LETTERS

Accommodating the Effect of Ancient DNA Damage on Inferences of Demographic Histories

Andrew Rambaut, Simon Y.W. Ho,† Alexei J. Drummond,‡ and Beth Shapiro§*

*Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh, United Kingdom; †Centre for Macroevolution and Macroecology, School of Botany and Zoology, Australian National University, Canberra ACT, Australia; ‡Department of Computer Science, University of Auckland, Auckland, New Zealand; and §Department of Biology, The Pennsylvania State University

DNA sequences extracted from ancient remains are increasingly used to generate large population data sets, often spanning tens of thousands of years of population history. Bayesian coalescent methods such as those implemented in the software package BEAST can be used to estimate the demographic history of these populations, sometimes resulting in complex scenarios of fluctuations in population size, which can be correlated with the timing of environmental events, such as glaciations. Recently, however, Axelsson et al. (Axelsson E, Willerslev E, Gilbert MTP, Nielsen R. 2008. The effect of ancient DNA damage on inferences of demographic histories. Mol Biol Evol 25:2181–2187.) claimed that many of these complex demographic trends are likely to be the result of postmortem DNA damage, a problem that they investigate by removing all sites involving transitions from ancient sequences prior to analysis. When this solution is applied to a previously published data set of Pleistocene bison, they show that the demographic signal of population expansion and decline disappears. Although some apparently segregating mutations in ancient sequences may be due to postmortem damage, we argue that discarding the data will result in loss of power to detect patterns of population change. Instead, to accommodate this problem, we implement a model in which sequences are the result of a joint process of molecular evolution and postmortem DNA damage within a probabilistic inference framework. Through simulation, we demonstrate the ability of this model to accurately recover evolutionary parameters, demographic history, and DNA damage rates. When this model is applied to the bison data set, we find that the rate of DNA damage is significant but low and that the reconstruction of population size history is nearly identical to previously published estimates.

Introduction

The development of techniques to extract and amplify DNA sequences from fossil remains has provided a novel perspective to molecular evolutionary analyses. Large ancient DNA (aDNA) data sets spanning nearly 60,000 years of population history have revealed dramatic fluctuations in genetic diversity across time and space, making it possible to test hypotheses about, for example, the long-term effects of climate change on large mammals (Shapiro et al. 2004; Drummond et al. 2005; Barnes et al. 2007; Debruyne et al. 2008) and the timing and nature of domestication bottlenecks (Edwards et al. 2007; Finlay et al. 2007; Ho et al. 2008). To infer demographic history, each of these analyses used the software package BEAST (Drummond and Rambaut 2007), which provides a flexible framework for hypothesis testing with time-structured molecular sequence data. In BEAST, changes in effective population size through time are inferred according to the assumptions of coalescent theory (Kingman 1982; Griffiths and Tavare 1994), using parametric or nonparametric models of variable population size. Because BEAST employs Bayesian Markov chain Monte Carlo (MCMC) to average over all possible evolutionary histories, the results of demographic analyses can be summarized as plots of effective population size through time (Drummond et al. 2005).

All phylogenetic methods make simplifying assumptions about the evolutionary process. Throughout the history of phylogenetics, violations of these assumptions have invariably arisen, eventually leading to the development of more sophisticated models that more accurately reflect the evolutionary process. In this way, the field has evolved to deal with difficulties such as unequal base frequencies (Felsenstein 1981), variation in nucleotide substitution rates (Lanave et al. 1984), and violations of the molecular clock (Thorne et al. 1998). In their recent paper, Axelsson et al. (2008) discuss a source of violations that is both specific to and ubiquitous in aDNA data sets: postmortem DNA damage. They correctly argue that DNA damage, which manifests as extraneous mutations along lineages leading to "ancient" specimens, has the potential to confound evolutionary and demographic analyses. Axelsson et al. then demonstrate through simulation how high rates of DNA damage can cause the inference of artifactually complex demographic histories where none are required. The dominant form of postmortem single-base modifications are C to T changes, which will be indistinguishable from G to A changes because damage can occur on either strand. Thus, Axelsson et al. investigate removing all sites containing transitions from alignments of aDNA sequences. Perhaps unsurprisingly, when all segregating transitions, regardless of their frequency in the population, are removed from a data set of Pleistocene and modern bison (Shapiro et al. 2004), the previously reported demographic signal disappears.

Here, we argue that simply removing data is an inappropriate response, as it will remove the majority of genuine signals of changes in effective population size so that the retrieved demographic will be governed mostly by the prior. Therefore, although the reconstruction of a population growth and crash for the bison data set is superficially similar to the artifact identified by Axelsson et al., this does not mean damage is the cause. Indeed, changes in fossil abundance and diversity throughout the Late Pleistocene as well as modern molecular evidence of a recent, severe,

Key words: ancient DNA, demographic history, coalescent, Bayesian MCMC.

E-mail: a.rambaut@ed.ac.uk.

Mol. Biol. Evol. 26(2):245–248. 2009
doi:10.1093/molbev/msn256
Advance Access publication November 11, 2008

Table 1
The Result of the Simulation Study

	Substitution Rate ($\times 10^{-7}$ Subst/Site/Year)	Kappa	Damage Rate ($\times 10^{-7}$ Errors/Site/Year)
Simulated value	1.5	10.0	0.7
No-damage model			
Mean [mean HPDs]	2.44 [1.95, 2.96]	23.2 [14.9, 32.6]	n/a
Type I error[a]	84.0%	99.5%	
Damage model			
Mean [mean HPDs]	1.51 [1.14, 1.90]	10.5 [6.4, 15.1]	0.71 [0.58, 0.83]
Type I error[a]	7.0%	5.5%	5.0%

[a] The proportion of simulations where the true simulated value fell outside the 95% HPD intervals.

population bottleneck provide strong corroborating evidence of such an evolutionary history; it was the timing of such events that was being investigated by Shapiro et al. (2004). Furthermore, as we will show here, when the process of postmortem damage (PMD) is modeled appropriately, the signal of growth and decline remains.

To avoid, or identify and correct, miscoding lesions in aDNA sequences, a variety of experimental protocols have been developed, most focusing on sequence replication. In producing the bison data set, overlapping fragments were amplified, cloned, and multiple polymerase chain reaction products were sequenced for >100 specimens. Replication of the entire extraction and amplification process was undertaken for nearly 15% of specimens at laboratories in Oxford, London, and Copenhagen. These measures will not necessarily avoid every incidence of PMD, but the results of these replication experiments all suggest remarkable preservation of the majority of specimens (as might be expected for permafrost-preserved remains), and thus, most of the observed segregating transitions are more likely to be real than the result of PMD.

To account for DNA damage in the bison and other aDNA data sets, we implemented in BEAST the model that Axelsson et al. (2008) used to simulate their test alignments, which we will refer to as the PMD model. This model allows each observed state in an alignment to probabilistically be the result of a PMD event. As DNA damage will accumulate through time, we assume that it is more likely for sequences derived from older specimens to have miscoding lesions. To model this, the probability that any given nucleotide remained undamaged is assumed to decay exponentially with sample age. This assumption can be adjusted to more accurately reflect preservation, for example, by parameterizing the damage model with the thermal ages of the samples, rather than radiocarbon ages.

Table 1 provides the results of a simulation study with 200 replicates (see Methods for details). Not accommodating damage when it is present results in overestimation of parameter values for the rate of evolution and the transition–transversion bias. As shown by Axelsson et al., PMD will also result in artifacts in demographic reconstruction (fig. 1*a*), where damage manifests as a period of steep population growth followed by decline. However, when the same simulated data are analyzed using the PMD model, the evolutionary parameters (table 1) and demographic reconstruction (fig. 1*b*) are recovered correctly.

Axelsson et al. suggest that the findings of Shapiro et al. (2004) and Drummond et al. (2005) are due to the artifact induced by PMD. However, the PMD model applied to these data gives nearly identical results to those previously published, both in terms of estimated parameter values (data not shown) and the reconstructed demographic history (fig. 2). The estimated rate of damage was 1.96×10^{-8} damaged sites per year (95% highest posterior density [HPD] interval: 0.90×10^{-8}, 2.99×10^{-8}) about a factor of 3 less than the level used in the simulation studies. This provides an expectation that the oldest (hence most damaged) sequences contain only 0.74 damaged sites. This level of PMD produced no qualitative change in the demographic reconstruction or our conclusions from it (fig. 2). Although PMD should, and now can, be accommodated in aDNA analyses, extensive, additional sequence replication as proposed by Axelsson et al. is not necessary if standard aDNA protocols are followed. In addition to accurately accommodating damage, the PMD model can be used to estimate the extent of damage in a data set so as to determine whether additional replication is required. The PMD model has been implemented in BEAST and is freely available from http://beast.bio.ed.ac.uk/.

Methods

Previous studies investigating the effects of PMD in the context of coalescent analyses (Ho et al. 2007; Mateiu and Rannala 2008) modeled this process as a fixed amount of extra evolution on each external branch in the tree. Here, we implement a model that allows the observed sequence to probabilistically be the result of a PMD event. This makes use of the fact that for the standard nucleotide likelihood formulation (Felsenstein 1981), it is necessary to specify the probability of the four nucleotide states for each site on an external node in the tree. Assuming the sequence is known without error, the probability for the observed state is 1 and for the other states is 0, but uncertainty can be included in the actual state observed at each site. This approach has been previously proposed as a mechanism to accommodate errors in the process of sequencing (Felsenstein 2004).

To model PMD, we assume the probability that a site is undamaged decays exponentially with rate r. For each site in each sequence, the probability that the site, S, is nucleotide state j given the observed state is i is given as

$$P(S = j|\text{obs} = i) = \begin{cases} e^{-rt} & \text{if } i = j, \\ 1 - e^{-rt} & \text{if } i \text{ and } j \text{ differ by a transition,} \\ 0 & \text{if } i \text{ and } j \text{ differ by a transversion.} \end{cases}$$

Within the Bayesian MCMC framework, BEAST (Drummond and Rambaut 2007), the rate of damage, r,

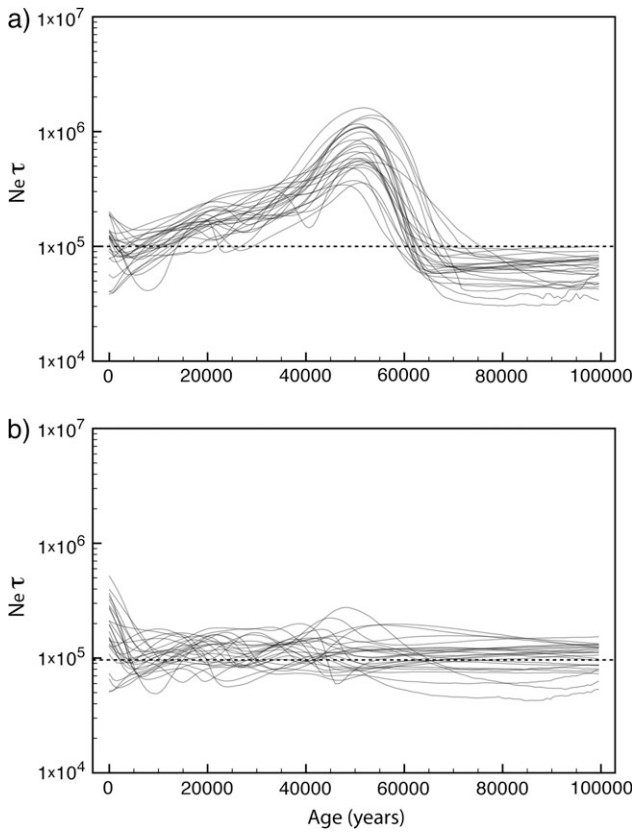

FIG. 1.—Reconstructed demographic histories for 25 alignments simulated under a constant population size model with PMD at a rate of 0.7×10^{-7} errors per site per year (see Methods for details). Reconstructions were performed *a*) assuming no postmortem DNA damage and *b*) incorporating the effect of PMD using the PMD model. The dashed lines show the "true" simulated demographic history.

is sampled to obtain an estimate of the marginal posterior probability density for this parameter.

Conditioned on the sampling times assumed by Axelsson et al. (2008), we generated genealogies using a constant-size coalescent process (with the product of effective population size and generation time set to 10^5). For each, sequences (of length 606 nucleotides) were simulated under a Kimura substitution model (Kimura 1980) with the transition–transversion parameter, kappa, set to 10. The rate of substitution was set to 1.5×10^{-7} substitutions/site/year. The process of PMD was simulated by inducing a transition at each site with a probability $1 - e^{-rt}$, where t is the age of the sequence and the damage rate, r, was assumed to be 0.7×10^{-7} errors per site per year. Although the parameter values for the bison simulations are not reported by Axelsson et al. (2008), the values used here match those used for their figure 5 (Axelsson E, personal communication). The simulated data were analyzed using BEAST v1.5, both with and without the PMD model. For the PMD model, a uniform prior between 0 and 1 was assumed for the damage rate parameter. The demographic function was modeled as a Bayesian Skyline with 10 sampling intervals (Drummond et al. 2005). All other priors and settings match the defaults for BEAST 1.4.8.

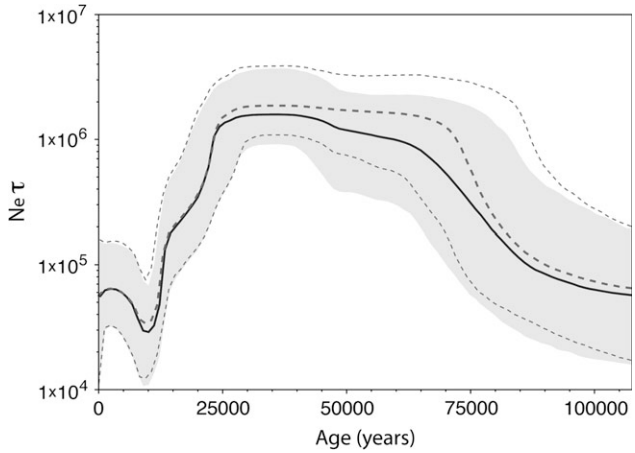

FIG. 2.—The effect on the reconstructed demographic history of bison using the PMD model. The dashed lines show the mean Bayesian Skyline reconstruction and 95% HPD intervals assuming no PMD. The solid line and intervals show the equivalent reconstruction including the PMD model.

Acknowledgments

A.R. is funded by The Royal Society. S.Y.W.H. is funded by the Australian Research Council. We would like to thank Erik Axelsson and Thomas Gilbert for some constructive dialog.

Literature Cited

Axelsson E, Willerslev E, Gilbert MTP, Nielsen R. 2008. The effect of ancient DNA damage on inferences of demographic histories. Mol Biol Evol. 25:2181–2187.

Barnes I, Shapiro B, Lister A, Kuznetsova T, Sher A, Guthrie D, Thomas MG. 2007. Genetic structure and extinction of the woolly mammoth, *Mammuthus primigenius*. Curr Biol. 17: 1072–1075.

Debruyne R, Chu G, King CE, et al. (21 co-authors). 2008. Out of America: ancient DNA evidence for a new world origin of Late Quaternary woolly mammoths. Curr Biol. 18:1–7.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 22:1185–1192.

Edwards CJ, Bollongino R, Scheu A, et al. (40 co-authors). 2007. Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. Proc Roy Soc B Biol Sci. 274: 1377–1385.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17:368–376.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Finlay EK, Gaillard C, Vahidi SM, Mirhoseini SZ, Jianlin H, Qi XB, El-Barody MA, Baird JF, Healy BC, Bradley DG. 2007. Bayesian inference of population expansions in domestic bovines. Biol Lett. 3:449–452.

Griffiths RC, Tavare S. 1994. Simulating probability distributions in the coalescent. Theor Popul Biol. 46:131–159.

Ho SY, Heupink TH, Rambaut A, Shapiro B. 2007. Bayesian estimation of sequence damage in ancient DNA. Mol Biol Evol. 24:1416–1422.

Ho SY, Larson G, Edwards CJ, Heupink TH, Lakin KE, Holland PW, Shapiro B. 2008. Correlating Bayesian date estimates with climatic events and domestication using a bovine case study. Biol Lett. 4:370–374.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

Kingman J. 1982. The coalescent. Stoch Proc Applic. 13:235–248.

Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol. 20:86–93.

Mateiu LM, Rannala B. 2008. Bayesian inference of errors in ancient DNA caused by post mortem degradation. Mol Biol Evol. 25:503–511.

Shapiro B, Drummond AJ, Rambaut A, et al. (27 co-authors). 2004. Rise and fall of the Beringian steppe bison. Science. 306:1561–1565.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol. 15:1647–1657.

Connie Mulligan, Associate Editor

Accepted October 30, 2008