at different rates, producing a displacement between transitions with time. This is in agreement with the observation that hydrolytic deamination of cytosine and its homolog 5-methyl cytosine to uracil and thymine, generating CG→TA transitions during replication, are among the major types of miscoding lesions in the genome of living human cells. These transitions are believed to occur at a rate about 30–50 times that of hydrolytic deamination of adenine to hypoxanthine, generating TA→CG transitions during replication (Lindahl 1993*b*).

The inclusion of the distribution of *Taq* polymerase errors in the statistical model causes a problem of overparameterization, which limits the opportunities for statistical analysis (table 3). Using high-fidelity polymerases such as the *Pfu* with an error rate of 2.0×10^{-6} to 6.5×10^{-7} per nucleotide per cycle (Flaman et al. 1994; André et al. 1997) would permit regular DNA polymerase errors to be completely ignored in the statistical model. This would allow for comparisons of factors such as the amounts of transitions and transversions within a clone data set and transition/transversion ratios among different data sets. Therefore, future amplification of DNA from fossil remains should be carried out using high-fidelity DNA polymerases, as has recently been proved possible (Willerslev et al. 1999).

In summary, the results provide statistical evidence for the assumption that heterogeneity observed in PCR products from fossil remains in general are due to regular DNA polymerase errors as well as miscoding lesions in the template DNA sequences (Pääbo 1990; Handt et al. 1994*b*, 1996; Krings et al. 1997). Furthermore, the results suggest that miscoding lesions in DNA sequences from fossil remains can occur with different rates generating a displacement of transitions with time.

Table 3
Transitions and Transversions

		Observed[a]		Expected[b]		χ^2 Contribution			
		$j = 1$	$j = 2$	$j = 1$	$j = 2$	$j = 1$	$j = 2$	χ^2	P
Test of H_1									
Taq polymerase	TS	76	15	70.93	19.79	0.36	1.16	7.4	
	TV	13	10	18.07	5.21	1.42	4.41		
Neanderthal bone	TS	16	25	15.94	21.38	0.00	0.61	4.9	
	TV	4	2	4.06	5.62	0.00	2.34		
Tyrolean ice man	TS	12	13	15.14	11.88	0.65	0.11	3.0	
	TV	7	2	3.86	3.12	2.56	0.41		
Hokokam Indian.	TS	2	4	3.98	3.96	0.99	0.00	3.7	
	TV	3	1	1.02	1.04	3.88	0.00		
								19.0	0.02
Test of H_2									
Neanderthal bone	TS	16	25	22.5	18.5	1.9	2.3	4.7	0.03
	R	187	142	180.5	148.5	0.2	0.3		
Tyrolean ice man	TS	12	13	13.2	11.8	0.1	0.1	0.3	0.6
	R	173	151	171.8	152.2	0.01	0.01		
Hokokam Indian.	TS	2	4	2.7	2.3	0.2	0.2	0.4	0.6
	R	191	157	190.3	157.7	0.00	0.00		

Note.—For H_1, the hypothesis that all substitutions could be ascribed to regular *Taq* polymerase errors; in the full model, probabilities are given by r_{ij}, where i = transitions (TS) or transversions (TV), and $j = 1$ or 2, and each row sums to one: $r_{TSj} + r_{TVj} = 1$. The r_{ij}'s are possibly distinct in the four data sets (obtained from tables 1 and 2). Hypothesis H_1 states that the r_{ij}'s are identical in the four sets. Note that the full model has eight free parameters (two for each set), whereas H_1 has two free parameters. For H_2, the hypothesis that AT→GC (TS1) and GC→AT(TS2) substitutions occur at the same rate; in the full model, probabilities are given by $r_{TSj} = p_{TSj} + d_{TSj} \cdot q$, $r_{Rj} = 1 - r_{TSj}$, where $j = 1$ or 2. This is a consequence of the additivity of substitution probabilities. p_{TSj} is the probability of a pre-PCR derived transition given that the consensus sequence is j, and q is the probability of a transition caused by a regular *Taq* polymerase error. The constant d_{TSj}, $j = 1, 2$, is the probability of an error of type TSj given that a TS error occurs, that is, d_{TSj} = no. of TSj/(no. of TS1 + no. of TS2) and is assumed to be known (see table 1). Thus, $d_{TSj} \cdot q$ is the probability of a transition caused by a regular *Taq* polymerase error if the consensus sequence is in state j. Hypothesis H_2 states that $p_{TS1} = p_{TS2}$. The full model is overparameterized (three parameters), and H_2 has two free parameters, i.e., the maximal number of parameters that can possibly be estimated uniquely. For both hypotheses, the chi-square statistics were applied: $\chi^2 = \text{sum}_{ij}(E_{ij} - N_{ij})^2/E_{ij}$, where N_{ij} is the count of type ij and E_{ij} is the expected count under the null hypothesis. The significance of the observed value x of χ^2 was evaluated through simulations under the null hypothesis. If $p = P(\chi^2 \geq x) \geq 5\%$, the null hypothesis was accepted; otherwise, it was rejected.

[a] For the test of H_1, the counts of transitions and transversions are shown. For the test of H_2, the counts of transversions and identities lumped (R) and the counts of transitions are shown. The lumped types are called R_j, $j = 1, 2$, and data are summarized by n_{TS1}, n_{TS2}, n_{R1}, and n_{R2}, with $n_{Rj} = n_{IDj} + n_{TVj}$.

[b] Expected counts under the hypotheses H_1 and H_2.

Acknowledgments

We are grateful to M.-A. Coutellec-Vreto, S. Mathiasen, J. Pritchard, and S. Sumner for critical reading of the manuscript. A.J.H. and E.W. were supported by the VELUX Foundation of 1981, Denmark, and C.W. was supported by grant BBSRC 43/MMI09788 and the Carlsberg Foundation, Denmark. A.J.H. and E.W. contributed equally to this work and should be regarded as joint first authors.

LITERATURE CITED

André, P., A. Kim, K. Khrapko, and W. Thilly. 1997. Fidelity and mutational spectrum of *Pfu* DNA polymerase on a human mitochondrial DNA sequence. Genome Res. **7**: 843–852.

Austin, J. J., A. B. Smith, and R. H. Thomas. 1997. Palaeontology in a molecular world: the search for authentic ancient DNA. TREE **12**:303–306.

Dunning, A. M., P. Talmud, and S. E. Humphries. 1988. Errors in the polymerase chain reaction. Nucleic Acids Res. **16**:10393.

Eckert, K. A., and T. A. Kunkel. 1990. The fidelity of DNA polymerase used in the polymerase chain reaction. Pp. 225–244 *in* M. J. McPherson, P. Quirke, and G. R. Taylor, eds. PCR: a practical approach. IRL Press, Oxford University Press, Oxford, England.

———. 1991. DNA polymerase fidelity and the polymerase chain reaction. PCR Methods Appl. **1**:17–24.

Flaman, J.-M., T. Frebourg, V. Moreau, F. Charbonnier, C. Martin, C. Ishioka, S. H. Friend, and R. Iggo. 1994. A rapid PCR fidelity assay. Nucleic Acids Res. **22**:3259–3260.

Gocke, C. D., F. A. Benko, and P. K. Rogan. 1998. Transmission of mitochondrial DNA heteroplasmy in normal pedigrees. Hum. Genet. **102**:182–186.

Greenwood, A. D., C. Capelli, G. Possnert, and S. Pääbo. 1999. Nuclear DNA sequences from late Pleistocene megafauna. Mol. Biol. Evol. **16**:1466–1473.

Handt, O., M. Höss, M. Krings, and S. Pääbo. 1994*a*. Ancient DNA: methodological challenges. Experientia **50**: 524–529.

Handt, O., M. Krings, R. H. Ward, and S. Pääbo. 1996. The retrieval of ancient human DNA sequences. Am. Hum. Genet. **59**:368–376.

Handt, O., M. Richards, M. Trommsdorff et al. (13 coauthors). 1994*b*. Molecular genetic analyses of the Tyrolean ice man. Science **264**:1775–1778.

Höss, M., P. Jaruga, T. H. Zastawny, M. Dizdaroglu, and S. Pääbo. 1996. DNA damage and DNA sequence retrieval from ancient tissues. Nucleic Acids Res. **24**:1304–1307.

Krings, M., A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking, and S. Pääbo. 1997. Neandertal DNA sequences and the origin of modern humans. Cell **90**:19–30.

Lindahl, T. 1993*a*. Recovery of antediluvian DNA. Nature **365**:700.

———. 1993*b*. Instability and decay of the primary structure of DNA. Nature **362**:709–715.
———. 1997. Facts and artifacts of ancient DNA. Cell **90**:1–3.
———. 2000. Fossil DNA. Curr. Biol. **10**:616.
PÄÄBO, S. 1990. Amplifying ancient DNA. Pp. 159–166 *in* M. A. INNIS, D. H. GELFAND, J. J. SNINSKY, and T. J. WHITE, eds. PCR protocols: a guide to methods and applications. Academic Press, San Diego.
PÄÄBO, S., R. G. HIGUCHI, and A. C. WILSON. 1989. Ancient DNA and the polymerase chain reaction. J. Biol. Chem. **264**:9709–9712.
PÄÄBO, S., and A. C. WILSON. 1991. Miocene DNA sequence—a dream come true? Curr. Biol. **1**:45–46.
SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, G. T. HORN, K. B. MULLIS, and H. A. ERLICH. 1988. Primer-directed enzymatic amplification of DNA with thermostable DNA polymerase. Science **239**:487–491.
TINDALL, K. R., and T. A. KUNKEL. 1988. Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase. Biochemistry **27**:6008–6013.
WILLERSLEV, E., A. J. HANSEN, B. CHRISTENSEN, J. P. STEFFENSEN, and P. ARCTANDER. 1999. Diversity of Holocene life forms in fossil glacier ice. Proc. Natl. Acad. Sci. USA **96**:8017–8021.

FUMIO TAJIMA, reviewing editor

Accepted October 9, 2000