Title

Revisiting the Briggs Ancient DNA Damage Model: A Fast Maximum Likelihood Method to Estimate Post-Mortem Damage.

Journal
Molecular Ecology Resources, 25(1)

Authors
Zhao, Lei
Henriksen, Rasmus
Ramsøe, Abigail
et al.

Publication Date
2025

DOI
10.1111/1755-0998.14029

Peer reviewed

Molecular Ecology Resources

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

Revisiting the Briggs Ancient DNA Damage Model: A Fast Maximum Likelihood Method to Estimate Post-Mortem Damage

Lei Zhao[1,2] | Rasmus Amund Henriksen[2] | Abigail Ramsøe[2] | Rasmus Nielsen[2,3] | Thorfinn Sand Korneliussen[2]

[1]School of Ecological and Environmental Sciences, East China Normal University, Shanghai, China | [2]Section for GeoGenetics, Globe Institute, University of Copenhagen, Copenhagen K, Denmark | [3]Department of Integrative Biology and Department of Statistics, University of California, Berkeley, California, USA

Correspondence: Lei Zhao (lzhao@des.ecnu.edu.cn) | Thorfinn Sand Korneliussen (tskorneliussen@sund.ku.dk)

Received: 21 November 2023 | **Revised:** 13 September 2024 | **Accepted:** 20 September 2024

Handling Editor: Frederic Austerlitz

Funding: This study was supported by the Carlsberg Foundation (CF19-0712 and CF20-0071) and the Lundbeck Foundation Centre for Disease Evolution (R302-2018-2155).

Keywords: ancient DNA | DNA deamination | high-throughput data | maximum likelihood estimation | posterior ancient probability | statistical inference

ABSTRACT

One essential initial step in the analysis of ancient DNA is to authenticate that the DNA sequencing reads are actually from ancient DNA. This is done by assessing if the reads exhibit typical characteristics of post-mortem damage (PMD), including cytosine deamination and nicks. We present a novel statistical method implemented in a fast multithreaded programme, ngsBriggs that enables rapid quantification of PMD by estimation of the Briggs ancient damage model parameters (Briggs parameters). Using a multinomial model with maximum likelihood fit, ngsBriggs accurately estimates the parameters of the Briggs model, quantifying the PMD signal from single and double-stranded DNA regions. We extend the original Briggs model to capture PMD signals for contemporary sequencing platforms and show that ngsBriggs accurately estimates the Briggs parameters across a variety of contamination levels. Classification of reads into ancient or modern reads, for the purpose of decontamination, is significantly more accurate using ngsBriggs than using other methods available. Furthermore, ngsBriggs is substantially faster than other state-of-the-art methods. ngsBriggs offers a practical and accurate method for researchers seeking to authenticate ancient DNA and improve the quality of their data.

1 | Introduction

Ancient DNA (aDNA) refers to the preserved genetic material of ancient organisms. Analysing aDNA has proven to be an essential mean for researchers to study the past. It has, for example, aided a deeper understanding of ancestral population history (Allentoft, Sikora, Fischer, et al. 2024; Allentoft, Sikora, Refoyo-Martínez, et al. 2024) and the dynamics of ancient ecosystems (Kjær et al. 2022; Fernandez-Guerra et al. 2023).

During the last few decades, the field of aDNA has seen an increase in both the quality and quantity of data, with the number of published ancient genomes surpassing 10,000 at the end of 2022 (Mallick et al. 2023).

Lei Zhao and Rasmus Amund Henriksen are first authors.

The primary structure of DNA consists of a linear sequence of nucleotides (A, T, C or G) connected together by a phosphate backbone. DNA is double-stranded—the two strands are connected by hydrogen bonds between the nucleotides. Due to the passage of time and prolonged exposure to various environmental conditions, DNA undergoes several conformational alterations, known as post-mortem damage (PMD). PMD in the double-stranded DNA molecule manifests mainly as nicks and deamination (Willerslev and Cooper 2005; Dabney, Meyer, and Pääbo 2013). A nick is a discontinuity in the backbone of either strand in a fragment. This can cause structural instability, which mediates a complete break where two smaller subfragments are formed (Willerslev and Cooper 2005). Nicks and subsequent breaks might explain why single-stranded regions at the termini of the fragment exist. A single-stranded region of the ancient molecule is termed an overhang, which can be defined as either the 5′ or 3′ overhang. Another hallmark of PMD is deamination of cytosine, which converts cytosine to uracil (Dabney, Meyer, and Pääbo 2013), with an increased frequency in the single-stranded part of a fragment compared to the double-stranded region. During the library preparation and PCR, uracil will be treated as thymine creating apparent C→T substitutions, a C→U change by deamination could also have a compounding effect, whereby blunt-end repair during library preparation leads to a complement G→A substitution, which manifest in 5′ overhangs near the focal (specific DNA segment in focus) fragment 3′ end (Briggs et al. 2007; Meyer and Kircher 2010). This is visualised simplistically in the first panel of Figure 1 representing the deamination directly as the C→T substitution, whereas the specific influence of each laboratory step and PCR amplification on the deamination pattern is visualised in greater detail in Figure S1. Lastly, these PMD characteristics make the truly ancient DNA distinguishable from its modern counterpart, supporting the separation of DNA from the (post-)depositional environment from endogenous DNA and allowing researchers to conduct their genetic analyses only on genuinely ancient materials (Willerslev and Cooper 2005; Dabney, Meyer, and Pääbo 2013). The process of obtaining DNA sequences from biological material follows a laboratory protocol that can be divided into (1) DNA extraction and purification, (2) library preparation and (3) DNA sequencing. The first step is sample-specific, whereas the second is sequencing platform-specific. The most common sequencing approach is the sequencing-by-synthesis, which requires that each original DNA fragment is ligated with known adapter sequences. Although several library preparations exist, historically the most common in aDNA studies, includes the following steps (Briggs et al. 2007; Meyer and Kircher 2010):

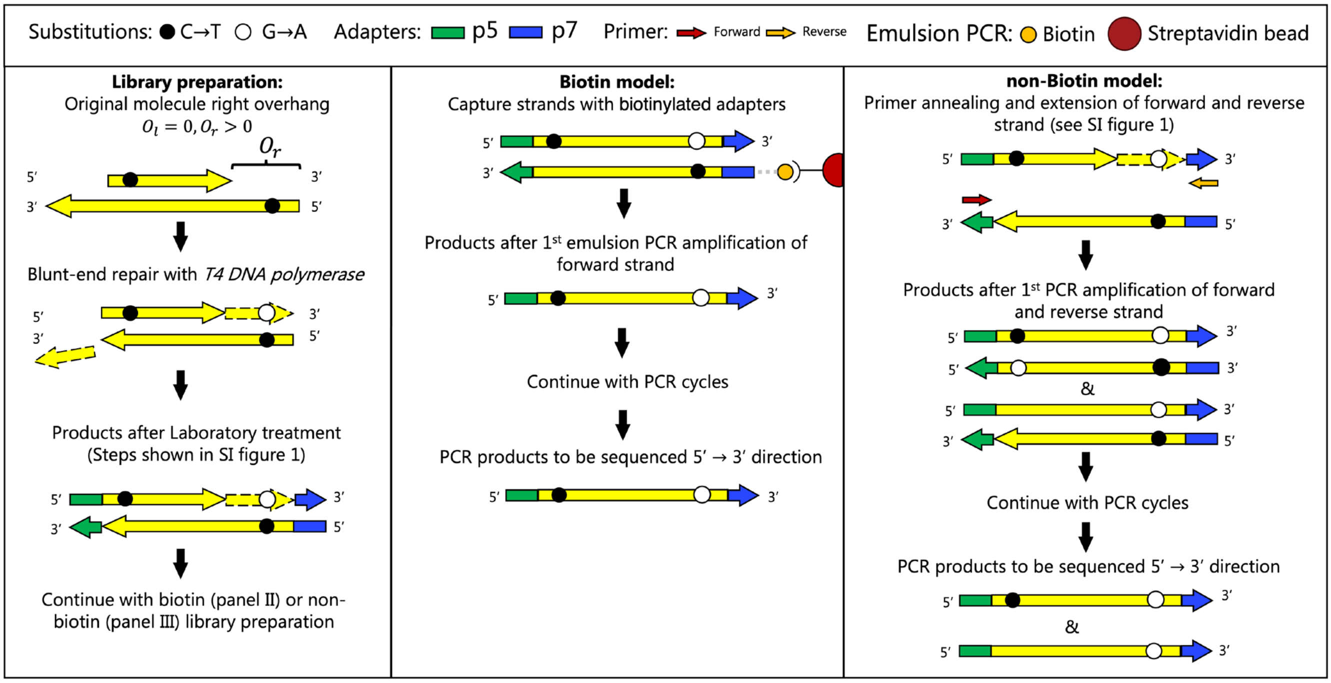

Substitutions: ● C→T ○ G→A
Adapters: p5 p7
Primer: Forward Reverse
Emulsion PCR: Biotin Streptavidin bead
Library preparation:
Original molecule right overhang
$O_l = 0, O_r > 0$
O_r
5′
3′
Blunt-end repair with *T4 DNA polymerase*
Products after Laboratory treatment
(Steps shown in SI figure 1)
Continue with biotin (panel II) or non-biotin (panel III) library preparation
Biotin model:
Capture strands with biotinylated adapters
Products after 1st emulsion PCR amplification of forward strand
Continue with PCR cycles
PCR products to be sequenced 5′ → 3′ direction
non-Biotin model:
Primer annealing and extension of forward and reverse strand (see SI figure 1)
Products after 1st PCR amplification of forward and reverse strand
&
Continue with PCR cycles
PCR products to be sequenced 5′ → 3′ direction

FIGURE 1 | Illustrative representation of an aDNA fragment with 5′ right overhang ($O_l = 0$, $O_r > 0$) prepared with two different laboratory protocols and the resulting unique deamination signal. The left panel (Library preparation) shows the original DNA and the double-stranded products following Steps 1–3 in the library preparation: Blunt-end Repair, Adapter Ligation and Adapter Fill-in, with each step shown in greater detail in Figure S1. During these steps, the 3′ overhang is removed during the blunt-end repairs ($O_l = 0$, left overhang), while the complementary substitution is observed in the 5′ overhang ($O_r > 0$, right overhang) before the sole molecules with a p5 and p7 ligated adapter are kept. The middle panel (biotin model) shows Step 4: PCR and denaturing when using the emulsion PCR amplification as done in 454 Roche sequencing (Briggs et al. 2007), capturing one strand with streptavidin beads using 5′ p7 biotinylation modifications. The PCR products and their deamination signal from the middle panel can be modelled using the ngsBriggs biotin model. The right panel shows Step 4 using the current Illumina PCR amplification approach (Meyer and Kircher 2010), amplifying both strands and sequencing the 5′ p5→p7 3′ (the first two PCR cycles are depicted in detail in Figure S1). The right panel can be modelled using ngsBriggs non-biotin. The PCR products of the biotin model are a subset of the PCR product obtained from the non-biotin model. In the example in this figure, the DNA fragment solely containing G→A near the 3′ end remains unique to non-biotin, which illustrates the different products following amplification, with multiple scenarios depicted in Data S1 (Supporting Information).

1. Blunt-end Repair: Removal of the 3′ overhang of the molecule through 3′→5′ exonuclease activity of *T4 DNA Polymerase* while also catalysing the 5′ overhangs fill-in synthesising the DNA in 5′→3′ direction (See part I of Figure 1; Tabor et al. 1997; Sambrook, Fritsch, and Maniatis 1989). With the subsequent 5′ phosphorylation of *T4 Polynucleotide Kinase* required for adapter ligation.
2. Adapter Ligation: Random ligation of the adapters (p5 or p7) with *T4 DNA Ligase* catalysing the phosphodiester bond with the blunt end. Those fragments ligated solely with the p5 or p7 adapter are non-functional, with only fragments ligated with both p5 and p7 adapters contributing to later processes.
3. Adapter Fill-in: Followed by the *Bst DNA Polymerase, Large Fragment* enzyme with strand displacement activity extending the nick present on one strand between the adapter and template (adapter fill-in; not visualised in left side of Figure 1, see Figure S1). This strand displacement activity will in the presence of single-stranded nicks in double-stranded inserts perform a similar downstream displacement during the DNA synthesis (nick-fill, as visualised with nicks in both strands Figure S3; nicks in a single strand Figure S4).
4. PCR and denaturing: The resulting double-stranded DNA is denatured, and both strands may act as the templates of PCR following different PCR protocols (e.g., emulsion PCR and Illumina sequencing) described below.

1.1 | Output PCR Templates

In the initial aDNA studies, the 454 sequencing protocol was commonly used whereas current studies utilise the Illumina sequencing platform. The deamination signal which we aim to model, depends on the chosen PCR protocol and does only differ following the adapter ligation and fill-in step as visualised in Figure S1.

In the 454 Roche sequencing protocol the PCR amplification is initiated by fixating the strand containing the biotin modified p7 adapter in the 5′ termini and the subsequent (emulsion) PCR will only amplify the complementary strand (see middle panel of Figure 1) and the final PCR products will therefore be a subset of the PCR products generated by the standard independent Illumina model which amplifies both strands.

This causes subtle differences in the deamination signal in the PCR products as visualised in Figure 1. Most obviously, our amplification will not be able to recover the molecules where we have damage on both original strands. More details can be found in Figure S1.

1.2 | Briggs Parameters and ngsBriggs

In the Briggs' 2007 article (Briggs et al. 2007), the authors mathematically model the effect of PMD from the historical Roche 454 sequencing platform, deducing four parameters (λ, δ_d, δ_s, ν; denoted throughout as Briggs parameters and their relationship to original aDNA fragments visualised at the bottom left corner of Figure 2):

λ: The parameter of a geometric distribution related to the 5′ overhang length distribution.

δ_d: Deamination level in the double-stranded region.

δ_s: Deamination level in the single-stranded region.

ν: Nick frequency.

In this article, we refer to the inference of these four parameters when considering that the PMD signal stems from PCR products from solely one strand (middle panel of Figure 1) as the biotin model. However, using the biotin model to describe deamination patterns of sequencing reads generated on modern Illumina platforms which amplifies both strands (as previously described) is not suitable. In this paper, we extend the idea of the biotin model (middle panel of Figure 1), and develop a non-biotin model that considers deamination signals from PCR protocols that amplified both strands of the double-stranded DNA aDNA fragment (right panel of Figure 1; Meyer and Kircher 2010).

Some methods have been developed for determining if sequencing reads are produced from ancient DNA. Some use overall nucleotide differences to a reference (e.g., mapDamage (Ginolhac et al. 2011)) at each position in a read but do not rely on an explicit model of damages. In contrast, mapDamage 2.0 (Jónsson et al. 2013) directly infers three of the four Briggs parameters by using a Bayesian Markov chain Monte Carlo approach (MCMC). Other methods, for example, PMDTools (Skoglund et al. 2014) which does not infer the Briggs parameters, can compute a test statistic at the sequence read level for discriminating between reads that exhibit PMD characteristic and those that do not.

To allow for the efficient estimation and calculation of the Briggs parameters, we present a novel statistical method called ngsBriggs that utilises a multinomial model with maximum likelihood fit (the workflow of ngsBriggs is visualised in Figure 2).

Our method estimates the four parameters of the Briggs model based on a set of sequencing reads. As such, ngsBriggs is the most current and relevant tool for estimating deamination patterns of modern Illumina ancient DNA libraries. Furthermore, within the same framework (and by using an external estimate of the contamination fraction), we can also compute, for each read, a probability of it originating from endogenous ancient DNA. This ability to differentiate between the truly ancient reads and the modern contamination combines the functionalities of previous bioinformatical tools in one coherent framework.

2 | Materials and Methods

Both the biotin and non-biotin models attribute any observed PMD signals to nick's placement, the degree of deamination within both the single- and double-stranded regions and the length of the 5′ overhangs (The detailed derivation of both models can be found in Section 2 and 3 in Data S1 and the terminology table is Table S1). λ determines the distribution of the 5′ overhang lengths[1] (both the left and right 5′ overhangs, as

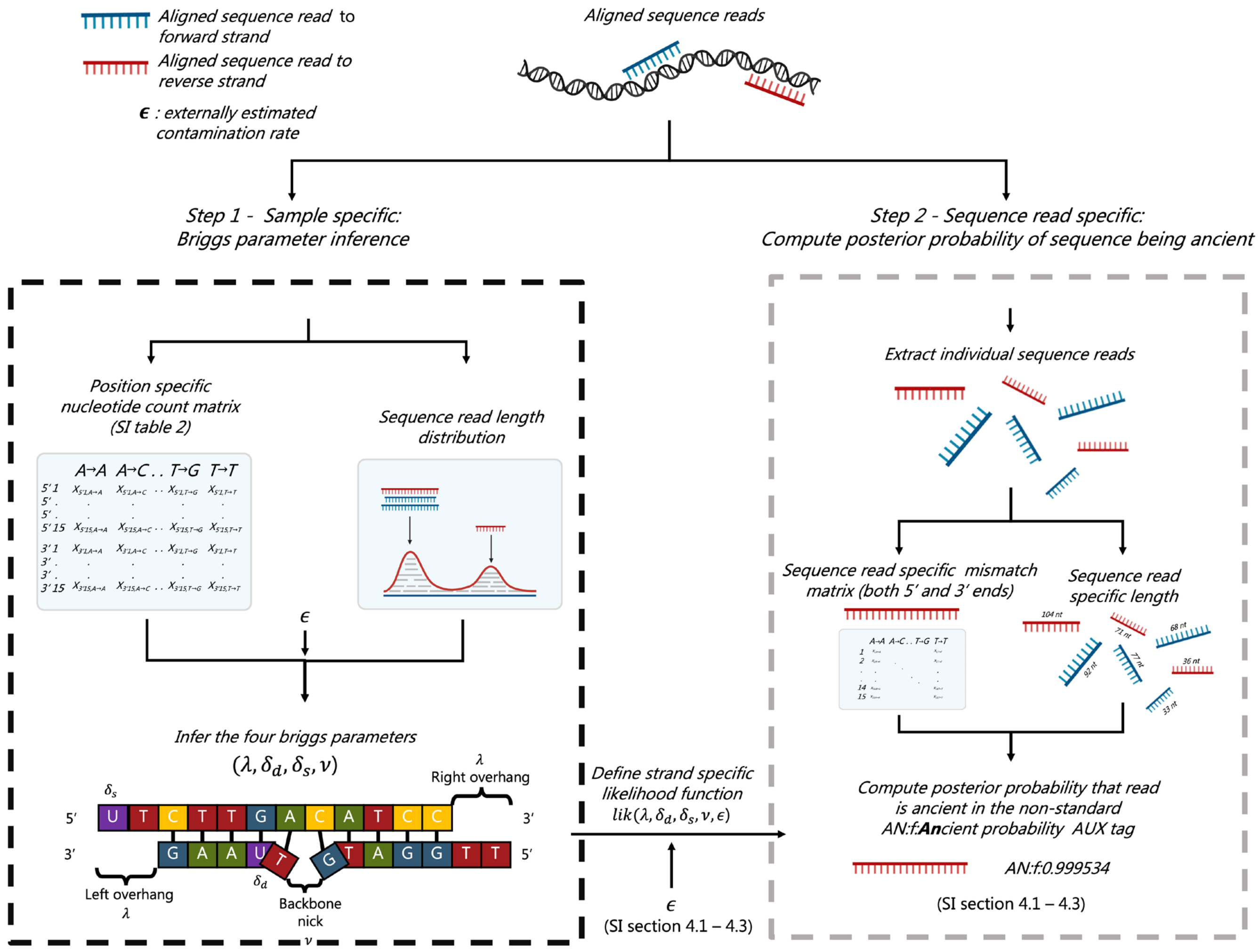

Aligned sequence read to forward strand
Aligned sequence read to reverse strand
ϵ : externally estimated contamination rate
Aligned sequence reads
Step 1 - Sample specific: Briggs parameter inference
Step 2 - Sequence read specific: Compute posterior probability of sequence being ancient
Position specific nucleotide count matrix (SI table 2)
Sequence read length distribution
A→A A→C . . T→G T→T
5′1
5′15
3′1
3′15
ϵ
Infer the four briggs parameters
$(\lambda, \delta_d, \delta_s, \nu)$
δ_s
λ
Right overhang
5′ U T C T T G A C A T C C 3′
3′ G A A U T G T A G G T T 5′
Left overhang
λ
δ_d
Backbone nick
ν
Define strand specific likelihood function
$lik(\lambda, \delta_d, \delta_s, \nu, \epsilon)$
ϵ
(SI section 4.1 – 4.3)
Extract individual sequence reads
Sequence read specific mismatch matrix (both 5′ and 3′ ends)
Sequence read specific length
104 nt
72 nt
68 nt
92 nt
77 nt
36 nt
33 nt
Compute posterior probability that read is ancient in the non-standard AN:f:Ancient probability AUX tag
AN:f:0.999534
(SI section 4.1 – 4.3)

FIGURE 2 | Workflow chart illustrating the two features of ngsBriggs: (1) inference of the four Briggs parameters (left panel in the black dotted square) and (2) sample decontamination (right panel in the grey dotted square). The Briggs inference parameter is sample-specific, utilising information across all aligned reads. First ngsBriggs computes the cycle-specific mismatch matrix for both 5′ and 3′ end, by comparing the sequencing reads to the aligned reference genome region (see Table S1 for full example). Second step when inferring the Briggs parameters is to create a sequence read length distribution. From this cyclic specific count matrix and sequence read length distribution the four Briggs parameters (λ, δ_d, δ_s, ν) can be inferred, with the relation between each parameter and an aDNA fragment shown in the bottom of the left panel. These estimates can also be inferred by providing an externally estimated contamination rate (ϵ) which in combination with cyclic-specific count matrix and read length can be used to define a likelihood function as shown in between the two panels (highlighted by dotted lines). Once a likelihood function is defined (further described in Sections 4.1–4.3 in Data S1), ngsBriggs Step 2 can proceed, which is sequence read specific as it calculates for each read the probability being ancient (grey dotted square). From these probabilities, modern-day sequence reads can be removed to decontaminate any given sample. The inference functionality can be used in a metagenomic framework by providing the mismatch matrix and sequence read length distribution information for each taxon in a metagenomic database.

defined in the original Briggs article (Briggs et al. 2007), share the same distribution:

$$\mathbf{P}(O = l) = 0.5\chi_{l=0} + 0.5\frac{(1-\lambda)^l \lambda}{1-(1-\lambda)^{L-1}} \quad (1)$$

where $\chi_{l=0}$ is an indicator function which takes a value 1 when $l = 0$ and 0 otherwise. L is the focal fragment's length and l is the overhang length. Notice that we make no distinction between the left (denoted as O_l) and right (denoted as O_r) overhang length distribution and assume that the biochemical process that generates the 5′ and 3′ overhangs are similar competing processes. However, we would only observe the 5′ overhang due to the limitations of the double-stranded library preparation, at which point blunt-end repair removes any 3′ overhang and only retains 5′ overhangs. Furthermore, we require the combined total length of the left 5′ overhang and the right 5′ overhang on the same aDNA fragment that cannot exceed $L-2$ (which was also assumed in the original Briggs paper (Briggs et al. 2007)). In practice, reads with long overhangs are not observed due to the configuration being too unstable and thus the fragment would not be chemically feasible.

In theory, nicks can occur at any position of an aDNA fragment. However, the nicks in the single-stranded region will not be observed, as these simply decrease the overhang length. As such, our model will solely consider the nicks within the double-stranded region. Nicks are assumed to occur uniformly along the genome, with a rate (ν) per site. In the original fragment (not the observed sequencing products) we might have multiple nicks; however, after library preparation, all nucleotides downstream from the first nick will be removed and replaced according to the opposite strand (See Figures S3 and S4 for cases when both strands of the fragment have nicks and only one strand of the fragment has nicks; details of nick placement are discussed in Subsection 2.3 in Data S1 and shown in Figures S5 and S6). In the model, we, therefore, focus on the first nick in the double-stranded region on each strand and we only show the first nick in the relevant figures and will refer to the first nick as 'the nick'.

2.1 | Inference of Parameters of the Briggs Model

Our method ngsBriggs uses a multinomial model with maximum likelihood fit, to infer the Briggs parameters, it will create and utilise a mismatch matrix across all reads. The mismatch matrix is the position-specific nucleotide substitution count relative to a reference genome (one example is shown in Table S2). The count of deamination-specific substitutions C→T and G→A in the mismatch matrix will be inflated by sequencing errors and true biological variation. ngsBriggs corrects for this by theoretically separating the effect of the sequencing errors and true biological variations and the C→T and G→A arising from PMD with the assumption that all nucleotide substitutions unrelated to PMD at the same cyclic position have an equal likelihood of occurring across all sequence reads, which can be calculated according to the mismatch matrix (Section 3.1.4 in Data S1).

We can classify the probabilities of four spatial relationships of the focal nucleotide position (n), potential nick and the left or right 5′ overhang:

1. Focal position n is within the double-strand region and downstream[2] of the possible first nick on this strand. Denoted as $p_1(n; L)$.
2. Focal position n is within the right 5′ overhang region. Denoted as $p_2(n; L)$.
3. Focal position n is within the double-strand region and upstream of the possible first nick (this also includes the case without any nick). Denoted as $p_3(n; L)$.
4. Focal position n is within the left 5′ overhang region. Denoted as $p_4(n; L)$.

The four spatial relationship probabilities are functions of not only n and the focal strand's length L, but also the four damage parameters, $\lambda, \delta_d, \delta_s$ and ν as seen in Section 3.1 in Data S1, with the probabilities satisfying $p_1(n; L) + p_2(n; L) + p_3(n; L) + p_4(n; L) = 1$.

If no sequencing errors or contaminating DNA strands are considered, a deaminated C→T given the reference nucleotide is C at position n (counted from 5′ end) on a randomly chosen original ancient strand (a potential template for both models) can only be observed when the spatial relationship satisfies either Type 3 ($p_3(n; L)$, with n being in the double-stranded region with the deamination level δ_d) or Type 4 ($p_4(n; L)$, with n in the single-stranded region, and deamination level δ_s). Hence, the chance of observing a C→T given a reference nucleotide of C is $p_3(n; L)\delta_d + p_4(n; L)\delta_s$. Similarly, the chance of observing a G→A given a reference nucleotide, G, is $p_1(n; L)\delta_d + p_2(n; L)\delta_s$.

However, when the focal position n is on a randomly chosen reverse complement of the original strand (a potential PCR template only for the non-biotin model), a C→T (or a G→A) is equivalent to a G→A (or a C→T) at position $L - n + 1$ on the original strand. Therefore the probability of observing either C→T or a G→A, at position n on the complementary strand, is given as $p_1(L - n + 1; L)\delta_d + p_2(L - n + 1; L)\delta_s$ and $p_3(L - n + 1; L)\delta_d + p_4(L - n + 1; L)\delta_s$, respectively.

Given these relationships, we can define the theoretical deamination frequencies for the model-specific PCR templates (see Figure 1) with the following formulae. Equations (2) and (3) represent the biotin model (abbreviated as b), with Equations (4) and (5) representing the non-biotin model (abbreviated as nb).

$$f_{\mathrm{C}\to\mathrm{T}|\mathrm{C}}(n, L; \mathrm{b}) = p_3(n; L)\delta_d + p_4(n; L)\delta_s \tag{2}$$

$$f_{\mathrm{G}\to\mathrm{A}|\mathrm{G}}(n, L; \mathrm{b}) = p_1(n; L)\delta_d + p_2(n; L)\delta_s \tag{3}$$

$$\begin{aligned} f_{\mathrm{C}\to\mathrm{T}|\mathrm{C}}(n, L;\mathrm{nb}) = 0.5\left[p_3(n;L) + p_1(L-n+1;L)\right]\delta_d \\ +0.5\left[p_4(n;L) + p_2(L-n+1;L)\right]\delta_s \end{aligned} \tag{4}$$

$$\begin{aligned} f_{\mathrm{G}\to\mathrm{A}|\mathrm{G}}(n, L;\mathrm{nb}) = 0.5\left[p_1(n;L) + p_3(L-n+1;L)\right]\delta_d \\ +0.5\left[p_2(n;L) + p_4(L-n+1;L)\right]\delta_s \end{aligned} \tag{5}$$

where at position n in a fixed-fragment-length (L) sample of a specified model (m), $f_{\mathrm{X}\to\mathrm{Y}|\mathrm{X}}(n, L;\ \mathrm{m})$ denotes the frequency of nucleotide change $\mathrm{X} \to \mathrm{Y}$ given X. Throughout the rest of this work, we will use the notation $f_{\mathrm{X}\to\mathrm{Y}|\mathrm{X}}(n, L)$ if the relevant formulae are applied to both models.

The above theoretical frequencies can be further incorporated with sequencing errors and potential contamination (as shown in Sections 3.1.1–3.1.4 of the Data S1). For consistency and simplicity, we will still use the same notations to represent the corresponding frequencies with errors and contamination. Using the mismatch matrix, ngsBriggs maximises the following log-likelihood,

$$\begin{aligned} l(\lambda, \delta_d, \delta_s, \nu) = \sum_n \Bigg\{ N_{\mathrm{C}\to\mathrm{T}|\mathrm{C}}(n) \log\left[\sum_L f_{\mathrm{C}\to\mathrm{T}|\mathrm{C}}(n, L) p_L\right] \\ + N_{\mathrm{G}\to\mathrm{A}|\mathrm{G}}(n) \log\left[\sum_L f_{\mathrm{G}\to\mathrm{A}|\mathrm{G}}(n, L) p_L\right] \Bigg\} \end{aligned} \tag{6}$$

where $N_{\mathrm{X}\to\mathrm{Y}|\mathrm{X}}(n)$ is the actual counts of nucleotide X→Y given X at position n of a randomly chosen fragment from the mismatch matrix, and p_L is the proportion of fragments of length L in the sample after the PCR.

2.2 | Ancient Read Probability in the Presence of Contamination

Once the four Briggs model parameters have been inferred, we can also compute a posterior probability of each read being ancient for samples contaminated with modern human DNA.

We expect the length distribution of ancient reads to be distinguishable from that of modern reads (ancient reads are generally shorter) and have an elevated C→T frequency at the ends; this leads us to the following two assumptions.

1. By assuming that the lengths of the endogenous and modern contamination follow distinguishable constrained normal distributions $\mathcal{N}_b(\mu_a, \sigma_a)$ and $\mathcal{N}_b(\mu_m, \sigma_m)$,[3] we can estimate the values of $(\mu_a, \sigma_a, \mu_m, \sigma_m)$ if the overall modern contamination amount r is provided.

$$\begin{aligned} l(\mu_a,\sigma_a,\mu_m,\sigma_m) &= \sum_k \log \mathbf{P}\left(L_k,O_k \middle| \mu_a,\sigma_a,\mu_m,\sigma_m,\hat{\lambda},\hat{\delta}_d,\hat{\delta}_s,\hat{\nu},\hat{\vec{\epsilon}}_k\right) \\ &\triangleq \sum_k \log\Big[rLik^m_{L_k}(\mu_m,\sigma_m)Lik^m_{O_k}\left(\hat{\vec{\epsilon}}_k\right) \\ &\quad +(1-r)Lik^a_{L_k}(\mu_a,\sigma_a)Lik^a_{O_k}\left(\hat{\lambda},\hat{\delta}_d,\hat{\delta}_s,\hat{\nu},\hat{\vec{\epsilon}}_k\right)\Big] \end{aligned} \tag{7}$$

where $\mathbf{P}(L_k, O_k | \cdots)$ represents the joint probability of ancient strand length L_k and the nucleotide misincorporation pattern O_k. Here $\hat{\vec{\epsilon}}_k$ is the sequencing error per position of strand k provided by the Phred-scaled base quality score, and $(\hat{\lambda}, \hat{\delta}_d, \hat{\delta}_s, \hat{\nu})$ are the inferred damage parameters based on the mismatch matrix. The notation $\text{Lik}_{\cdot}^{\cdot}$ represents the likelihood function associated with the subscripted data (i.e., either L_k or O_k) and the superscript specifies whether the fragment is assumed to be ancient (a) or modern (m). The derivation of $\text{Lik}_{\cdot}^{\cdot}$ is in Sections 4.1 and 4.2 of the Data S1.

2. Based on the estimates $(\hat{\lambda}, \hat{\delta}_d, \hat{\delta}_s, \hat{\nu})$ and $(\hat{\mu}_a, \hat{\sigma}_a, \hat{\mu}_m, \hat{\sigma}_m)$, we can now calculate the posterior probability of being ancient for each strand given the observed nucleotide misincorporation pattern and strand length, for example, for strand k, the posterior probability can be written as follows,

$$\mathbf{P}(a|L_k,O_k) = \frac{(1-r)Lik^a_{L_k}(\hat{\mu}_a,\hat{\sigma}_a)Lik^a_{O_k}\left(\hat{\lambda},\hat{\delta}_d,\hat{\delta}_s,\hat{\nu},\hat{\vec{\epsilon}}_k\right)}{\mathbf{P}\left(L_k,O_k \middle| \hat{\mu}_a,\hat{\sigma}_a,\hat{\mu}_m,\hat{\sigma}_m,\hat{\lambda},\hat{\delta}_d,\hat{\delta}_s,\hat{\nu},\hat{\vec{\epsilon}}_k\right)} \tag{8}$$

The detailed derivation of this expression is given in Section 4 of the Data S1.

2.3 | Published Data and Simulated Files

2.3.1 | Simulated Files for Inference of Briggs Parameters

To determine the accuracy of our estimation framework, we used ngsBriggs inference on simulated files. Using the simulation software NGSNGS (Henriksen et al. 2023), we generated 100 files for the biotin- and 100 for the non-biotin model, equally separated into five groups with a varying number of reads, that is, 10^3, 10^4, 10^5, 10^6 and 10^7 to test different scenarios. All of these files were simulated with a set of 'default' deamination parameters as estimated in the original Briggs article (Briggs et al. 2007), that is, 0.36, 0.0097, 0.68 and 0.024 (λ, δ_d, δ_s, ν, Section 5 in Data S1 for more details).

2.3.2 | Published Data for Different Populations for Inference of Briggs Parameters

We also applied our programme on previously published ancient samples across different time periods, that is, 121 individuals from Barros Damgaard et al. (2018) (termed DA), 219 individuals from Allentoft, Sikora, Refoyo-Martínez, et al. (2024) (NEO), 67 individuals from Allentoft, Sikora, Fischer, et al. (2024) (RISE) and finally 379 individuals from Margaryan et al. (2020) (VK). The aDNA extracted from these human samples originates from diverse tissue types under different preservation conditions with varying ages, all of which contribute to unique PMD signals. As such, we could test our inference models on samples with potentially different deamination patterns (See accession numbers for the published populations in Table 1).

2.3.3 | Simulated Files for Contamination Scenarios

To investigate the effect of contamination, we simulated files with a mixture of ancient- and modern sequencing reads, using contamination proportions of 10%, 20%, 30%, 40% or 50%. For each of the contamination levels, we simulated numerous PMD signals (either biotin or non-biotin) by varying the overhang length λ and single-stranded deamination rate δ_s as a way to signify different levels of ancientness (Section 5 in Data S1).

2.3.4 | Simulated Files for Read-Specific Ancient Probability for Decontamination

To test our ability to discriminate between ancient and modern sequencing reads, we performed two analyses with simulated data, the first testing the accuracy by varying the deamination signal and the second testing varying the overlap between the ancient- and modern-day sequence read lengths. In the first analysis we simulated files with a total of 10^6 reads with a fixed contamination level of 10% and various PMD signals (by varying λ, and δ_s). The lengths of the modern reads were sampled

TABLE 1 | Genetic datasets and their accession numbers.

Populations	Accession number
DA (Barros Damgaard et al. 2018)	PRJEB26349 and ERP107300
NEO (Allentoft, Sikora, Refoyo-Martínez, et al. 2024; Allentoft, Sikora, Fischer, et al. 2024)	PRJEB64656
RISE (Allentoft et al. 2015)	PRJEB9021
VK (Margaryan et al. 2020)	PRJEB37976

from a normal distribution $\mathcal{N}(130,5)$, truncated to be within the interval [30, 145] and the lengths of the ancient read from a log-normal distribution (4, 0.5) truncated to be within the interval [30, 125]. In the second analysis, we simulated datasets with a total of 10^5 reads with 10% contamination, with the same Briggs parameters but varying the distribution parameters of the modern-day contamination including $\mathcal{N}(130, 20)$, $\mathcal{N}(130, 5)$, $\mathcal{N}(110, 20)$, $\mathcal{N}(110, 5)$, $\mathcal{N}(95, 20)$ and $\mathcal{N}(95, 5)$.

3 | Results

3.1 | Inferring Briggs Parameters on Simulated Files

We evaluate the performance of the new method by normalising the root mean square difference between the method-specific inferred values of the Briggs parameters (λ, δ_s, δ_d and ν) and the true values to the true values of the specific parameter, denoted as NRMSE in Figure 3. The simulation commands can be found in Section 5 in Data S1, the corresponding results are presented as Figures 3 and S7–S14.

We also compare the performance of ngsBriggs to mapDamage 2.0 under all scenarios. For both methods, we observe that with a lower number of reads (i.e., a lower depth-of-coverage), the interquartile range and normalised root mean square error increases, signifying more uncertainty in the inferred parameters compared to the true deamination patterns.

In Figure 3, we observe slightly higher NRMSE values for ngsBriggs compared to mapDamage 2.0 when estimating λ and δ_s for 10^3 reads. However with an increasing number of reads ngsBriggs exhibits a lower NRMSE (Figure 3a,b,e,f). We observe the largest difference between the mapDamage 2.0 and ngsBriggs for the δ_d parameter (Figure 3c,d), where mapDamage 2.0 exhibits lower accuracy and precision than ngsBriggs in all scenarios. The estimates of ν have the highest statistical uncertainty for both the biotin and non-biotin model (Figures S10 and S14).

3.2 | Inferring Briggs Parameters on Empirical Data

The results of inferring Briggs parameters on the deamination pattern from empirical data (Figure S15) are presented in Figures S16–S20 and Table S3. Scatter plots of inferred parameters from mapDamage 2.0 and ngsBriggs for all the ancient data (RISE, DA, VK and NEO) can be found in Figure S16. We observed a clear correlation between the corresponding statistics for ngsBriggs and mapDamage 2.0. However, given that the data are empirical, we do not know the true values of the parameters and hence cannot judge which method is more accurate.

Grouping all the ancient samples in different datasets, we conducted Jonckheere-Terpstra test (Jonckheere 1954) and observed significant increasing trends of ngsBriggs estimated λ, δ_d and δ_s as the radiocarbon date of the sample increases (see Figures S19 and S20). While the increasing λ observation is not a direct measurement of deamination, it suggests the older samples have shorter single-stranded regions, possibly due to the instability of short fragments with large single-stranded regions, as previously mentioned. Whereas the increased δ_d and δ_s observations may support the hypothesis that with increasing archaeological age, ancient specimens tend to accumulate deamination. However, there is likely to be a large contribution from type-type, location-specific preservation or project-specific treatment.

With regard to the efficiency of the tools, across all populations, we observed when creating a mismatch matrix that ngsBriggs is several magnitudes faster than both mapDamage 2.0 and PMDtools. When inferring the parameters, ngsBriggs is likewise magnitudes faster than mapDamage 2.0 (Figures S26 and S27, Tables 2 and S4).

3.3 | Benchmarking in the Presence of Contamination

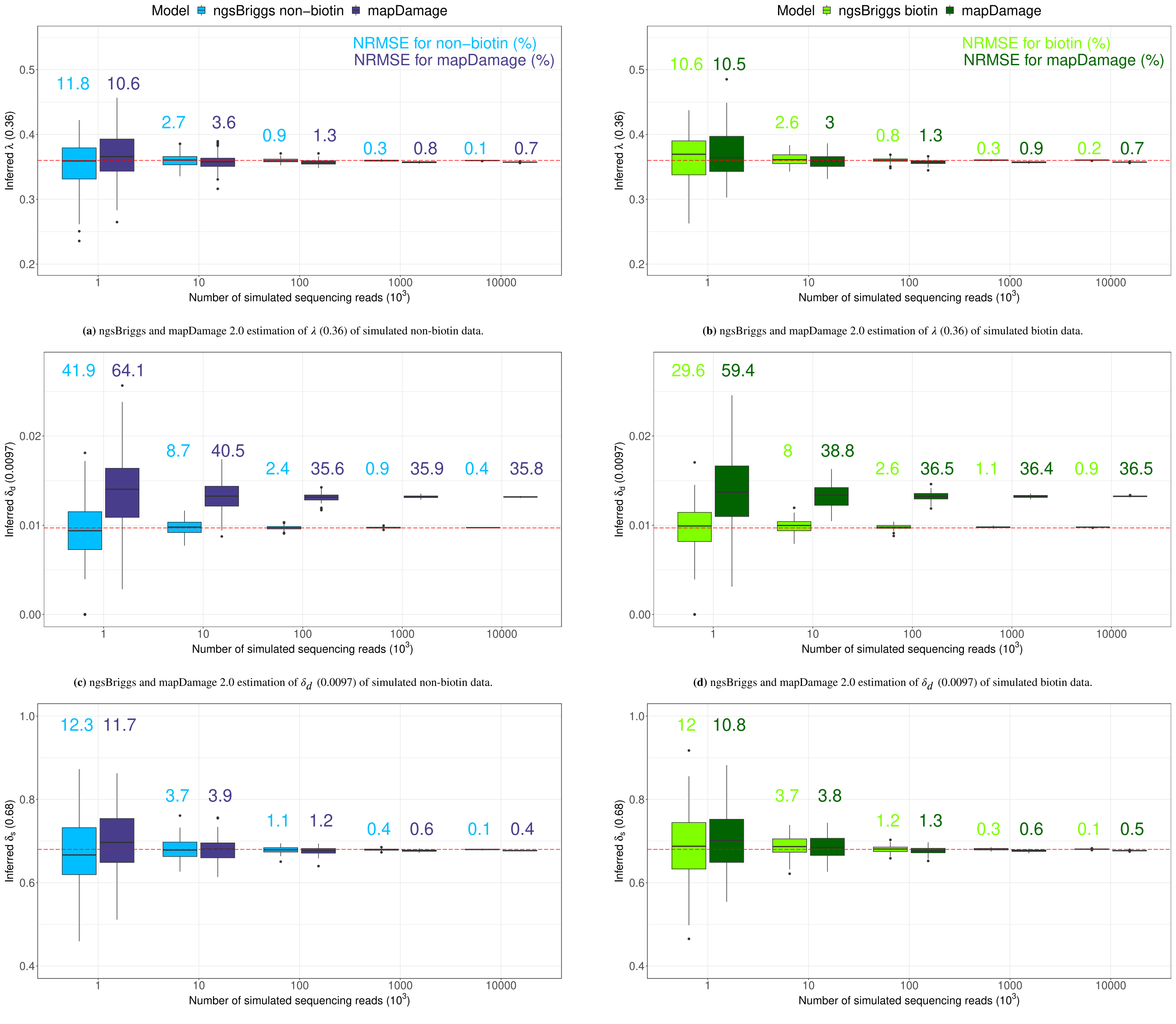
To measure the effect of contamination with modern human DNA, we compared parameter estimates of mapDamage 2.0, which assumes all reads are of ancient origin, to the ngsBriggs estimates, with- or without providing prior knowledge of contamination level (ϵ, Section 6.5 in Data S1, Figures 4 and S21–S25).

When including contamination we observe stable estimates of λ across all investigated scenarios but have a tendency to be biased, which is to be expected, as λ can only be accurately inferred in the decreasing patterns of deamination signals (5′ C→T or 3′ G→A) and therefore remains more unaffected by the extent of contamination. We observe a significant impact from contamination on the three other estimated parameters δ_d, δ_s and ν. MapDamage 2.0 estimates of δ_d, δ_s decreases, due to its inability to take into account contamination from modern sources. When assuming no contamination with ngsBriggs we observe the same trend but allowing for $\epsilon > 0$ it becomes possible to obtain essentially unbiased and accurate results (Figures S21 and S22).

3.4 | Discrimination of Ancient and Modern Reads

We assessed the performance of ngsBriggs when discriminating between ancient and modern reads by comparing them to PMDtools. This was done by using the receiver operating characteristic (ROC) curve plotting the true-positive rate (TPR) against the false-positive rate (FPR) at different classification thresholds (Figure 4). Points above the diagonal signify better classification than could be obtained randomly. With a perfect classification, the line would have a TPR of 1 and FPR of 0 across all thresholds.

We benchmarked the tools as depicted in Figure 4 with the simulated files with 10% contamination, with the ancient component (Section 5.3 in Data S1) having the default deamination settings, similar to the ones used for benchmarking in Figure 3, provided by the original Briggs article (Briggs et al. 2007) (λ: 0.36, δ_d: 0.0097, δ_s: 0.68, ν: 0.024). The sequence reads from the modern component (Section 5.3 in Data S1) follows $\mathcal{N}$ (130, 5) with an upper limit of 145, whereas the ancient sequence reads



(a) ngsBriggs and mapDamage 2.0 estimation of λ (0.36) of simulated non-biotin data.

(b) ngsBriggs and mapDamage 2.0 estimation of λ (0.36) of simulated biotin data.

(c) ngsBriggs and mapDamage 2.0 estimation of δ_d (0.0097) of simulated non-biotin data.

(d) ngsBriggs and mapDamage 2.0 estimation of δ_d (0.0097) of simulated biotin data.

(e) ngsBriggs and mapDamage 2.0 estimation of δ_s (0.68) of simulated non-biotin data.

(f) ngsBriggs and mapDamage 2.0 estimation of δ_s (0.68) of simulated biotin data.

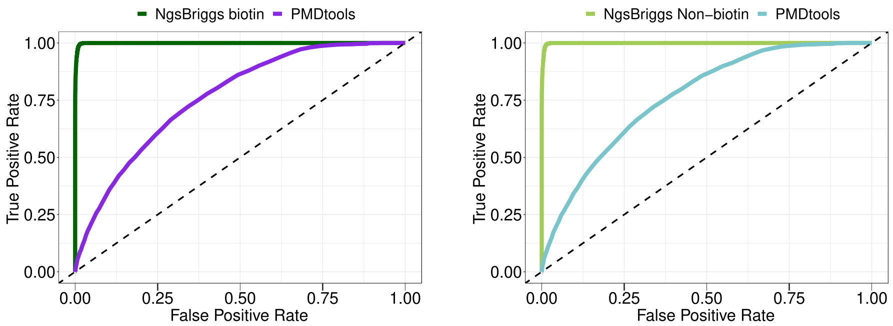
FIGURE 3 | Each subfigure contains the inferred values of a single parameter using ngsBriggs and mapDamage 2.0 across 100 replicates. The percentages above each boxplot represent an error measurement, across the multiple replicates, calculated by normalising the root mean square difference (NRMSE) between the inferred values with the true value of the specific parameter. The subfigures (a), (c) and (e) are from simulated data using the non-biotin deamination model (Henriksen et al. 2023), whereas (b), (d) and (f) are from simulated data using the biotin model. The true parameter value is shown in each subfigure as a red horizontal dotted line intersecting with the inferred values in the y-axis and defined in the y-axis legend and figure caption. The x-axis represent the number of sequencing reads multiplied with 10^3.

follow a constrained log-normal distribution (4, 0.5) with a lower and upper limit of 30–125 respectively. Additional deamination scenarios with variations in the λ and δ_s parameters are presented in Section 6.5.5 in Data S1.

We observe that PMDtools' classification in all ROC curves across the various simulation scenarios remains almost unaffected, as it in most cases can distinguish between ancient DNA reads containing the PMD signal and human contaminants with sequencing error masking as C→T or G→A. However, the remaining ancient DNA fragments with a small fragment length but without a deamination signal remain unclassified. As described in Section 4 in Data S1, ngsBriggs combines the strand length information and PMD signal to infer a potential sequence read length distribution for the contaminants. The ROC curve corresponding to the ngsBriggs method in Figure 4 is closer to the top-left corner. This indicates ngsBriggs almost perfectly discriminates between modern and ancient reads under the explored simulation conditions. The results presented in Figure 4 and Section 6.5.5 in Data S1 only have varying PMD signals. Therefore we also benchmarked the tool with simulated datasets with varying degrees of overlap between the modern and ancient sequencing reads, testing both with a wide and narrow modern-day contamination distribution. The PMD signal of the ancient component are similar to the ones used for benchmarking in Figures 3 and 4 and the sequence reads from the modern-day component were sampled from $\mathcal{N}$ (130, 20) and $\mathcal{N}$ (95,

TABLE 2 | The first three lines represent the mean wall clock running time used to generate the nucleotide matrices of 5′ CT and 3′ G→A deamination frequencies. The next two lines represent the wall clock running time of the Briggs parameter inference (Runtimes are given in seconds).

	Population			
Application	**DA**	**NEO**	**RISE**	**VK**
Mismatch matrix				
mapDamage 2.0	17,446.88	10,884.68	20,500.45	12,359.35
PMDtools	30,428.19	14,202.02	24,856.68	36,412.81
ngsBriggs	257.54	167.91	207.02	187.71
Parameter inference				
mapDamage 2.0	1079.61	1486.84	1052.58	927.57
ngsBriggs	2.33	1.64	2.39	1.53



(a) ngsBriggs biotin (dark green) and PMDtools (purple)

(b) ngsBriggs non-biotin (light green) and PMDtools (blue)

FIGURE 4 | ROC curve illustrating the performance of the classification models, PMDtools and both ngsBriggs models.

20) and to mimic potential decades-old contamination which could be slightly fragmented as seen in Figure 5. With several additional distributions presented in Figure S25. With the greater overlap between the modern contaminants and ancient sequencing reads, we observe the power of ngsBriggs classification is reduced as opposed to Figure S12, as our initial assumption of distinguishable distributions is violated. Despite this decrease, ngsBriggs is still able to classify all sequencing reads exhibiting PMD signals, as well as a proportion of shorter reads without any signal.

3.5 | Runtime

When measuring the wall clock time for the analyses of all empirical data sets, ngsBriggs was considerably faster than mapDamage 2.0, both when creating the mismatch matrix and when inferring the Briggs parameters (Table 2).

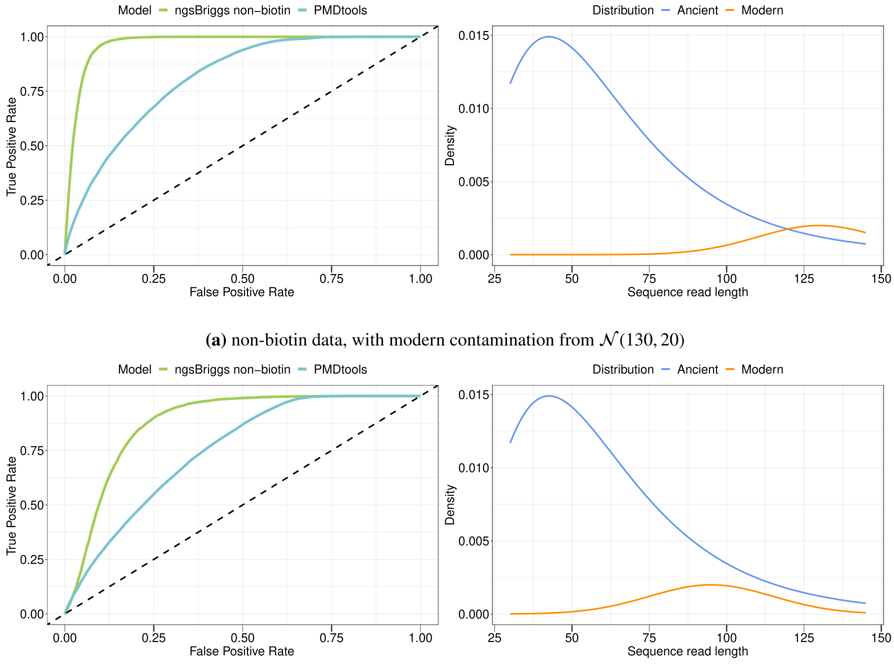
This time disparity is a consequence of the different frameworks of mapDamage 2.0, PMDtools and ngsBriggs. mapDamage 2.0 (Jónsson et al. 2013) combines Python and R and requires multiple independent steps to compute the aDNA-specific metrics (deamination frequency, statistical estimations and visualising the results). PMDtools processes and calculates the PMD using standard output produced by samtools (Danecek et al. 2021) which parses the SAM file. The extra step of processing the samtools output accounts for most of the time difference. However, ngsBriggs computes all these metrics in one step without requiring multiple I/O operations.

4 | ngsBriggs Implementation

The presented methods are implemented in a metagenomic toolkit (metadamage/metaDMG (Michelsen et al. 2022)) as a standalone fast multi-threaded C/C++ function with htslib as a dependency. The code and documentation is available on (https://github.com/RAHenriksen/ngsBriggs,https://www.popgen.dk/software). We use the BFGS algorithm for parameter estimation in our multinomial model. For computing the mismatch matrix, we support the MD:Z tag in the AUX part of the samtools specification, removing the need for parsing the reference (Li et al. 2009). Additionally, the sub-functionality for computing the posterior probability that a read is ancient extends the AUX section of the alignment by adding a new non-standard tag AN:f:ANcient probability.

5 | Discussion

The tool presented in this paper, ngsBriggs, introduces a novel approach for accurately quantifying the post-mortem signal in reads sequenced by both older and modern sequencing platforms. Additionally, ngsBriggs represents a significant advancement in the authentication of ancient samples, as it combines



(a) non-biotin data, with modern contamination from $\mathcal{N}(130, 20)$

(b) non-biotin data, with modern contamination from $\mathcal{N}(95, 20)$

FIGURE 5 | ROC curves (left panel) illustrating the performance of the classification models, PMDtools and ngsBriggs non-biotin model with different length distributions of the contamination (right panel). For simplicity, we did not truncate the length distributions in this figure.

the functionalities of mapDamage 2.0 of estimating damage parameters (λ, δ_d, δ_s and unique to ngsBriggs ν) and PMDtools by computing posterior probabilities to discriminate ancient from modern reads. ngsBriggs creates a global mismatch matrix and infers the parameters with a significantly faster wall clock running time compared to previous methods (Table 2) while classifying the ancient from modern reads with higher accuracy. These factors make ngsBriggs highly suitable for large-scale high-throughput analyses of aDNA data.

By accounting for sequencing errors and true biological variation, ngsBriggs ensured accurate estimates of the Briggs parameters. Once the Briggs parameters are inferred, researchers can effectively decontaminate and refine a given sample, removing reads with a low computed probability of being ancient. This enables researchers to make efficient use of the limited archaeological samples. During DNA extracting and library preparation, these finite ancient samples are inevitably destroyed, raising ethical implications. Although these ethical concerns might not necessarily impede scientific progress, our tool makes it possible for researchers to consider this, by mitigating the need for repeated and often destructive sampling thus contributing to the continued sustainable paleogenomics practice.

External estimates of contamination levels can mitigate the negative effect of exogenous DNA on the PMD signal, and ngsBriggs retains high accuracy during the inference of Briggs parameters and the assignment of read-specific contamination probabilities when such estimates are available.

Our current implementation framework shows accurate inference of the Briggs parameters and we show through extensive simulations, also using the Briggs model, that we are able to obtain essentially unbiased estimates of our statistics. An obvious limitation is that the PMD signal is poorly understood and the biochemical properties that are modelled directly by the four Briggs parameters will still pose an issue.

In the presence of contamination, we achieved more accurate estimates from ngsBriggs when providing a known contamination level. Hence, by leveraging ngsBriggs alongside tools like ContaMix (Fu et al. 2013), Schmutzi (Renaud et al. 2015) or ANGSD (Korneliussen, Albrechtsen, and Nielsen 2014), it is possible to estimate PMD features with high accuracy, even in the presence of significant contamination. It should be noted that the existing methods, for example, PMDtools, can only distinguish ancient reads from their modern counterparts based on the observed PMD signals, hence it will be difficult for them to identify those reads originating from ancient material but without PMD patterns. In contrast, ngsBriggs combines information on length distribution and PMD signals. By assuming the differences in lengths between modern and ancient DNA reads, it gains more power to identify ancient reads, even without evidence of PMD. However, one potential issue with ngsBriggs arises in the presence of contamination when the assumption of distinguishable length distribution for the endogenous and modern contamination DNA is violated. When employing sequencing platforms like NovaSeq X with a 50 nucleotide cycle length, most sequenced reads, aDNA and potential modern contamination alike, will exhibit a similar length distribution, thus violating this assumption. While ngsBriggs may be successful in distinguishing some modern and ancient reads, the presented advantage as seen with the ROC curves would diminish.

Other sequencing platforms and laboratory procedures might also present issues, as alternative library protocols could likewise influence the deamination pattern violating the originally quantified deamination (Briggs et al. 2007) which ngsBriggs and mapDamage 2.0 models. Since the original paper (Briggs et al. 2007), the usage of 454 Roche sequencing of aDNA has diminished; however, our biotin model also accommodates the PCR products of other NGS platforms, which would similarly attach fragmented DNA to beads and amplify the sequences by emulsion PCR, such as ION torrent (Akintunde, Tucker, and Carabetta 2023). Importantly, pre-amplifying the library with standard PCR would mimic the non-biotin approach even though emulsion PCR is performed subsequently.

Alternative library construction methods used include the New England Biolabs' NEBNext Ultra II kit (Gansauge et al. 2017), the y-shaped adapter double-stranded libraries which is further discussed in the Section 7 in Data S1. Several specialised laboratory techniques have been developed to increase the potential number of ancient DNA molecules. One of the latest, SCR/SRSLY single-stranded libraries (Bennett et al. 2014) for Illumina machines, might prove advantageous in conjunction with the software AuthentiCT (Peyrégne and Peter 2020), which can estimate contamination from DNA substitutions, similarly to the second functionality of ngsBriggs (as depicted in the right panel of Figure 1). However, ngsBriggs is not applicable to single-stranded library sequencing data. As discussed in detail in Section 7 in Data S1, the estimation of four Briggs parameters for single-stranded libraries can be misleading. Additional laboratory techniques focus on reducing the influence of aDNA errors by preparing libraries with the USER enzymes (Rohland et al. 2015). The USER treatment removes the deamination signal with uracil–DNA–glycosylase (UDG) (Briggs et al. 2010) and cleaves the 3′ backbone of the abasic site with an endonuclease VIII, limiting the number of apparent nucleotide substitutions caused by deamination. USER-treated data cannot be analysed by ngsBriggs, as it specifically aim at identifying ancient DNA using damage patterns.

As previously stated, ngsBriggs is implemented as part of a metaDMG framework to extend the mismatch matrix and the Briggs parameter inference functionalities to each taxonomic identifier ('taxid') found in metagenomic data aligned to a reference database. By leveraging ngsBriggs within the metaDMG framework, researchers can authenticate and gain valuable insights into the taxonomic composition of ancient environmental DNA samples (Fernandez-Guerra et al. 2023). This could enable researchers to delve deeper into the timeline of ancient samples and potentially gain new insight. This will broaden the scope of research in multiple fields, including paleogenomics, phylogenomics, metagenomics and historical ecology.

Acknowledgements

We thank Dr. Lasse Vinner, Dr. Hugh McColl and Dr. Søren Overballe for many discussions regarding the laboratory protocols. This work has been supported by the Carlsberg Foundation Young Researcher Fellowship awarded by the Carlsberg Foundation in 2019 [CF19-0712 to R.A.H, T.S.K and L.Z]. The Lundbeck Foundation Centre for Disease Evolution: [R302-2018-2155 L.Z] and Carlsberg Foundation (Queen Margrethe's and Vigdís Finnbogadóttir's Interdisciplinary Research Centre on Ocean, Climate, and Society, [CF20-0071 to A.R]).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Genetic data: Raw sequence reads and alignment files for the four annotated populations (DA, NEO, RISE and VK) have all been deposited in the Sequence Read Archive with the following identifiers (See Table 1). Code available in (https://github.com/RAHenriksen/ngsBriggs, https://www.popgen.dk/software).

Endnotes

[1] The length distribution of 5′ overhangs is a mixed distribution: $50\,\%$ of the chances are a focal 5′ overhang has the length 0, while the other $50\,\%$ of the chances are the length of this 5′ overhang follow a geometric distribution with the parameter λ (See Equation 1). Intuitively, a smaller λ indicates a short mean overhang length.

[2] The terms position n, downstream and upstream are in the sense of 5′ end to 3′ end.

[3] The assumption of normality is strong but for efficient computation, it can be viewed to define a score to distinguish reads. In Figures 4 and S24 and S25, we have proved it works for different cases, for example, log-normal length distributions or with a greater overlap between the two distributions, which violate this assumption.

References

Akintunde, O., T. Tucker, and V. J. Carabetta. 2023. "The Evolution of Next-Generation Sequencing Technologies." *ArXiv*.

Allentoft, M. E., M. Sikora, A. Refoyo-Martínez, et al. 2024. "Population Genomics of Post-Glacial Western Eurasia." *Nature* 625: 301–311. https://doi.org/10.1038/s41586-023-06865-0.

Allentoft, M. E., M. Sikora, A. Fischer, et al. 2024. "100 Ancient Genomes Show Repeated Population Turnovers in Neolithic Denmark." *Nature* 625: 329–337. https://doi.org/10.1038/s41586-023-06862-3.

Allentoft, M. E., M. Sikora, K. G. Sjögren, et al. 2015. "Population Genomics of Bronze Age Eurasia." *Nature* 522, no. 7555: 167–172.

Barros Damgaard, d P., R. Martiniano, J. Kamm, et al. 2018. "The First Horse Herders and the Impact of Early Bronze Age Steppe Expansions Into Asia." *Science* 360, no. 6396: eaar7711.

Bennett, E. A., D. Massilani, G. Lizzo, J. Daligault, E. M. Geigl, and T. Grange. 2014. "Library Construction for Ancient Genomics: Single Strand or Double Strand?" *BioTechniques* 56, no. 6: 289–300.

Briggs, A. W., U. Stenzel, P. L. Johnson, et al. 2007. "Patterns of Damage in Genomic DNA Sequences From a Neandertal." *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 37: 14616–14621.

Briggs, A. W., U. Stenzel, M. Meyer, J. Krause, M. Kircher, and S. Pääbo. 2010. "Removal of Deaminated Cytosines and Detection of In Vivo Methylation in Ancient DNA." *Nucleic Acids Research* 38, no. 6: e87.

Dabney, J., M. Meyer, and S. Pääbo. 2013. "Ancient DNA damage." *Cold Spring Harbor Perspectives in Biology* 5, no. 7: a012567.

Danecek, P., J. K. Bonfield, J. Liddle, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10, no. 2: giab008.

Fernandez-Guerra, A., G. Borrel, T. O. Delmont, et al. 2023. "A 2-Million-Year-Old Microbial and Viral Communities From the Kap København Formation in North Greenland." *bioRxiv*.

Fu, Q., A. Mittnik, P. L. Johnson, et al. 2013. "A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes." *Current Biology* 23, no. 7: 553–559.

Gansauge, M. T., T. Gerber, I. Glocke, et al. 2017. "Single-Stranded DNA Library Preparation From Highly Degraded DNA Using T4 DNA Ligase." *Nucleic Acids Research* 45, no. 10: e79.

Ginolhac, A., M. Rasmussen, M. T. P. Gilbert, E. Willerslev, and L. Orlando. 2011. "mapDamage: Testing for Damage Patterns in Ancient DNA Sequences." *Bioinformatics* 27, no. 15: 2153–2155.

Henriksen, R. A., L. Zhao, T. S. Korneliussen, and NGSNGS. 2023. "Next Generation Simulator for Next Generation Sequencing Data." *Bioinformatics* 39: btad041. https://doi.org/10.1093/bioinformatics/btad041.

Jonckheere, A. R. 1954. "A Distribution-Free k-Sample Test Against Ordered Alternatives." *Biometrika* 41, no. 1–2: 133–145. https://doi.org/10.1093/biomet/41.1-2.133.

Jónsson, H., A. Ginolhac, M. Schubert, P. L. Johnson, and L. Orlando. 2013. "mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters." *Bioinformatics* 29, no. 13: 1682–1684.

Kjær, K. H., M. Winther Pedersen, B. De Sanctis, et al. 2022. "A 2-Million-Year-Old Ecosystem in Greenland Uncovered by Environmental DNA." *Nature* 612, no. 7939: 283–291.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen. 2014. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics* 15, no. 1: 356.

Li, H., B. Handsaker, A. Wysoker, et al. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25, no. 16: 2078–2079.

Mallick, S., A. Micco, M. Mah, et al. 2023. "The Allen Ancient DNA Resource (AADR): A Curated Compendium of Ancient Human Genomes." *bioRxiv*.

Margaryan, A., D. J. Lawson, M. Sikora, et al. 2020. "Population Genomics of the Viking World." *Nature* 585, no. 7825: 390–396.

Meyer, M., and M. Kircher. 2010. "Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing." *Cold Spring Harbor Protocols* 2010, no. 6: pdb.prot5448.

Michelsen, C., M. W. Pedersen, A. Fernandez-Guerra, L. Zhao, T. C. Petersen, and T. S. Korneliussen. 2022. "metaDMG-A Fast and Accurate Ancient DNA Damage Toolkit for Metagenomic Data." *bioRxiv*.

Peyrégne, S., and B. M. Peter. 2020. "AuthentiCT: A Model of Ancient DNA Damage to Estimate the Proportion of Present-Day DNA Contamination." *Genome Biology* 21, no. 1: 246.

Renaud, G., V. Slon, A. T. Duggan, and J. Kelso. 2015. "Schmutzi: Estimation of Contamination and Endogenous Mitochondrial Consensus Calling for Ancient DNA." *Genome Biology* 16, no. 1: 224.

Rohland, N., E. Harney, S. Mallick, S. Nordenfelt, and D. Reich. 2015. "Partial Uracil–DNA–Glycosylase Treatment for Screening of Ancient DNA." *Philosophical Transactions of the Royal Society B: Biological Sciences* 370, no. 1660: 20130624.

Sambrook, J., E. Fritsch, and T. Maniatis. 1989. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Skoglund, P., B. H. Northoff, M. V. Shunkov, et al. 2014. "Separating Endogenous Ancient DNA From Modern Day Contamination in a Siberian Neandertal." *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 6: 2229–2234.

Tabor, S., K. Struhl, S. J. Scharf, and D. H. Gelfand. 1997. "DNA-dependent DNA polymerases." *Current Protocols in Molecular Biology* 37, no. 1: 3–5.

Willerslev, E., and A. Cooper. 2005. "Ancient dna." *Proceedings of the Royal Society B: Biological Sciences* 272, no. 1558: 3–16.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.