

How many clones need to be sequenced from a single forensic or ancient DNA sample in order to determine a reliable consensus sequence?

Mim A. Bower*, Matthew Spencer¹, Shuichi Matsumura, R. Ellen R. Nisbet² and Christopher J. Howe²

Archaeogenetics Laboratory, McDonald Institute for Archaeological Research, University of Cambridge, Downing Street, Cambridge CB2 3ER, UK, ¹Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, B3H 3J5, Canada and ²Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, UK

Received March 10, 2005; Revised April 7, 2005; Accepted April 15, 2005

ABSTRACT

Forensic and ancient DNA (aDNA) extracts are mixtures of endogenous aDNA, existing in more or less damaged state, and contaminant DNA. To obtain the true aDNA sequence, it is not sufficient to generate a single direct sequence of the mixture, even where the authentic aDNA is the most abundant (e.g. 25% or more) in the component mixture. Only bacterial cloning can elucidate the components of this mixture. We calculate the number of clones that need to be sampled (for various mixture ratios) in order to be confident (at various levels of confidence) to have identified the major component. We demonstrate that to be >95% confident of identifying the most abundant sequence present at 70% in the ancient sample, 20 clones must be sampled. We make recommendations and offer a free-access web-based program, which constructs the most reliable consensus sequence from the user's input clone sequences and analyses the confidence limits for each nucleotide position and for the whole consensus sequence. Accepted authentication methods must be employed in order to assess the authenticity and endogeneity of the resulting consensus sequences (e.g. quantification and replication by another laboratory, blind testing, amelogenin sex versus morphological sex, the effective use of controls, etc.) and determine whether they are indeed aDNA.

INTRODUCTION

Ancient DNA (aDNA) template is commonly a mixture of molecules comprising varying amounts of the correct endogenous sequence, damaged endogenous sequence, contaminant sequence and damaged contaminant sequence. Therefore, PCR products from aDNA template are also likely to be a mixture of misamplified damaged template and contaminant template. Although protein analysis, amino acid racemization and DNA quantification can act as a proxy for assessing DNA survival in the starting template, there are currently no effective methods for assessing the resulting template mixture prior to PCR.

Thus, the sequencing of individual PCR products, which have been ligated into a suitable vector and transformed into a bacterial host (i.e. bacterial cloning), has been put forward as an essential step by which we can identify the extent and components of aDNA template mixture (i.e. by sequencing many clones, each corresponding to a single molecule in the PCR product mixture). A consensus can then be constructed from the selected clone sequences once data from possible jumping PCR events have been identified and removed. But how many bacterial clones do we need to sequence for this consensus to be a reliable representation of the most abundant sequence?

The exact number of clones necessary to achieve a given level of confidence in the consensus depends on the frequency of each incorrect nucleotide at any given nucleotide position in the sequence. In this paper, we model ways to help researchers choose the number of clones they need to sequence to obtain the most representative consensus sequences (i.e. the most abundant sequence in the template mixture) in the presence of template mixture (i.e. the presence of contaminating DNA

*To whom correspondence should be addressed. Tel: +44 1223 339297; Fax: +44 1223 339285; Email: mab1004@cam.ac.uk

molecules from different organisms of the same species or damaged DNA molecules resulting from decay processes).

Additionally, we launch a web-based tool (the Consensus Confidence program, which can be found at <http://www.mcdonald.cam.ac.uk>) to help researchers assess confidence levels in their consensus sequences and decide how many additional clones, if any, they will need to sequence to reach acceptable confidence levels (e.g. 95% confidence that the consensus sequence they have constructed is the most abundant sequence present in their aDNA template mixture).

The Consensus Confidence program assesses the quality of the consensus sequence derived from a given PCR, it does not verify whether the consensus sequence is authentic aDNA and cannot replace the use of standard aDNA authentication methods. Accepted authentication methods must be employed in order to assess the authenticity and endogeneity of the resulting consensus sequences (e.g. quantification and replication by another laboratory, blind testing, amelogenin sex versus morphological sex, the effective use of controls, etc.) and determine whether they are indeed aDNA.

Sources and consequences of template mixture

There are several points at which template mixture can occur: during post-depositional decay processes (e.g. diagenesis of DNA molecular structure following the death and burial of the organism), excavation and post-excavation events (e.g. handling by archaeologists and museum staff) and more rarely, providing that standard aDNA laboratory protocols are followed, laboratory processing (e.g. extraction and amplification of DNA). Quantifying same-species contamination of fossil human remains is a colossal challenge (1), but even when working with non-human template, contamination from DNA of the same species cannot be ruled out, e.g. where research material is derived from museum collections or where reagents or plasticware are contaminated during manufacture (2). Pre-laboratory and laboratory contamination can be greatly reduced by following established protocols, i.e. effective and informative controls and contamination avoidance strategies (3). However, several authors have demonstrated that, in some cases, even the most stringent controls can fail to prevent or detect low-level contamination (1).

Cloning of PCR products is currently the only way of effectively elucidating the extent of contamination type mixture. However, even without taking contamination into account aDNA template is a mixture of individual DNA molecules, which contain differing levels of damage (e.g. depurination, deamination, strand breakage, cross-linking, etc.) according to their micro-taphonomic history. Some forms of DNA damage prevent PCR amplification altogether (e.g. strand breakage and cross-linking); however, where amplification of damaged DNA molecules is possible, it can result in misincorporation of nucleotides during PCR (4,5) and is probably the most common source of incorrect nucleotide in any given sequence. These misincorporations can be significant at any nucleotide position, and particularly so if they occur at nucleotide positions where phylogenetic variation is expected as this can affect subsequent analysis (6).

It has been argued that direct sequencing minimizes the chance of detecting sequences of minor contaminants or of amplification errors (7,8). This argument may apply in

particular circumstances where external evidence is sufficiently strong to validate the direct sequence (9). In less ideal circumstances, cloning is necessary (1,5,10–16). A consensus derived by direct sequencing could differ from the true aDNA sequence, in particular when the true sequence constitutes <50% of the mixture. Sequencing a number of clones, on the other hand, can reveal the most abundant nucleotide at a position so long as that nucleotide occurs in >25% of clones. Further standard authentication criteria must then be applied (17–20) to determine whether the most abundant nucleotide is plausibly the true aDNA. If it can be assumed that the most abundant nucleotide is also the true aDNA nucleotide (e.g. by fulfilling all other standard authentication criteria), then the true ancient nucleotide can be ascertained with a quantifiable degree of confidence.

METHODS

To calculate the number of cloned bacterial colonies to sequence, we take a random sample of n DNA molecules from a population of molecules with known proportions, where 1 nt is correct (with proportion P_C) and the other three represent contamination, damage or PCR errors (with proportions P_{I1} , P_{I2} and P_{I3}). The probability of drawing a sample of $S = (j, k, l, m)$ molecules from the population of molecules is the multinomial

$$P(S) = \frac{n!}{j! k! l! m!} P_C^j P_{I1}^k P_{I2}^l P_{I3}^m \quad 1$$

For a given sample, we define the consensus choice as the most abundant nucleotide in the sample, with ties broken at random. Breaking ties at random means that if there are two most abundant nucleotides in the sample, we will choose either one with equal probability. We assume that there is only one most abundant nucleotide in the population, which is reasonable if the population (e.g. the set of all molecules present after the final round of PCR) is large enough to treat proportions as continuous. We also define the representative consensus as the nucleotide that is most abundant in the population. Then the probability that we will choose the representative consensus nucleotide is

$$D(S) = \frac{\delta(j, \max(j, k, l, m))}{\sum_{s \in S} \delta(s, \max(j, k, l, m))}, \quad 2$$

where δ is the Kronecker delta function,

$$\delta(j, \max(j, k, l, m)) = \begin{cases} 1, & \text{if } j = \max(j, k, l, m), \\ 0, & \text{otherwise.} \end{cases}$$

Putting together Equations 1 and 2 and summing over all possible samples, the probability of getting the representative consensus is

$$P(\text{correct consensus}) = \sum_{j=0}^n \sum_{k=0}^{n-j} \sum_{l=0}^{n-j-k} P(S) D(S) \quad 3$$

with $m = n - j - k - l$. So long as the correct nucleotide is more frequent than any other (even if it constitutes <50%), the probability of obtaining a representative consensus approaches 1 as the number of molecules sampled becomes large.

When determining a representative consensus over an entire experiment, we do not distinguish between clones from single extractions and clones from multiple extractions. The latter design would perhaps be more appropriate, as repeated extractions and PCRs, which continuously give the same result, would additionally increase our confidence that a sequence is endogenous. However, this is impossible to model mathematically.

RESULTS

We used Equation 3 to calculate the probability of obtaining the representative consensus for two different mixture scenarios and a range of numbers of molecules sampled (i.e. clone colonies picked). Our results relate only to obtaining the most abundant sequence. We need other information (e.g. effectively applied extraction and PCR controls, phylogenetic plausibility, etc.) to decide whether this most abundant sequence is an authentic aDNA.

In the first scenario (Figure 1a), we assumed that the correct nucleotide at a site is mixed with a single other nucleotide. The probability of obtaining the representative consensus for a given proportion of the correct nucleotide in the template increases (or at least does not decrease) as the number of

clones increases, provided the correct nucleotide has a frequency $>50\%$. For example, if the correct nucleotide has a frequency of 70% and the incorrect nucleotide has a frequency of 30% , then the chance of obtaining the representative consensus is 90% with 10 clones and 97% with 20 clones.

If the correct nucleotide is mixed with three other nucleotides having equal frequencies (Figure 1b), the probability of obtaining the representative consensus is higher because each of the incorrect nucleotides is less common. If the correct nucleotide has a frequency of 70% and each incorrect nucleotide has a frequency of 10% , the probability of obtaining the representative consensus is 98% with 10 clones and $>99\%$ with 20 clones. In this case, the probability of obtaining the representative consensus increases (or at least does not decrease) with the number of clones as long as the correct nucleotide has a frequency $>25\%$. Thus, we may have a reasonable chance of obtaining the representative consensus even with high rates of contamination and damage. For example, if the correct nucleotide has a frequency of 40% and the three incorrect nucleotides each have a frequency of 20% , the chance of getting the representative consensus is 62% from 10 clones or 75% from 20 clones.

In principle, we could use Equation 3 to determine the probability that we have obtained the representative consensus

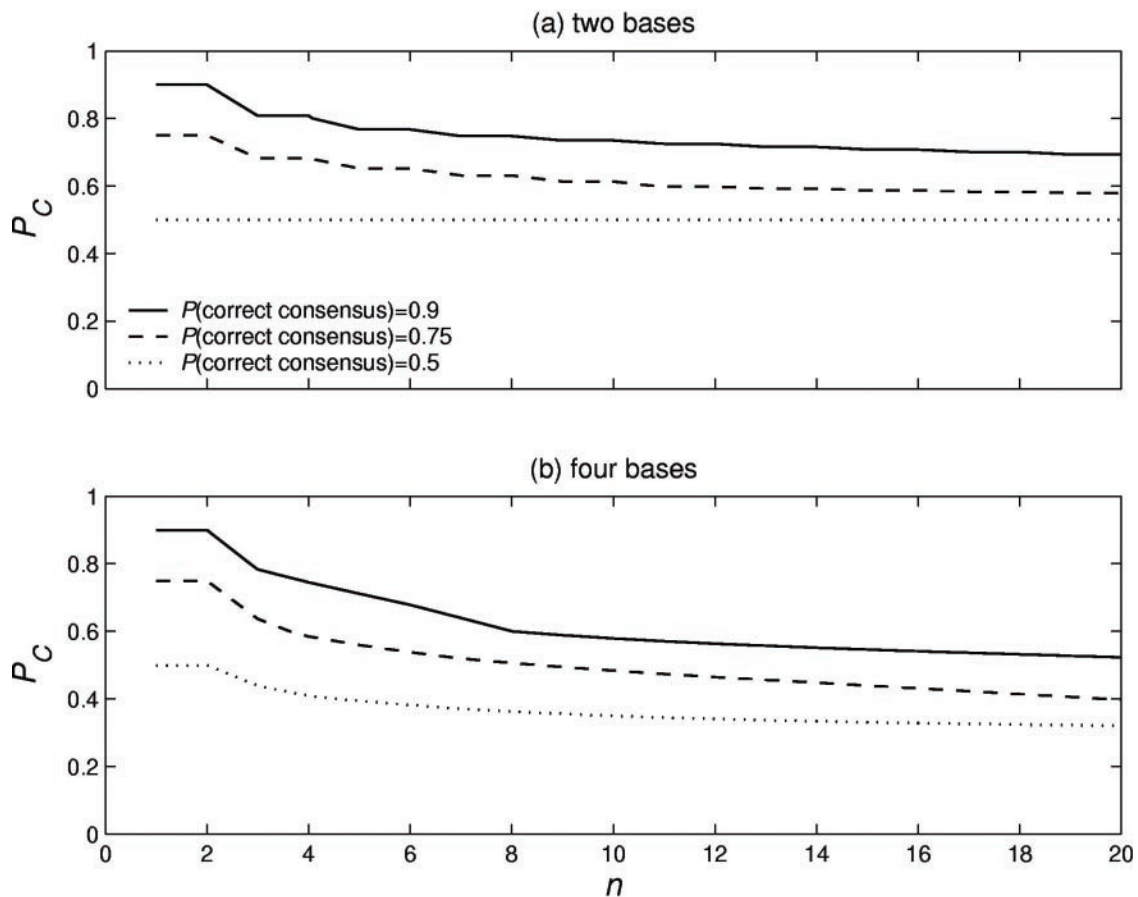


Figure 1. The probability of obtaining the representative consensus when the initial template contains: (a) a mixture of the correct nucleotide and one other; (b) a mixture of the correct nucleotide and three others, with the three incorrect nucleotides having equal frequency. P_C is the frequency of the correct nucleotide, and n is the number of molecules sampled to form the consensus. The lines are smoothed contours of the probability of obtaining the representative consensus being 90% (solid line), 75% (dashed line) or 50% (dotted line).

from a given sample. However, we will not know the probabilities P_C , P_{I1} , P_{I2} and P_{I3} exactly, so we would have to integrate over all possible values of these. A simpler approach is to use the approximate confidence intervals for multinomial proportions developed by Goodman (21). The 100 $(1 - \alpha)\%$ confidence interval for the i th kind of molecule out of k different kinds is

$$[B + 2n_i \pm (B(B + 4n_i(N - n_i)/N))^{1/2}]/[2(N + B)], \quad 4$$

where N is the number of molecules sampled, n_i is the number of these that were of the i th kind, and B is the upper 100 (α/k) th percentile of the χ^2 distribution with one degree of freedom. If we were looking at the distribution of nucleotides at a single site, k would be 4. The Goodman (21) confidence intervals are approximate, and may not be very reliable for small samples. Nevertheless, they are simple to calculate. Better but more complicated intervals exist (22). We conducted a number of Monte Carlo simulations and found that the Goodman intervals generally work well for large samples, but do not work well for fewer than 12 samples.

Goodman (21) also shows how to calculate approximate simultaneous confidence intervals for all pairwise differences between proportions. For the most abundant molecule in the sample, all these differences will be positive. If the 100 $(1 - \alpha)\%$ confidence intervals for all these differences do not include zero, then one or more of these differences will be negative (we will have wrongly identified the consensus) no more than $100\alpha\%$ of the time. The confidence intervals are

$$d_{ij} \pm [C(p_i + p_j - d_{ij}^2)/N]^{1/2}, \quad 5$$

where d_{ij} is $p_i - p_j$, p_i is the proportion of observations falling in category i , C is the upper 100 (α/K) th percentile of the χ^2 distribution with one degree of freedom, and $K = k(k - 1)/2$.

For example, if we take a sample of 10 clones and observe 7 molecules of type 1 and one each of three other types, then the 95% confidence intervals on the proportions are [0.3289, 0.9174] for the type of molecule we observed seven times, and [0.0124, 0.4949] for the three types we observed once. The lower 95% confidence limit for the difference between the proportions of the type of molecule we observed seven times and each of the other types of molecule is 0.0466. Thus, we would expect to identify the representative consensus $>95\%$ of the time. On the other hand, if we observed seven molecules of type 1, 3 of type 2, and none of types 3 or 4, the 95% confidence intervals on the proportions would be [0.3289, 0.9174], [0.0826, 0.6711], [0, 0.3842] and [0, 0.3842], respectively. The lower 95% confidence limits for the differences between the proportions of the type of molecule we observed eight times and the types of molecule we observed 2, 0 and 0 times are -0.0674 , 0.4663 and 0.4663, respectively. We would expect to identify the representative consensus $<95\%$ of the time. If we observed the same proportions, but doubled the sample size (giving 16, 4, 0 and 0 observations), the lower 95% confidence limits for the differences in proportions between the most abundant type of molecule and each other type would be 0.1281, 0.5640 and 0.5640. Thus, either a larger sample or a less even distribution of molecules among types increases our confidence that the consensus is really the most abundant type of molecule.

If we take a single site, we will have wrongly identified the consensus $<5\%$ of the time if all the three lower 95% confidence limits do not include zero. However, when we have 40 nt positions in a sequence, we would expect to have two wrong consensus positions on average, even if all the lower 95% confidence limits are positive. A simple way to avoid this problem is to use a more conservative significance level, α/s (s : sequence length), for each nucleotide position. This gives us a conservative estimate for the probability that we have identified the representative sequence as a whole. We may need slightly more clones to have a significant result at the sequence level. For example, if we had 20 clones of a sequence with 40 nt and observed the same proportions as above (i.e. 16, 4, 0 and 0), one of the lower 95% confidence limits would be negative. However, if we had 25 clones (20, 5, 0 and 0), all the 95% confidence limits would be positive.

RECOMMENDATIONS: A FIVE STEP STRATEGY

Sequence at least 12 clones from each PCR amplification

The number of clones the researcher chooses to sequence is a trade-off between expense and the minimum number required for appropriate statistical treatment. Based on our simulations using the Goodman equation, we recommend that in the first instance a minimum number of 12 clones are sequenced to create a consensus sequence for each PCR product of interest.

Cloning is the only way to elucidate the extent of template mixture as each cloned bacterial colony represents the sequence of a single PCR amplicon. Comparing the sequences of a given number of single-amplicon colonies and forming a consensus sequence from them is a good way to get an accurate consensus sequence where the template is mixed (sequencing each clone more than once would be slightly more accurate). Cloning has some disadvantages. Besides being time-consuming and expensive, each clone replicates a single DNA molecule and thus only a small number of molecules are sampled (20 colonies picked = 20 template molecules or fewer). Therefore, it is essential to sample and sequence a sufficient number of cloned colonies to have a good chance of obtaining the representative consensus. Taking a consensus of too few clones has the potential for higher error rates than simply direct sequencing [e.g. some researchers sequence only two clones (23)].

We can put a confidence interval on the probability that our consensus choice is an accurate reflection of the most abundant sequence present in the template mixture. This confidence level decreases as the relative frequency of the most abundant nucleotide decreases, i.e. the greater the mixture, the lower the probability of discovering the most abundant sequence. The number of clones needed to attain a given level of confidence in the consensus depends on the extent of the template mixture, which is unknown until sequencing of clones has begun.

Enter the clone sequence data in the Consensus Confidence program and calculate the confidence levels for the consensus sequence

The percentage confidence levels of the consensus sequence from the selected clones can be simply and swiftly calculated using our Consensus Confidence program, which can be found at <http://www.mcdonald.cam.ac.uk> (Figure 2). Sequence data

Please enter your clone sequences into the box and press OK.

Click here for the Help Files

----- Consensus Confidence Program (ver. 1.00) -----

Unfortunately, the probability that you have identified the correct consensus is below 95% at some sites.

Number of clones: 12; Total number of sites: 50

check this box if you have deletions/insertions

Data Input (Clone Sequences)

S001:AGCTGCGCTANGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC (example)

S001:AGCTGCGCTANGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC
 S002:AGTTGCGCTATGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC
 S003:AGCTGCGCTATGTGTAGCTTTGATCATGCGCATGTGTAGATCCGTAATAC
 S004:AGCTGCGCTATGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC
 S005:AGCTGCGCTATGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC
 S006:AGCTGCGCTATGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC
 S007:AGCTGCGCTATGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC
 S008:AGCTGCGCTATGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC
 S009:AGTTGCGCTATGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC
 S010:AGCTGCGCTATGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC

Consensus

10 20 30 40 50 60 70 80 90 100

AGCTGCGCTATGTGTAGCTTTGATCAAGCGCATGTGTAGATCCGTAATAC

70%< *****
 90%< *****
 95%< *****

95%< ** ***** (sequence level significance)

Details:

Site	Frequency					95% Confidence Intervals for Proportions				Cons.	Lower 95% Confidence Limit for Pair-Wise Differences			
	A	G	C	T	N	---A---	---G---	---C---	---T---					
1:	12	0	0	0	0	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	A	{A-G 1.0*}	{A-C 1.0*}	{A-T 1.0*}	
2:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
3:	0	0	10	2	0	[0.00,0.34]	[0.00,0.34]	[0.47,0.96]	[0.03,0.52]	C	{C-A 0.54}	{C-G 0.54}	{C-T 0.09}	
4:	0	0	1	11	0	[0.00,0.34]	[0.00,0.34]	[0.01,0.44]	[0.55,0.98]	T	{T-A 0.70}	{T-G 0.70}	{T-C 0.41}	
5:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
6:	0	0	12	0	0	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	C	{C-A 1.0*}	{C-G 1.0*}	{C-T 1.0*}	
7:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
8:	0	0	12	0	0	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	C	{C-A 1.0*}	{C-G 1.0*}	{C-T 1.0*}	
9:	0	0	0	12	0	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
10:	12	0	0	0	0	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	A	{A-G 1.0*}	{A-C 1.0*}	{A-T 1.0*}	
11:	0	0	0	11	1	[0.00,0.36]	[0.00,0.36]	[0.00,0.36]	[0.63,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
12:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
13:	0	0	0	12	0	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
14:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
15:	0	0	0	12	0	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
16:	9	3	0	0	0	[0.39,0.93]	[0.06,0.60]	[0.00,0.34]	[0.00,0.34]	A	{A-G -0.15}	{A-C 0.42}	{A-T 0.42}	
17:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
18:	0	0	12	0	0	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	C	{C-A 1.0*}	{C-G 1.0*}	{C-T 1.0*}	
19:	0	0	0	12	0	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
20:	0	0	0	12	0	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
21:	0	0	0	12	0	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
22:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
23:	12	0	0	0	0	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	A	{A-G 1.0*}	{A-C 1.0*}	{A-T 1.0*}	
24:	0	0	0	12	0	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
25:	0	0	12	0	0	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	C	{C-A 1.0*}	{C-G 1.0*}	{C-T 1.0*}	
26:	10	2	0	0	0	[0.47,0.96]	[0.03,0.52]	[0.00,0.34]	[0.00,0.34]	A	{A-G 0.09}	{A-C 0.54}	{A-T 0.54}	
27:	8	0	3	1	0	[0.32,0.89]	[0.00,0.34]	[0.06,0.60]	[0.01,0.44]	A	{A-G 0.30}	{A-C -0.23}	{A-T 0.09}	
28:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
29:	0	0	12	0	0	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	C	{C-A 1.0*}	{C-G 1.0*}	{C-T 1.0*}	
30:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
31:	0	0	12	0	0	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	C	{C-A 1.0*}	{C-G 1.0*}	{C-T 1.0*}	
32:	12	0	0	0	0	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	A	{A-G 1.0*}	{A-C 1.0*}	{A-T 1.0*}	
33:	0	0	0	12	0	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
34:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
35:	0	0	0	12	0	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
36:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	
37:	0	0	0	12	0	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	[0.65,1.00]	T	{T-A 1.0*}	{T-G 1.0*}	{T-C 1.0*}	
38:	12	0	0	0	0	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	[0.00,0.34]	A	{A-G 1.0*}	{A-C 1.0*}	{A-T 1.0*}	
39:	0	12	0	0	0	[0.00,0.34]	[0.65,1.00]	[0.00,0.34]	[0.00,0.34]	G	{G-A 1.0*}	{G-C 1.0*}	{G-T 1.0*}	

Figure 2. The Consensus Confidence program (<http://www.mcdonald.cam.ac.uk>) showing three dialogue boxes: the data input box, the results box and the details box. To use, cut and paste pre-aligned sequence data with a five character sequence identifier from a text file or Word document and click OK to run the calculation (based on Equations 4 and 5). The results (a consensus sequence, confidence levels for each nucleotide and a whole sequence confidence level) can be copied back into a text file or Word document if required. For full details on how to use the Consensus Confidence program, please consult the help files, which can be found on the webpage.

must be entered pre-aligned and can be easily cut and pasted from a text or word file (the program allows a five character sequence identifier of your choice). The program requires sequences from at least 10 clones to function accurately for the reasons stated above, and the system will easily analyse up to 100 sequences, of a length up to 800 nt. For full details of how to use the Consensus Confidence program, please refer to the help files.

The Consensus Confidence program constructs a consensus sequence based on the input clone sequences and calculates a percentage probability that each individual nucleotide position occurs statistically most frequently with a confidence level between 70% and 95% using Equations 4 and 5 above. Additionally, the program estimates the probability that we have identified the representative sequence as a whole by calculating the 100 (1 - α/s)% lower confidence limits for the difference (s : sequence length, $\alpha = 0.05$) for each nucleotide position.

Figure 3a shows the Consensus Confidence program's calculation of the confidence levels for the consensus sequence derived from 20 clones from two PCRs of the Neanderthal type specimen [data from Figure 1, (24)]. As can be seen, 2 nt positions have <70% confidence levels [positions 00132 and 00141 of the Hyper-variable region II of the mitochondrial D-loop, numbering according to Anderson *et al.* (25)]. In this case, more clone sequences need to be added to resolve the ambiguous nucleotide positions (Figure 3b).

In ambiguous situations how many additional clones should be sequenced in total? Statistically there is no upper limit but practicality and experience suggests a limit of 30 clones. If 30 clones are insufficient to resolve a particular nucleotide position at >95% confidence, then the most abundant nucleotide at that position should be accepted, but highlighted in any publication as having a low confidence level, with the alternative

nucleotide(s) and confidence level(s) published alongside. If the position of the doubtful nucleotide falls at a point mutation that is phylogenetically significant, this consensus sequence should not be included in any phylogenetic analysis, as any inference will be based on weak data. In Figure 3b, 10 additional clone sequences have been added to the confidence level calculation for the region encompassing the ambiguous nucleotide positions. The results show that the ambiguous nucleotide positions have been resolved to a 95% confidence level, although the sequence level significance is still <95% at one position [position 00141 (25), where in the 30 clones input there are 23 Cs and 7 Ts at that position].

Although we would expect to have the highest confidence levels in the consensus sequences we publish (e.g. 95% or more), if the result is of particular importance or interest it may be appropriate to publish it, even if the confidence levels are lower (e.g. 80% or less) as long as this confidence level is reached by an appropriate number of clones (say, at least 30 per PCR) and that the result and any interpretation arising from it is tempered with due caution.

Apply the standard range of external criteria for validating the nature of your consensus sequence

Comparing cloned colonies from a single PCR is clearly insufficient to authenticate the endogeneity of the aDNA template of a given sample. The Consensus Confidence program will only allow the researcher to assess the confidence levels of the consensus sequence for each given PCR. Therefore, having identified the most abundant amplicon in a given PCR in step 2, the user then needs to determine whether its sequence is truly an aDNA, a modern contaminant, or the product of misincorporated nucleotides resulting from severely damaged template.

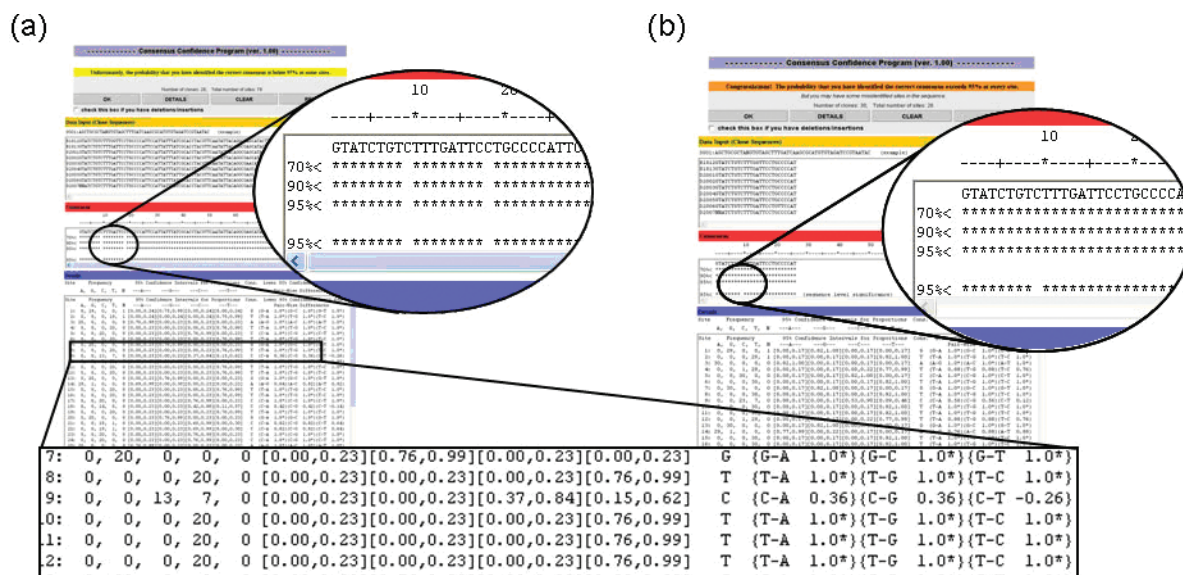


Figure 3. We input published clone sequences for the HVII region of the Neanderthal type specimen [Figure 1 (24)] into the Consensus Confidence program [clones B18.1–B18.13 and D20.1–D20.7; a total of 20 clone sequences, covering nucleotide positions 00124–00201, numbering after Anderson *et al.* (25)]. (a) The results show 2 nt positions where the confidence level of that particular nucleotide representing the most abundant one in the PCR was <95% [nucleotide positions 00132 and 00141 (25), here shown as position 9 and 18]. (b) We added the additional clone sequences which overlap the ambiguous (low confidence level) nucleotide positions so that the total number of clones was 30 [C19.1–C19.2, nucleotide positions 00124–00149 (25)]. The results show that the confidence levels for the ambiguous nucleotide positions are now at 95%, although the sequence level confidence is still <95%.

A wide range of external authentication criteria must be applied to make the case that the consensus sequence gained is indeed more likely than not to be the authentic endogenous DNA of the ancient sample in question. We refer the reader to a comprehensive list of protocols and recommendations for the authentication of aDNA available in the literature: the effective use of negative controls (20), quantification, replication of results, both by the primary researcher and by an external laboratory (18) blind testing (17), amelogenin sex versus morphological sex (19). No doubt, advances in aDNA research and technology will uncover more efficient and more powerful endogenous DNA validation tools in the future.

Report the number of clones and the confidence levels of your published consensus sequence

Our recommendation to include the individual clone sequences and their confidence levels in a publication should be self-evident. However, in reviewing the literature, we sometimes found it difficult to ascertain the number of clones that made up the published consensus sequences and what the extent of template mixture was (26,27). Given that we have argued that template mixture is almost unavoidable, greater transparency would be welcome if we are to judge each other's work effectively (28).

Report all the factors that might contribute to or minimize the probability of template mixture occurring

All the factors that might contribute to or minimize the probability of template mixture occurring (e.g. probable extent of contamination and use of cloning to identify template mixture) should be recorded. For example, Willerslev and co-workers (29) outline very carefully the care they took to understand the full extent of the template mixture from which they derived their final sequences. This is particularly important where aDNA is being used to address large and significant research questions or where the research claims are potentially controversial (such as the relationship between Neanderthals and modern humans, or the analysis of forensic specimens).

CONCLUSION

It is probably not technically feasible to guarantee that the aDNA sequences we publish are completely error free and this has to be understood and accepted (30). Therefore, we should consider that our results come with given levels of confidence derived from a range of factors, from the taphonomy and preservation history of the artefact and the micro-taphonomy of each extracted aDNA molecule, to excavation and post-excavation handling and laboratory protocols. Many of these confidence levels are impossible to quantify accurately. However, the simulations we have carried out and the Consensus Confidence program presented here allow the confidence level of the final consensus sequence to be accurately assessed.

We stress again that the Consensus Confidence program assesses the quality of the consensus sequence derived from a given PCR, it does not verify whether that consensus sequence is an authentic aDNA and cannot replace the use of standard aDNA authentication methods. It should be used in conjunction with the accepted authentication methods, which must be employed if the authenticity and endogeneity of the

resulting consensus sequences is to be argued. By using a wide array of authentication methods, the aDNA researcher must build a case that allows them to argue that the aDNA sequences published are more likely to be authentic and endogenous than not. Unfortunately, we will never be able to state unequivocally that our sequences are 100% guaranteed to be authentic, and we should not try to do this. However, one should use all appropriate methods to increase the probability that the DNA is an authentic aDNA. Our methods allow one to state that the consensus sequence is the representative of the total amplicon population in a given PCR, and is thus, however valuable, only one small step in the authentication process.

ACKNOWLEDGEMENTS

The authors are very grateful to Peter Forster for comments on earlier drafts, Martin Jones and Tibor Kálmár for advice and discussions and Dan and Zoë Leighton for technical support. M.A.B. was supported by the McDonald Institute for Archaeological Research, University of Cambridge, M.S. by the Arts and Humanities Research Board, S.M. by the Sloan Foundation and RERN by the Biotechnology and Biological Sciences Research Council. Funding to pay the Open Access publication charges for this article was provided by the JISC.

Conflict of interest statement. None declared.

REFERENCES

- Kolman, C.J. and Tuross, N. (2000) Ancient DNA analysis of human populations. *Am. J. Phys. Anthropol.*, **111**, 5–23.
- Schmidt, T., Hummel, S. and Herrmann, B. (1995) Evidence of contamination in PCR laboratory disposables. *Naturwissenschaften*, **82**, 423–431.
- Willerslev, E., Hansen, A.J., Binladen, J., Brand, T.B., Gilbert, M.T.P., Shapiro, B., Bunce, M., Wiuf, C., Gilchinsky, D.A. and Cooper, A. (2003) Diverse plant and animal genetic records from holocene and pleistocene sediments. *Science*, **300**, 791–795.
- Pääbo, S., Irwin, D.M. and Wilson, A.C. (1990) DNA damage promotes jumping between templates during enzymatic amplification. *J. Biol. Chem.*, **265**, 4718–4721.
- Goloubinoff, P., Pääbo, S. and Wilson, A.C. (1993) Evolution of maize inferred from sequence diversity of an Adh2 gene segment from archaeological specimens. *Proc. Natl Acad. Sci. USA*, **90**, 1997–2001.
- Gilbert, M.T.P., Willerslev, E., Hansen, A.J., Barnes, I., Rudbeck, L., Lynnerup, N. and Cooper, A. (2003) Distribution patterns of postmortem damage in human mitochondrial DNA. *Am. J. Hum. Genet.*, **72**, 32–47.
- Adcock, G.J., Dennis, E.S., Easteal, S., Huttley, G.A., Jermain, L.S., Peacock, W.J. and Thorne, A. (2001) Mitochondrial DNA sequences in ancient Australians: implications for modern human origins. *Proc. Natl Acad. Sci. USA*, **98**, 537–542.
- Keyser-Tracqui, C., Crubezy, E. and Ludes, B. (2003) Nuclear and mitochondrial DNA analysis of a 2000-year-old necropolis in the Egyin Gol Valley of Mongolia. *Am. J. Hum. Genet.*, **73**, 247–260.
- Anslinger, K., Weichhold, G., Keil, W., Bayer, B. and Eisenmenger, W. (2001) Identification of the skeletal remains of Martin Bormann by mtDNA analysis. *Int. J. Legal Med.*, **114**, 194–196.
- Handt, O., Hoss, M., Krings, M. and Pääbo, S. (1994a) Ancient DNA—methodological challenges. *Experientia*, **50**, 524–529.
- Handt, O., Krings, M., Ward, R.H. and Pääbo, S. (1996) The retrieval of ancient human DNA sequences. *Am. J. Hum. Genet.*, **59**, 368–376.
- Krings, M., Stone, A., Schmitz, R.W., Krainitzki, H., Stoneking, M. and Pääbo, S. (1997) Neanderthal DNA sequences and the origin of modern humans. *Cell*, **90**, 19–30.
- Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. and Pääbo, S. (2001b) DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.*, **29**, 4793–4799.

14. Barnes,I., Matheus,P., Shapiro,B., Jensen,D. and Cooper,A. (2002) Dynamics of pleistocene population extinctions in Beringian brown bears. *Science*, **295**, 2267–2270.
15. Gilbert,M.T.P., Hansen,A.J., Willerslev,E., Rudbeck,L., Barnes,I., Lynnerup,N. and Cooper,A. (2003) Characterization of genetic miscoding lesions caused by postmortem damage. *Am. J. Hum. Genet.*, **72**, 48–61.
16. Lalueza-Fox,C., Gilbert,M.T.P., Martinez-Fuentes,A.J., Calafell,F. and Bertranpetit,J. (2003) Mitochondrial DNA from pre-Columbian Ciboneys from Cuba and the prehistoric colonization of the Caribbean. *Am. J. Phys. Anthropol.*, **121**, 97–108.
17. Yang,H., Golenberg,E. and Shoshani,J. (1997) A blind testing design for authenticating ancient DNA sequences. *Mol. Phylogenet. Evol.*, **7**, 261–265.
18. Cooper,A. and Poinar,H.N. (2000) Ancient DNA: do it right or not at all. *Science*, **289**, 1139.
19. Clisson,I., Keyser,C., Francfort,H.P., Crubezy,E., Samashev,Z. and Ludes,B. (2002) Genetic analysis of human remains from a double inhumation in a frozen kurgan in Kazakhstan (Berel site, Early 3rd Century BC). *Int. J. Legal Med.*, **116**, 304–308.
20. Spencer,M. and Howe,C. (2004) Authenticity of ancient DNA results: a statistical approach. *Am. J. Hum. Genet.*, **75**, 240–250.
21. Goodman,L.A. (1965) On simultaneous confidence intervals for multinomial proportions. *Technometrics*, **7**, 247–254.
22. Hou,C.D., Chiang,J.T. and Tai,J.J. (2003) A family of simultaneous confidence intervals for multinomial proportions. *Comput. Statist. and Data Anal.*, **43**, 29–45.
23. Vernesi,C., Caramelli,D., Dupanloup,I., Bertorelle,G., Lari,M., Cappellini,E., Moggi-Cecchi,J., Chiarelli,B., Castri,L., Casoli,A. *et al.* (2004) The Etruscans: a population-genetic study. *Am. J. Hum. Genet.*, **74**, 694–704.
24. Krings,M., Geisert,H., Schmitz,R.W., Krainitzki,H. and Pääbo,S. (1999) DNA sequence of the mitochondrial hypervariable region II from the Neanderthal type specimen. *Proc. Natl Acad. Sci. USA*, **96**, 5581–5585.
25. Anderson,S., Bankier,A., Barrell,B., de Bruijn,M., Coulson,A., Drouin,J., Eperon,I., Nierlich,D., Roe,B., Sanger,F. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457–465.
26. Jaenicke-Després,V., Buckler,E.S., Smith,B.D., Gilbert,M.T.P., Cooper,A., Doebley,J. and Pääbo,S. (2003) Early allelic selection in maize as revealed by ancient DNA. *Science*, **302**, 1206–1208.
27. Montiel,R., Garcia,C., Canadas,M.P., Isidro,A., Guijo,J.M. and Malgosa,A. (2003) DNA sequences of *Mycobacterium leprae* recovered from ancient bones. *FEMS Microbiol. Lett.*, **226**, 413–414.
28. Poinar,H., Kuch,M., McDonald,G., Martin,P. and Pääbo,S. (2003) Nuclear gene sequences from a late pleistocene sloth coprolite. *Curr. Biol.*, **12**, 1150–1152.
29. Willerslev,E., Hansen,A.J., Christensen,B., Steffensen,J.P. and Arctander,P. (1999) Diversity of holocene life forms in fossil glacier ice. *Proc. Natl Acad. Sci. USA*, **96**, 8017–8021.
30. Rohl,A., Brinkmann,B., Forster,L. and Forster,P. (2001) An annotated mtDNA database. *Int. J. Legal Med.*, **115**, 29–39.