Am. J. Hum. Genet. 75:240–250, 2004

Authenticity of Ancient-DNA Results: A Statistical Approach

Matthew Spencer and Christopher J. Howe

Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

Although there have been several papers recommending appropriate experimental designs for ancient-DNA studies, there have been few attempts at statistical analysis. We assume that we cannot decide whether a result is authentic simply by examining the sequence (e.g., when working with humans and domestic animals). We use a maximum-likelihood approach to estimate the probability that a positive result from a sample is (either partly or entirely) an amplification of DNA that was present in the sample before the experiment began. Our method is useful in two situations. First, we can decide in advance how many samples will be needed to achieve a given level of confidence. For example, to be almost certain (95% confidence interval 0.96–1.00, maximum-likelihood estimate 1.00) that a positive result comes, at least in part, from DNA present before the experiment began, we need to analyze at least five samples and controls, even if all samples and no negative controls yield positive results. Second, we can decide how much confidence to place in results that have been obtained already, whether or not there are positive results from some controls. For example, the risk that at least one negative control yields a positive result increases with the size of the experiment, but the effects of occasional contamination are less severe in large experiments.

Introduction

To ensure the authenticity of ancient-DNA results is difficult because the target DNA is degraded and is present in small quantities, and there is a high risk of contamination. The accepted criteria for ancient-DNA work, therefore, include reproducibility and the use of negative controls (e.g., Stoneking 1995; Cooper and Poinar 2000; Hummel 2003, p. 150), in an attempt to demonstrate that the amplified DNA did not come from contaminants introduced during the experiment. A negative control is a PCR reaction that is known not to contain authentic ancient DNA. Examples include extraction blanks, to which no ancient material was added, and PCR blanks, to which water was added instead of a sample extract. Reproducibility and the use of negative controls are sensible criteria, but there have been few attempts at quantitative analysis of ancient-DNA experiments. For example, although the absence of positive results from negative controls makes us more confident that positive results from samples are authentic, we cannot say exactly how confident we are. It is always possible that no negative control but at least one sample was contaminated or that contaminant DNA was present both in samples and negative controls but was amplified only in samples. In consequence, we cannot make rational decisions about the number of independent samples needed, and we cannot assess the reliability of experiments that have been done in the past. In this article, we use maximum-likelihood methods to address these problems.

We consider two definitions of an authentic result. The first definition is a sequence amplified entirely from sample DNA (DNA that was present in the sample before the experiment began) or a sequence that is a mixture of amplified sample and contaminant DNA (where a contaminant is any DNA introduced after the experiment began). The second, stricter definition is a sequence amplified entirely from sample DNA, with mixtures of sample and contaminant DNA excluded. Throughout, we use "amplified" to mean "amplified to a detectable level," but we do not use any quantitative information on the number of molecules.

In most cases, we cannot distinguish between DNA that was originally present in the sample and DNA that was not originally present but was introduced before the beginning of the experiment (the exception is when we have controls that have the same history as the samples but that are known not to contain authentic DNA). We assume that we cannot decide for certain whether a result is sample or contaminant simply by examining the sequence. For example, we have been attempting to amplify a fragment of the mitochondrial cytochrome-b gene from parchment samples to identify the species of animal used to make the parchment (M.S., C. de Hamel, and C.J.H., unpublished data). Since most of the potential species are common domestic animals, and some laboratory disposables are contaminated with cow cytochrome b (Hummel 2003, p. 140), the authentic sequences we are likely to obtain are also plausible con-

Received January 29, 2004; accepted for publication May 21, 2004; electronically published June 15, 2004.

Address for correspondence and reprints: Dr. Matthew Spencer, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, B3H 3J5, Canada. E-mail: matts@mathstat.dal.ca

© 2004 by The American Society of Human Genetics. All rights reserved.

0002-9297/2004/7502-0009$15.00

taminants. Authentic ancient-DNA sequences may be damaged in recognizable ways (Gilbert et al. 2003), and we expect that only short fragments of authentic ancient DNA will be amplifiable (e.g., Handt et al. 1994). It may, therefore, be possible to identify long undamaged sequences as modern contaminants. On the other hand, a short damaged fragment is not necessarily authentic, because laboratory techniques, such as autoclaving and cleaning with sodium hypochlorite (which produces peroxide free radicals), can result in similar DNA damage (Willerslev et al. 2004).

If the rate of positive results is not significantly higher in samples than in controls, we would be unlikely to believe the results are authentic. Thus, we could use contingency tables to evaluate the null hypothesis of no difference in the rate of positives (Agresti 2002, chapters 2–3). Because the number of observations will usually be small, exact tests based on permutation will be the most appropriate (Mehta and Patel 1997). This approach is very simple. Here, we suggest a slightly more complicated maximum-likelihood method. A little more effort is required in computation, but we obtain more information. For example, if a contingency-table analysis suggests a higher rate of positives in samples, we still do not know the probability that a positive result is authentic. By explicitly modeling the possible ways we can obtain a positive result, our method supplies this additional information.

In the "Theory" section, we first describe in detail the necessary statistical methods for the simple case in which there is only one kind of control. We estimate the probability that a positive result from a sample is an amplification of sample DNA alone or of a mixture of sample and contaminant DNA. We show how to compare the likelihoods of the hypotheses that this probability is greater than zero or that this probability is zero. We also calculate a 95% CI for this probability. Then, we give formulas for three other cases: experiments with one kind of control, where we want the probability that a positive result from a sample is an amplification of sample DNA alone; experiments with two kinds of control (extraction blanks and PCR blanks), where we want the probability that a positive result from a sample is either an amplification of sample DNA alone or of a mixture of sample and contaminant DNA; and experiments with two kinds of control, where we want the probability that a positive result from a sample is an amplification of sample DNA alone. The methods used in these other cases are the same as in the simple case, but the analysis is slightly more complicated.

In the "Applications" section, we give three examples. First, we calculate how our confidence in the results of an ancient-DNA experiment is affected by the size of the experiment and by the proportion of samples giving positive results, when no negative controls yield positive results. Second, we show, for two different experiment sizes, how much the presence of positive results in some negative controls reduces our confidence in the results. Third, we analyze the results of experiments with ancient-DNA extraction from parchment.

Theory

One Kind of Control, Probability that a Positive Result Is at Least Partly from Sample DNA

We first discuss the simple case in which there is only one kind of negative control and in which we want the probability that a positive result from a sample is either an amplification of sample DNA alone or of a mixture of sample and contaminant DNA. The other possible outcome is that a positive result from a sample is the result of contaminant DNA alone, and the sum of these three probabilities is 1. Let $P(r)$ be the probability of a positive result from a sample extract (either a band on a gel or a sequence result). The result may be due to sample or contaminant DNA, or both. Let $P(s)$ be the probability that there is sample DNA in the sample, and let $P(r_s \mid s)$ be the probability that sample DNA is amplified to a detectable level if present. Let $P(c)$ be the probability that there is contaminant DNA present in the sample, and let $P(r_c \mid c)$ be the probability that contaminant DNA is amplified to a detectable level if present. The probabilities that DNA is amplified to a detectable level if present will depend on the number of molecules initially present, the PCR conditions (e.g., the efficiency of the polymerase enzyme and the number of cycles), and the presence of carrier effects and PCR inhibitors. Both sample and contaminant DNA may be present and amplified, and we assume that we cannot distinguish between amplifications of sample and contaminant DNA from the sequences alone. Contaminants that can be identified with certainty should be treated as negative results.

We want the conditional probability $P(r_s \mid r)$ that a positive result is (at least in part) an amplification of sample DNA, whether or not there is also contaminant DNA present, and amplified

$$P(r_s \mid r) = \frac{P(r_s)}{P(r)} \ . \tag{1}$$

In this first case, we include amplifications that contain both sample and contaminant DNA. If we cannot tell the components apart, it sometimes makes sense to treat these cases as positive results (e.g., if we are working with sequences that allow us to identify species but not individuals). We later examine cases where only sample DNA is included. We assume that the events s (authentic

sample DNA is present) and c (contaminant DNA is present) occur independently. To enumerate the three different ways we could obtain a positive result,

$$P(r) = P(r_s|s)P(s)[1 - P(c)] + P(r_c|c)P(c)[1 - P(s)] + P(r|s,c)P(s)P(c) \ , \quad (2)$$

where $P(r|s,c)$ is the probability of a positive result if both sample and contaminant DNA are present. In this sum, the first term is the probability that sample DNA is present and is amplified to a detectable level and that no contaminant DNA is present. The second term is the probability that contaminant DNA is present and is amplified to a detectable level and that no sample DNA is present. The third term is the probability that both sample and contaminant DNA are present and that at least one of them is amplified to a detectable level.

We further assume that the amplification and detection of sample and contaminant DNA are independent. This does not mean we assume that they are amplified with equal efficiency, only that the efficiency of amplification of one kind of DNA is unaffected by the presence of the other kind. This may be reasonable if DNA polymerase, primers, and dNTPs are available in excess during the amplification; samples do not contain large amounts of PCR inhibitors; and there are no strong carrier effects (enhancements of the probability of contaminant amplification by components of the sample extract). We revisit this assumption in the "Discussion" section. Then,

$$P(r|s,c) = P(r_s|s)[1 - P(r_c|c)] + P(r_c|c)[1 - P(r_s|s)] + P(r_c|c)P(r_s|s) \ . \quad (3)$$

In this sum, the first term is the probability that only sample DNA is amplified to a detectable level, the second term is the probability that only contaminant DNA is amplified to a detectable level, and the third term is the probability that both are amplified to a detectable level. Without the assumption that amplification and detection of sample and contaminant DNA are independent, we would need a separate estimate of $P(r \mid s,c)$, which we cannot obtain from this kind of experiment. Combining equations (1), (2), and (3) and writing $P(r_s|s)P(s) = \theta_1$, $P(r_c|c)P(c) = \theta_2$, and $P(r_s|r) = \gamma$ for conciseness, we have

$$\gamma = \frac{\theta_1}{\theta_1 + \theta_2 - \theta_1\theta_2} \ . \quad (4)$$

Suppose we run an experiment from which we obtain data $\mathbf{D} = [n_b,b_b,n_s,b_s]$, where there are n_b negative controls, of which b_b give positive results, and n_s samples, of which b_s give positive results. We assume that each replicate is an independent unit. Among other things, this means that each sample is a separate extraction and that all factors that might influence the results either are held constant over all replicates (e.g., by use of the same stock solution for all) or are randomized (e.g., by use of a separate cutting tool to prepare each sample). Designs in which replicates are grouped into blocks (e.g., repeat amplifications from the same extraction, or groups of extractions, each with a different set of tools) are more difficult to analyze.

Under the assumption that the occurrence and amplification of sample and contaminant DNA are independent, we can combine the data from samples and controls to estimate the parameters of the model. We know that the controls contain only contaminant DNA (θ_1 and γ are zero in the controls). The samples provide an estimate of $\theta_1 + \theta_2 - \theta_1\theta_2$, because they may contain either sample or contaminant DNA. The likelihood of the data, given the parameters θ_1 and θ_2, is the product of two binomials:

$$P(\mathbf{D} \mid \theta_1,\theta_2) = \binom{n_b}{b_b}\theta_2^{b_b}(1 - \theta_2)^{n_b - b_b} \times \binom{n_s}{b_s}(\theta_1 + \theta_2 - \theta_1\theta_2)^{b_s} \times (1 - \theta_1 - \theta_2 + \theta_1\theta_2)^{n_s - b_s} \ . \quad (5)$$

Here, the first binomial is the contribution to the likelihood from the controls. This term contains only θ_2, because we know there is no authentic DNA in the controls. The second binomial is the contribution to the likelihood from the samples and contains both θ_1 and θ_2, because both authentic and contaminant DNA may be present. We want to rewrite the likelihood so that we can find the maximum-likelihood estimate of γ, the value that maximizes equation (5). By rearranging equation (4), we can eliminate either θ_1 or θ_2. It is probably more sensible to eliminate θ_1, because θ_2 can be estimated more directly from the data. Thus, rearranging (4) gives

$$\theta_1^* = \frac{\gamma\theta_2}{1 - \gamma(1 - \theta_2)} \ , \quad (6)$$

the value of θ_1 needed to obtain a specified value of γ, given θ_2. We then substitute equation (6) into (5),

$$P(\mathbf{D} \mid \gamma,\theta_2) = \binom{n_b}{b_b}\theta_2^{b_b}(1 - \theta_2)^{n_b - b_b} \times \binom{n_s}{b_s}\left[\frac{\theta_2}{1 - \gamma(1 - \theta_2)}\right]^{b_s} \times \left[1 - \frac{\theta_2}{1 - \gamma(1 - \theta_2)}\right]^{n_s - b_s} \quad (7)$$

(note that if $\theta_2 = 0$, we simply substitute $\theta_1 = \gamma$ into eq. [5]), and differentiate to find the maximum-likelihood estimates of θ_2 and γ:

$$\begin{cases} \hat{\theta}_2 = \dfrac{b_b + b_s}{n_b + n_s}, \hat{\gamma} = 0 & \text{if } \dfrac{b_s}{n_s} < \dfrac{b_b}{n_b} \\ \hat{\theta}_2 = \dfrac{b_b}{n_b}, \hat{\gamma} = \dfrac{b_s n_b - b_b n_s}{b_s(n_b - b_b)} & \text{otherwise} \end{cases} \tag{8}$$

(throughout, we use hat symbols to indicate maximum-likelihood estimates). θ_2 is a nuisance parameter; we are not very interested in its value, but we cannot eliminate it.

These estimates make intuitive sense and have a simple interpretation. If the fraction of positives in samples is less than that in the controls, we do not believe any of the positive results are authentic. Then, we set $\hat{\gamma}$ to zero, and our estimate of $\hat{\theta}_2$ is the total number of positives divided by the total number of samples and controls. Otherwise, our estimate of $\hat{\theta}_2$ is the fraction of positives in controls, and we can rewrite our estimate of $\hat{\gamma}$ as:

$$\hat{\gamma} = \frac{\left(\dfrac{b_s}{n_s} - \dfrac{b_b}{n_b}\right)}{\left[\dfrac{b_s}{n_s}\left(1 - \dfrac{b_b}{n_b}\right)\right]} . \tag{9}$$

We can see from equation (2) that the numerator of equation (9) is an estimate of $P(r_s)[1 - P(r_c)]$ and that the denominator is an estimate of $P(r)[1 - P(r_c)]$. Therefore, (9) is a natural estimate of $P(r_s|r) = P(r_s)/P(r)$, the conditional probability that a positive result is, at least in part, an amplification of sample DNA (eq. [1]).

Hypothesis Tests and CIs

If we analyze a single sample and a single extraction blank and obtain a positive result only from the sample ($n_b = 1$, $b_b = 0$, $n_s = 1$, and $b_s = 1$), the maximum-likelihood estimate of γ is 1. We get the same maximum-likelihood estimate if we analyze 10 samples and 10 extraction blanks, obtaining positive results from all 10 samples and from none of the blanks ($n_b = 10$, $b_b = 0$, $n_s = 10$, and $b_s = 10$). It is clear that we should be more confident that the results from the second experiment were genuine ancient-DNA amplifications. Hypothesis tests and CIs for γ quantify this intuition.

A hypothesis test for γ can be based on the question: how much more likely are the data, given the maximum-likelihood estimates $\hat{\gamma}$ and $\hat{\theta}_2$, than they are under the null hypothesis that $\gamma = 0$? If the null hypothesis is true, then the log-likelihood–ratio statistic

$$R = -2\Big(L[\mathbf{D} \mid \gamma = 0, \hat{\theta}_2(\gamma = 0)] - L[\mathbf{D} \mid \hat{\gamma}, \hat{\theta}_2]\Big) \tag{10}$$

has an asymptotic χ^2 distribution with 1 df (McCullagh and Nelder 1989, p. 476; Hilborn and Mangel 1997, pp. 153–154). Here, L is the log likelihood (the natural log of eq. [7]), and $\hat{\theta}_2(\gamma = 0)$ is the conditional maximum-likelihood estimate of θ_2 when $\gamma = 0$. From equation (8), this estimate is simply $(b_b + b_s)/(n_b + n_s)$. Note that the binomial coefficients in equation (7) cancel out in equation (10), so there is no need to evaluate them. For example, with $n_b = 1$, $b_b = 0$, $n_s = 1$, and $b_s = 1$, $R = -2[\log(1/4) - \log(1)] = 2.77$. From standard tables of the χ^2 distribution, we cannot reject the null hypothesis that $\gamma = 0$ ($P = .096$). With $n_b = 2$, $b_b = 0$, $n_s = 2$, and $b_s = 2$, $R = -2[\log(1/16) - \log(1)] = 5.55$, and we can reject the null hypothesis ($P = .019$). With $n_b = 10$, $b_b = 0$, $n_s = 10$, and $b_s = 10$, $R = -2[\log(1/2^{20}) - \log(1)] = 27.73$, and we can very strongly reject the null hypothesis ($P = 1.4 \times 10^{-7}$).

A CI for γ can be constructed by a similar principle and gives us more information, at the cost of more complicated computation. The $100(1 - \alpha)\%$ profile likelihood CI for γ, with one kind of control, is the set of values of γ for which

$$-2\Big(L[\mathbf{D} \mid \gamma, \hat{\theta}_2(\gamma)] - L[\mathbf{D} \mid \hat{\gamma}, \hat{\theta}_2]\Big) \leq q(1 - \alpha) , \tag{11}$$

where L is the log likelihood, $\hat{\theta}_2(\gamma)$ is the conditional maximum-likelihood estimate of θ_2 given γ, and $q(1 - \alpha)$ is the $1 - \alpha$ quantile of the χ^2 distribution with 1 df (e.g., Venzon and Moolgavkar 1988; Hilborn and Mangel 1997, pp. 162–167). In other words, this is the set of values of γ that would not be rejected by a likelihood-ratio test when the other parameters are set to their conditional maximum-likelihood estimates. We can calculate the conditional likelihood analytically and can use numerical minimization to find the boundaries of the set. The formulas for the conditional maximum-likelihood estimates are long, but they are easily found using computer algebra.

One Kind of Control, Probability that a Positive Result Is Entirely from Sample DNA

We might also be interested in the probability that a positive result from a sample is an amplification of sample DNA only. In this case, we want to estimate

$$\gamma_2 = \frac{\theta_1(1 - \theta_2)}{\theta_1 + \theta_2 - \theta_1\theta_2} .$$

By the same methods as above, the maximum-likelihood estimates are

$$\begin{cases} \hat{\theta}_2 = \dfrac{b_b + b_s}{n_b + n_s}, \hat{\gamma}_2 = 0 & \text{if } \dfrac{b_s}{n_s} < \dfrac{b_b}{n_b} \\ \hat{\theta}_2 = \dfrac{b_b}{n_b}, \hat{\gamma}_2 = \dfrac{b_s n_b - b_b n_s}{b_s n_b} & \text{otherwise} \end{cases},$$

and we can use equation (11) to find a 95% CI. The only important difference is the additional constraint $\gamma_2 \leqslant 1 - \theta_2$, because $\theta_1 \leqslant 1$.

Two Kinds of Control, Probability that a Positive Result Is at Least Partly from Sample DNA

It is common to use two different kinds of control: extraction blanks (the extraction procedure is performed with no sample material added) and PCR blanks (water is added, instead of extract, during PCR setup). If contamination occurs, the two controls make it easier to locate the source. Analyzing this kind of experiment is slightly more complicated than the cases above, but it is done in a similar way.

Let $P(c_e)$ be the probability that there is contamination introduced during extraction, and let $P(c_p)$ be the probability that contamination is introduced during PCR setup. Let $P(r_{ce}|c_e)$ and $P(r_{cp}|c_p)$ be the probabilities that contaminants introduced during extraction and PCR setup are amplified to detectable levels if present. Under the assumption of independent occurrence and amplification of sample and contaminant DNA, as above, and with $P(r_s|s)P(s) = \theta_1$, $P(r_{ce}|c_e)P(c_e) = \theta_2$, and $P(r_{cp}|c_p)P(c_p) = \theta_3$, we obtain

$$\gamma = \frac{\theta_1}{\pi_s},$$

where $\pi_s = \theta_1\theta_2\theta_3 + \theta_1 + \theta_2 + \theta_3 - \theta_1\theta_2 - \theta_1\theta_3 - \theta_2\theta_3$. An experiment with this design gives data $\mathbf{D} = [n_e, b_e, n_p, b_p, n_s, b_s]$, where the subscripts e and p represent extraction and PCR blanks, respectively. The extraction blanks provide an estimate of $\pi_e = \theta_2 + \theta_3 - \theta_2\theta_3$, because they may contain contamination introduced during either extraction or PCR setup. The PCR blanks provide an estimate of θ_3, because they may contain contamination introduced during PCR setup only. The likelihood of the data, given the parameters, is the product of three binomials:

$$P(\mathbf{D}|\theta_1,\theta_2,\theta_3) = \binom{n_p}{b_p}\theta_3^{b_p}(1-\theta_3)^{n_p-b_p} \times \binom{n_e}{b_e}\pi_e^{b_e}(1-\pi_e)^{n_e-b_e} \times \binom{n_s}{b_s}\pi_s^{b_s}(1-\pi_s)^{n_s-b_s}.$$

Substituting

$$\theta_1^* = \frac{\gamma\pi_e}{1-\gamma(1-\pi_e)}$$

(the value of θ_1 needed to obtain a specified γ, given θ_2 and θ_3) and differentiating, we obtain the maximum-likelihood estimates

$$[\hat{\theta}_2, \hat{\theta}_3, \hat{\gamma}] = \begin{cases} \left[\dfrac{b_p n_e - b_e n_p}{n_e(b_p - n_p)}, \dfrac{b_p}{n_p}, \dfrac{b_e n_s - b_s n_e}{b_s(b_e - n_e)}\right] \\ \quad \text{if} \sim c_1 \text{ and } \sim c_2 \\ \left[0, \dfrac{b_e + b_p}{n_e + n_p}, \dfrac{n_s(b_e + b_p) - b_s(n_e + n_p)}{b_s(b_e + b_p - n_e - n_p)}\right] \\ \quad \text{if } c_1, \sim c_2, \text{ and } \sim c_3 \\ \left[\dfrac{b_p(n_e + n_s) - n_p(b_e + b_s)}{(b_p - n_p)(n_e + n_s)}, \dfrac{b_p}{n_p}, 0\right] \\ \quad \text{if} \sim c_1, c_2, \text{ and } \sim c_4 \\ \left[0, \dfrac{b_e + b_p + b_s}{n_e + n_p + n_s}, 0\right] \\ \quad \text{otherwise} \end{cases} \tag{12}$$

under the conditions

$$c_1 = \left(\frac{b_p}{n_p} > \frac{b_e}{n_e}\right),\ c_2 = \left(\frac{b_e}{n_e} > \frac{b_s}{n_s}\right),$$
$$c_3 = \left(\frac{b_e + b_p}{n_e + n_p} > \frac{b_s}{n_s}\right), \text{ and } c_4 = \left(\frac{b_p}{n_p} > \frac{b_e + b_s}{n_e + n_s}\right), \tag{13}$$

where $\sim c_1$ means "c_1 false." We now have two nuisance parameters, θ_2 and θ_3, and both must be set to their conditional maximum-likelihood estimates to find the 95% CI. There are some cases where γ is not identifiable. For example, the data [1,1,1,1,1,1] have likelihood 1 for any values of γ and θ_2, as long as θ_3 is 1.

Two Kinds of Control, Probability that a Positive Result Is Entirely from Sample DNA

In this case, we want to estimate

$$\gamma_2 = \frac{\theta_1(1-\pi_e)}{\pi_s}.$$

Using the same methods as above, we obtain

$$\left[\hat{\theta}_2,\hat{\theta}_3,\hat{\gamma}_2\right] = \begin{cases} \left[\dfrac{b_p n_e - b_e n_p}{n_e(b_p - n_p)}, \dfrac{b_p}{n_p}, \dfrac{b_s n_e - b_e n_s}{b_s n_e}\right] \\ \qquad \text{if} \sim c_1 \text{ and } \sim c_2 \\ \left[0, \dfrac{b_e + b_p}{n_e + n_p}, \dfrac{b_s(n_e + n_p) - n_s(b_e + b_p)}{b_s(n_e + n_p)}\right] \\ \qquad \text{if } c_1, \sim c_2, \text{ and } \sim c_3 \\ \left[\dfrac{b_p(n_e + n_s) - n_p(b_e + b_s)}{(b_p - n_p)(n_e + n_s)}, \dfrac{b_p}{n_p}, 0\right] \\ \qquad \text{if} \sim c_1, c_2, \text{ and } \sim c_4 \\ \left[0, \dfrac{b_e + b_p + b_s}{n_e + n_p + n_s}, 0\right] \\ \qquad \text{otherwise} \end{cases} . \tag{14}$$

Only the first two cases are different from equation (12). The conditions are the same as in equation (13). There is the additional constraint that $\gamma_2 \leqslant 1 - \pi_e$. This makes estimating the 95% CI slightly more complicated because the constraint region is not rectangular, and there are many points along the curved boundary where the derivatives of the likelihood function, with respect to θ_2 and θ_3, are undefined because of division by zero. Thus, it may be necessary to choose the maximum over a grid of points along the boundary if the conditional maximum-likelihood estimates are on the edges of the possible parameter space.

Applications

Here, we discuss two simple cases: the interpretation of results when no controls give positive results and how the presence of positive results in a few controls changes our confidence about positive results from samples. We then present an example from our own experiments. For all cases, we used Maple 6 (Waterloo Maple) to solve the likelihood equations and Matlab release 13 (The Mathworks) for numerical calculations.

Can We Trust Results When No Controls Gave Positive Results?

If we get positive results from no negative controls and at least one sample, the maximum-likelihood estimates of both γ (the probability that a positive result is owing, at least in part, to sample DNA) and γ_2 (the probability that a positive result is entirely owing to sample DNA) are 1. The upper 95% CL is also 1. Nevertheless, we cannot necessarily rule out contamination. Figure 1 shows the lower 95% CL in the best possible case, when all of $n = n_s$ samples and none of an equal number of extraction blanks give positive results. In this case, we can be almost certain that a positive result is at least partly owing to sample DNA (*solid line* in fig. 1) when we have at least five of five sample positives, because the lower 95% CL for $n = 5$ is 0.96. With only a single sample, we cannot reject confidently the possibility that a positive result is only contamination, because the lower 95% CL for $n = 1$ is 0. When designing an experiment, therefore, we certainly need at least two samples and controls, and there are large benefits in adding more samples, at least up to five samples and controls. It makes no difference in this case whether we include PCR blanks as well as extraction blanks, since the PCR blanks do not appear in the equation for γ (eq. [12], first case). PCR blanks will affect the CI in some other cases and will help track down the source of contamination if it occurs.

For a given sample size, we can be much less certain that a positive result is an amplification of sample DNA alone than that it is at least partly an amplification of sample DNA (*dashed line* in fig. 1). For example, we would need a very large experiment ($n \approx 40$, if all samples and no controls give positive results) for the lower 95% CL on γ_2 to be >0.95. This is because γ_2 excludes cases where both sample and contaminant DNA were amplified.

Our ability to rule out contamination depends on the reproducibility of the amplification as well as the number of extractions. If only one of n samples gave a positive result (and no blanks gave a positive result), the 95% CI for γ or γ_2 is 0–1 for any n, and we cannot reject the possibility of contamination. Figure 2 shows how the lower 95% CLs for γ and γ_2 vary with the number of samples (of 10 total samples) giving positive results, under the assumption that none of 10 extraction blanks and PCR blanks give positive results. The curves in figure 2 climb more slowly than those in figure 1. For example, we can be more confident that a sample sequence was amplified only from sample DNA if we get 3 of 3 positives from samples (lower 95% CL for γ_2 is 0.53, under the assumption of three extraction blanks with no positives) than if we get 4 of 10 positives from samples (lower 95% CL for γ_2 is 0.48, under the assumption of 10 extraction blanks with no positives). This underlines the importance of reproducibility as a criterion for the authenticity of ancient-DNA results (e.g., Cooper and Poinar 2000).

Can We Trust Results When Some Controls Revealed Contamination?

Completely eliminating contamination is impossible, so there is always a chance that at least one negative control will yield a positive result. This risk increases as

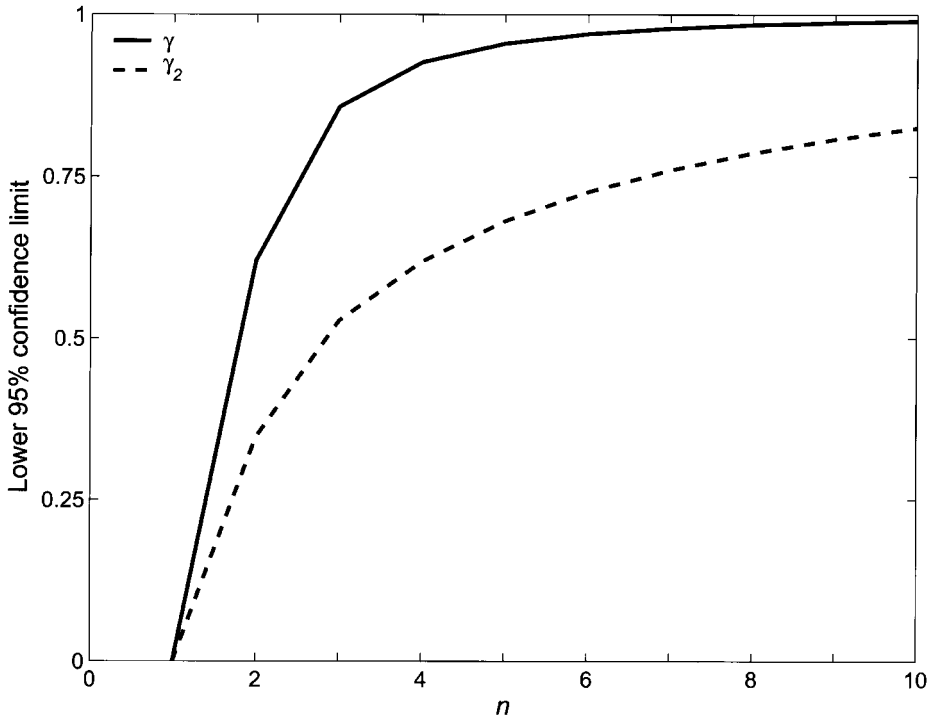

Figure 1 Effects of sample size on lower 95% profile likelihood CLs for γ (*solid line*, probability that a positive result from a sample is an amplification of sample DNA or of a mixture of sample and contaminant DNA) and for γ_2 (*dashed line*, probability that a positive result from a sample is an amplification of sample DNA only). In this example, a positive result is observed in all n sample extracts and none of n extraction blanks and n PCR blanks. The CIs are the same if we run only n blanks of a single kind. The maximum-likelihood estimates of γ and γ_2 and their upper 95% profile CLs are 1, in all cases.

the size of the experiment increases. Nevertheless, if the rate of positive results is much higher in samples than it is in negative controls, we would still be inclined to believe that the majority of positive results are due, at least in part, to sample DNA. As the number of positive results from extraction blanks increases, the maximum-likelihood estimate of γ or γ_2 decreases rapidly. Figure 3 shows a simple example, in which half of the samples and some of the extraction blanks yield positive results. With a larger experiment (20 instead of 10 samples) and the same proportions of samples and extraction blanks giving positive results, the maximum-likelihood estimates are unchanged, but the 95% CIs are narrower. Thus, if one tenth of the extraction blanks and half of the samples give positive results (fig. 3*A*; $b_e/n_e = 0.1$), the CI for γ includes almost the entire range from 0 to 1, if 10 each of extraction blanks, PCR blanks, and samples are run (fig. 3*A*). For the same rates of positives—but 20 each of extraction blanks, PCR blanks, and samples—the lower 95% CL for γ is only slightly below 0.5 (fig. 3*A*). The pattern is similar for γ_2, except that the maximum-likelihood estimate approaches zero linearly and more rapidly as the rate of contamination increases (fig. 3*B*).

For a large experiment with a reasonably high rate of positive results from samples, it may not be necessary to reject the entire experiment if there is a low rate of positive results from negative controls. Nevertheless, as the rate of positive results in negative controls increases, the results of the experiment quickly become unreliable.

Experimental Data: DNA from Parchment

Currently, we are experimenting with methods for extracting DNA from parchment (M.S., C. de Hamel, and C.J.H., unpublished data). To improve our technique, we have been working with five legal documents, dated between 1730 and 1830. We have been using the primers CyBa and CyBb (Burger et al. 2001), which amplify a 147-bp segment of the mitochondrial cytochrome-b gene from most nonhuman mammals. Between July 2003 and March 2004, our main extraction method gave positives from 6 of 62 extraction blanks, 0 of 52 PCR blanks, and 15 of 59 samples (excluding repeat amplifications of the same sample, second-round PCR, and experiments designed to screen out sources of contamination). All were bands of the expected size, and all of those that we sequenced matched *Bos taurus/indicus* cytochrome b.

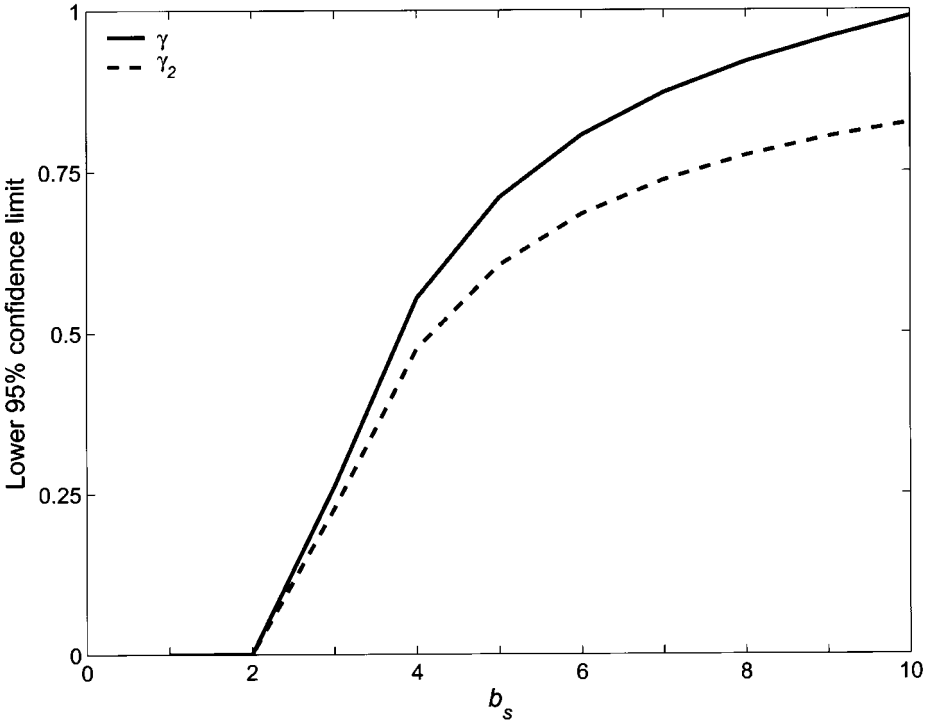

Figure 2 Effects of reproducibility on lower 95% profile likelihood CLs for γ (*solid line*, probability that a positive result from a sample is an amplification of sample DNA or of a mixture of sample and contaminant DNA) and γ_2 (*dashed line*, probability that a positive result from a sample is an amplification of sample DNA only). In this example, a positive result is observed in b_s of 10 sample extracts and none of 10 extraction blanks and 10 PCR blanks. The CIs are the same if we run only 10 blanks of a single kind. The maximum-likelihood estimates of γ and γ_2 and their upper 95% profile CLs are 1, in all cases.

B. taurus is a plausible species for all parchment samples, on the basis of their size and general appearance (C. de Hamel and C. Checkley-Scott, personal communication). *B. taurus* has been reported elsewhere as a contaminant in negative controls when these primers are used (Burger et al. 2001). Applying equations (12) and (14), we obtain the estimates $\hat{\theta}_2 = 0.10$, $\hat{\theta}_3 = 0$, $\hat{\gamma} = 0.69$ (95% CI 0.16–0.90), and $\hat{\gamma}_2 = 0.62$ (95% CI 0.13–0.86). We conclude that about two thirds of our sample positives are likely to result, at least in part, from DNA present in the samples before each experiment began. We can confidently reject the hypothesis that none are genuine, because the lower 95% CLs for both γ and γ_2 are well above zero. We can also confidently reject the hypothesis that all are genuine, because the upper 95% CLs for both are <1. DNA extraction, rather than PCR setup, appears to be the cause of contamination. This may be because the equipment and reagents used for extraction were contaminated or because cleaning and cutting samples cause cross-contamination. We also used a second extraction method, in which we obtained 2 of 25 positives from extraction blanks, 0 of 25 positives from PCR blanks, and 6 of 25 positives from samples. These data give $\hat{\theta}_2 = 0.08$, $\hat{\theta}_3 = 0$, $\hat{\gamma} = 0.72$ (95% CI 0–0.96), and $\hat{\gamma}_2 = 0.67$ (95% CI 0–0.95). For this method, the parameter estimates are similar, but we cannot reject the hypothesis that none of the sample positives are genuine, because of the small number of data. Finally, we performed a few experiments with commercial kits, which gave three of eight positives from extraction blanks, zero of eight positives from PCR blanks, and zero of eight positives from samples; these data give $\hat{\theta}_2 = 0.19$, $\hat{\theta}_3 = 0$, $\hat{\gamma} = 0$ (95% CI 0–0.65), and $\hat{\gamma}_2 = 0$ (95% CI 0–0.59). The high rate of positives in extraction blanks is probably the result of contaminated reagents in the kits.

Discussion

Our aim is to suggest a statistical framework for the interpretation of ancient-DNA results in cases for which we cannot definitely reject sequences as being obvious contaminants. Calculation of the probability that a positive result represents (either in part or entirely) DNA present in the sample before the experiment began is useful in two situations. First, it helps us decide how large an experiment is needed to achieve a given level of confidence. Second, it helps us interpret the results of

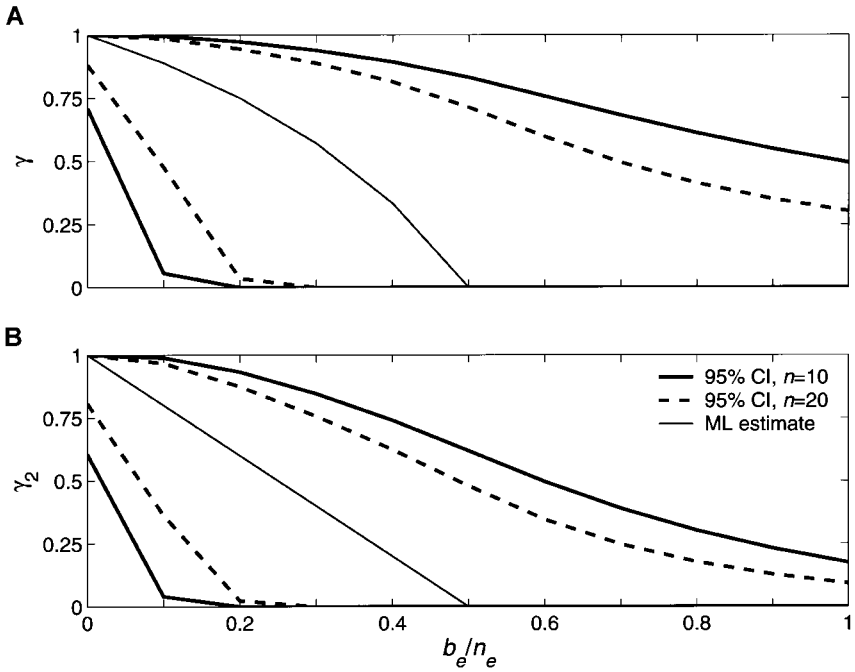

Figure 3 Effects of sample size and of the rate of positives in extraction blanks on 95% CIs for (*A*) γ (probability that a positive result from a sample is an amplification of sample DNA or of a mixture of sample and contaminant DNA) and (*B*) γ_2 (probability that a positive result from a sample is an amplification of sample DNA only) for two different sizes of experiment: $n = 10$ (two thick solid lines indicate lower and upper values of 95% CI) or $n = 20$ (two dashed lines indicate lower and upper values of 95% CI], with n each of extraction blanks, PCR blanks, and samples. In each case, the maximum-likelihood (ML) estimate is the thin solid line (the same for both sample sizes). In this example, half the samples and none of the PCR blanks gave positive results, and the rate of positive results in extraction blanks is b_e/n_e.

experiments that have been done in the past. We think that it would be sensible to include some statistical analysis with all ancient-DNA work. Molecular biologists have not resorted traditionally to statistical analysis, because their experiments are believed to be highly reproducible. In contrast, ancient-DNA work is often difficult to replicate (e.g., Austin et al. 1997) and is highly vulnerable to contamination. In this respect, perhaps ancient-DNA work has more in common with sciences such as ecology, in which statistical analysis is essential. We are aware of only two related approaches. Weiss and von Haeseler (1997) show how to estimate the probability that all initial template molecules were the same, given a set of clones from a single amplification. Willerslev et al. (2003) use a bootstrap test to show that samples of sediment, originating close together in time, contained similar distributions of sequences.

Our results lead us to suggest some changes in the acceptable standards for experimental design. For example, the DNA Commission of the International Society for Forensic Genetics (Bär et al. 2000) makes the following recommendations for quality assurance in the analysis of human mtDNA from degraded forensic samples (a problem with many similarities to those of ancient-DNA studies): "Although single analyses can produce reliable results, it is desirable to carry out analysis twice on separate occasions to better interpret the effects of contamination.... If either the extraction reagent blank or the PCR negative control yields a sequence that is the same as that of the evidence sample, the results from the evidence sample must be rejected and the analysis repeated" (Bar et al. 2000, p. 194). We disagree on two points. First, we showed that with a single successful extraction, we cannot confidently exclude the hypothesis that the result is owing to contamination, even in the absence of positive results from negative controls. Thus, at least two extractions are essential, not just desirable. Second, if some negative controls give positive results, we can quantify the probability that positive results from samples are owing to contamination. If more samples cannot be obtained, we still may be able to gain some information from the experiment.

More complicated experiments include the addition of artifacts with a similar history to the samples but not containing sample DNA. It should be possible to extend the methods described here to deal with these experiments. A more important case is replicate amplifications from the same extractions. To make the parameters of the model identifiable, we assumed that the efficiency

of amplification of one kind of extract is unaffected by the presence of the other kind of extract. If, instead, we amplify the same extractions several times, we might be able to estimate parameters without this assumption, with the use of a binomial mixture model (McLachlan and Peel 2000, p. 164). This model would help deal with the presence of PCR inhibitors and carrier effects, in which components of the sample extract reduce or increase the probability of amplifying contaminant DNA, respectively (e.g., Cooper 1994, pp. 154–156, 158). Our current model ignores these possibilities. Therefore, it will underestimate the probability that a positive result from a sample is authentic if the extract contains PCR inhibitors, and it will overestimate the probability that a positive result is authentic if there are carrier effects.

There are many other ways one could analyze the results of ancient-DNA experiments. For example, we treated sample and contaminant DNA as if they were either present or absent. If the results of an experiment are quantitative, it might be possible to model the concentrations of each kind of DNA. If contaminants are present at a much lower level than sample DNA, quantifying the yield from PCR may allow us to distinguish between amplifications of sample and contaminant DNA (Yang et al. 2003). We are not aware of any attempts at statistical analysis of this situation.

Another important factor that is not included in our model is that the likelihood of a result being authentic should depend on the identity of the sequence. For example, a human mitochondrial sequence found in one of the investigators is a more likely contaminant than a sequence not found in any of the investigators. Similarly, a common human sequence is a more likely contaminant than a rare sequence. Forensic studies use large databases of human DNA sequences to estimate the probabilities of events such as paternity (e.g., Rolf et al. 2001). Estimation of the probability that a sequence contaminates an ancient-DNA sample is more difficult, because we do not know how the frequency of a sequence in a database relates to the probability of that sequence appearing as a contaminant and because relevant databases are not available for most nonhuman species. Prior information from experiments performed under similar conditions could solve this problem; however, if contamination is a rare event, we would need a large amount of prior information to obtain accurate estimates. We could use this information in a Bayesian manner to integrate over the nuisance parameters θ, rather than setting them to their conditional maximum-likelihood estimates (Berger et al. 1999). This method can give a better indication of the uncertainty in parameters. Incorporating sequence identity information will be a valuable direction for further work.

In summary, we are not claiming that our method is the best possible way to analyze the results of ancient-DNA experiments. Nevertheless, we hope to draw attention to the need for statistical analysis and to suggest how it might be done.

Acknowledgments

This work was funded by the Arts and Humanities Research Board. We are grateful to Mim Bower, Peter Forster, Ellen Nisbet, and Ed Susko, for discussions, and to Christopher de Hamel and Caroline Checkley-Scott, for examining our parchment samples. Terry Brown, Alan Cooper, and an anonymous reviewer made helpful comments on the manuscript.

References

Agresti A (2002) Categorical data analysis. John Wiley & Sons, Hoboken, NJ

Austin JJ, Ross AJ, Smith AB, Fortey RA, Thomas RH (1997) Problems of reproducibility: does geologically ancient DNA survive in amber-preserved insects? Proc R Soc Lond B 264: 467–474

Bär W, Brinkmann B, Budowle B, Carracedo A, Gill P, Holland M, Lincoln PJ, Mayr W, Morling N, Olaisen B, Schneider PM, Tully G, Wilson M (2000) DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Int J Legal Med 113:193–196

Berger JO, Liseo B, Wolpert RL (1999) Integrated likelihood methods for eliminating nuisance parameters. Stat Sci 14: 1–28

Burger J, Pfeiffer I, Hummel S, Fuchs R, Brenig B, Herrmann B (2001) Mitochondrial and nuclear DNA from (pre)historic hide-derived material. Anc Biomol 3:227–238

Cooper A (1994) DNA from museum specimens. In: Herrmann B, Hummel S (eds) Ancient DNA. Springer-Verlag, New York, pp 149–165

Cooper A, Poinar HN (2000) Ancient DNA: do it right or not at all. Science 289:1139

Gilbert MTP, Hansen AJ, Willerslev E, Rudbeck L, Barnes I, Lynnerup N, Cooper A (2003) Characterization of genetic miscoding lesions caused by postmortem damage. Am J Hum Genet 72:48–61

Handt O, Richards M, Trommsdorff M, Kilger C, Simanainen J, Georgiev O, Bauer K, Stone A, Hedges R, Schaffner W, Utermann G, Sykes B, Pääbo S (1994) Molecular genetic analyses of the Tyrolean Ice Man. Science 264:1775–1778

Hilborn R, Mangel M (1997) The ecological detective. Princeton University Press, Princeton, NJ

Hummel S (2003) Ancient DNA typing: methods, strategies and applications. Springer-Verlag, Berlin

McCullagh P, Nelder JA (1989) Generalized linear models. Chapman and Hall, London

McLachlan G, Peel D (2000) Finite mixture models. John Wiley & Sons, New York

Mehta CR, Patel NR (1997) Exact inference for categorical data. Harvard University and Cytel Software Corporation, Cambridge, MA

Rolf B, Keil W, Brinkmann B, Roewer L, Fimmers R (2001)

Paternity testing using Y-STR haplotypes: assigning a probability for paternity in cases of mutations. Int J Legal Med 115:12–15

Stoneking M (1995) Ancient DNA: how do you know when you have it and what can you do with it? Am J Hum Genet 57:1259–1262

Venzon DJ, Moolgavkar SH (1988) A method for computing profile-likelihood-based confidence intervals. Appl Statist 37:87–94

Weiss G, von Haeseler A (1997) A coalescent approach to the polymerase chain reaction. Nucleic Acids Res 25:3082–3087

Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, Shapiro B, Bunce M, Wiuf C, Gilichinsky DA, Cooper A (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science 300:791–795

Willerslev E, Hansen AJ, Poinar HK (2004) Isolation of nucleic acids and cultures from fossil ice and permafrost. Trends Ecol Evol 19:141–147

Yang DY, Eng B, Saunders SR (2003) Hypersensitive PCR, ancient human mtDNA, and contamination. Hum Biol 75: 355–364