Sequence analysis

Inference and visualization of DNA damage patterns using a grade of membership model

Hussein Al-Asadi^{1,2,*†}, Kushal K. Dey^{2,*†}, John Novembre^{1,3} and Matthew Stephens^{2,3}

¹Committee on Evolutionary Biology, ²Department of Statistics and ³Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: John Hancock

Received on June 6, 2018; revised on August 11, 2018; editorial decision on August 29, 2018; accepted on September 4, 2018

Abstract

Motivation: Quality control plays a major role in the analysis of ancient DNA (aDNA). One key step in this quality control is assessment of DNA damage: aDNA contains unique signatures of DNA damage that distinguish it from modern DNA, and so analyses of damage patterns can help confirm that DNA sequences obtained are from endogenous aDNA rather than from modern contamination. Predominant signatures of DNA damage include a high frequency of cytosine to thymine substitutions (C-to-T) at the ends of fragments, and elevated rates of purines (A & G) before the 5' strand-breaks. Existing QC procedures help assess damage by simply plotting for each sample, the C-to-T mismatch rate along the read and the composition of bases before the 5' strand-breaks. Here we present a more flexible and comprehensive model-based approach to infer and visualize damage patterns in aDNA, implemented in an R package *aRchaic*. This approach is based on a 'grade of membership' model (also known as 'admixture' or 'topic' model) in which each sample has an estimated grade of membership in each of K damage profiles that are estimated from the data.

Results: We illustrate *aRchaic* on data from several aDNA studies and modern individuals from 1000 Genomes Project Consortium (2012). Here, *aRchaic* clearly distinguishes modern from ancient samples irrespective of DNA extraction, lab and sequencing protocols. Additionally, through an *in-silico* contamination experiment, we show that the *aRchaic* grades of membership reflect relative levels of exogenous modern contamination. Together, the outputs of *aRchaic* provide a concise visual summary of DNA damage patterns, as well as other processes generating mismatches in the data.

Availability and implementation: *aRchaic* is available for download from <https://www.github.com/kkdey/aRchaic>.

Contact: halasadi@uchicago.edu or kkdey@uchicago.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Ancient DNA (aDNA) research has seen rapid growth with the recent advancements in recovery of short DNA fragments, increased throughput, and lower per-base cost in sequencing (Shapiro and

Hofreiter, 2014). Both the number and size of aDNA datasets have grown rapidly over the last five years, and several recent studies sequenced hundreds of ancient individuals (Allentoft *et al.*, 2015; Lazaridis *et al.*, 2016; Lipson *et al.*, 2017; Mathieson *et al.*, 2015, 2018; Olalde *et al.*, 2018).

This rapid recent growth in aDNA research has provided novel insights into human history. However, working with aDNA remains challenging. For example, ancient samples often contain very little endogenous DNA because in many environments DNA degrades rapidly post-mortem (Sawyer *et al.*, 2012). Furthermore, ancient samples are often contaminated by microbes and exogenous human DNA (Malmström *et al.*, 2007). Both these factors mean that many sequence reads generated by an aDNA study may not actually come from the ancient sample.

Because of these challenges, aDNA researchers pay careful attention to quality control (QC), including checking sequencing reads from each sample for signatures of endogenous aDNA. These signatures include: short fragment length, an enrichment of purines before strand breaks and a high frequency of cytosine to thymine substitutions (C-to-T) at the ends of fragments (Briggs *et al.*, 2007; Ginolhac *et al.*, 2011; Jónsson *et al.*, 2013; Sawyer *et al.*, 2012; Skoglund *et al.*, 2014b). One common QC procedure is to plot, for each sample, the C-to-T mismatch rate as a function of position from the end of the read, and to look for an elevated rate near the ends of reads as an indication of the presence of endogenous aDNA. Another common procedure is to look for elevated rates of purines (A & G) before the 5' strand-breaks. Both these procedures are implemented in the software *mapdamage* (Ginolhac *et al.*, 2011; Jónsson *et al.*, 2013), for example.

These commonly-used QC checks, though simple and useful, have several limitations. For example, they produce a plot for each sample, which can be inconvenient to work with and difficult to compare across many samples. This issue becomes increasingly important with the growing size of aDNA datasets. These plots can also be difficult to interpret, in part because they do not contrast observed patterns with expected patterns in modern samples. Finally, these approaches can detect only pre-defined damage signatures, and may fail to capture other key features or artifacts in the data.

Here we introduce methods to help address these problems. These methods start with a Binary Alignment Map (BAM) file, obtained by aligning each read to a reference genome. The BAM file includes information on the mismatches that occur in each read (versus the reference). We characterize each mismatch by several relevant features, including its type (e.g. C-to-G, etc.), flanking bases and distance from the end of the read. We then use these features to cluster the mismatches into groups, which we call *mismatch profiles*. Intuitively, a mismatch profile associated with post-mortem damage is expected to show high levels of C-to-T mismatches at the ends of reads. On the other hand, a mismatch profile that is typical of modern DNA polymorphism will show a different pattern, such as a transition to transversion ratio of 2:1 (Goldman and Yang, 1994). Finally we estimate the relative frequency of each mismatch profile in each sample, which we refer to as the 'Grade of membership' (Erosheva, 2006) of that sample in that mismatch profile. These grades of membership should reflect which processes generated mismatches in each sample. For example ancient samples should have a high grade of membership in mismatch profiles characteristic of post-mortem DNA damage. Grade of membership models are widely used to infer structure in admixed populations (Pritchard *et al.*, 2000), document collections (Blei *et al.*, 2003), RNA-seq data Dey *et al.* (2017b) and somatic mutation data (Shiraishi *et al.*, 2015) for example.

We have implemented methods to fit this model, and visualize the results in a software package, *aRchaic*. For example, the grades of membership for all samples are succinctly displayed in a single STRUCTURE plot (Rosenberg *et al.*, 2002), and each mismatch

profile is displayed using simple intuitive plots (Dey *et al.*, 2017a). Together these plots provide a concise visual summary of DNA damage patterns, as well as other processes generating mismatches in the data.

2 Materials and methods

For each sample $i = 1, \dots, I$ we first obtain a BAM file. From this BAM file we extract information on the mismatches (versus a reference) that occur in reads from the sample. First we filter out low-quality reads (mapping ≤ 30), low-quality mismatches (base quality ≤ 20) and mismatches that occur more than 20 bp from the end of a read [since these are unlikely to reflect damage patterns (Briggs *et al.*, 2007)]. When a read carries more than one mismatch we treat these as independent (an assumption we verified by checking that the probability of a mismatch conditional on the occurrence of another mismatch on the same read is not significantly different from the marginal probability, P -value = 0.43).

Let J_i denote the total number of remaining mismatches. For each mismatch $j = 1, \dots, J_i$, we first identify the strand in the reference genome that carries a C or T allele; we call this strand the 'reference strand' and denote it by S_j . Let $x_{ij} = (x_{ij,1}, x_{ij,2}, x_{ij,3}, x_{ij,4}, x_{ij,5})$ denote the following features of the mismatch (see Supplementary Fig. S1 for illustration):

1. $x_{ij,1} \in \{T \text{ to } -A, T \text{ to } -C, T \text{ to } -G, C \text{ to } -A, C \text{ to } -G, C \text{ to } -T\}$ denotes the mismatch (with respect to the reference strand S_j).
2. $x_{ij,2} \in \{A, C, G, T\}$ denotes the nucleotide immediately 5' to the mismatch on S_j .
3. $x_{ij,3} \in \{A, C, G, T\}$ denotes the nucleotide immediately 3' to the mismatch on S_j .
4. $x_{ij,4} \in \{0, \dots, 20\}$ denotes the distance (in base-pairs) from the mismatch to the nearest end of the read.
5. $x_{ij,5} \in \{A, C, G, T\}$ denotes the nucleotide that occurs one base upstream from the 5' end of the reference strand that is closest to the location of mismatch. Since mismatches resulting from damage primarily occur at 5' ends of the reads, this features captures the enrichment of purines immediately one base 5' upstream of the strand-break for samples with sufficient DNA damage (Briggs *et al.*, 2007; Sawyer *et al.*, 2012).

These features are designed to reflect the major modes of DNA damage (Briggs *et al.*, 2007; Prüfer *et al.*, 2014; Sawyer *et al.*, 2012a,b). For each feature $l \in \{1, \dots, 5\}$ we let M_l denote the number of possible values of $x_{ij,l}$, and for notational convenience we relabel the possible outcomes such that $x_{ij,l}$ takes on integer values in $\{1, \dots, M_l\}$. For example we represent $x_{ij,1} = T \text{ to } -A$ by $x_{ij,1} = 1$.

Our model assumes that each mismatch in each individual arose from one of K mismatch profiles ('clusters'). We introduce latent variables $z_{ij} \in \{1, \dots, K\}$ to denote the profile that gave rise to mismatch j in individual i . We assume

$$\Pr(z_{ij} = k) = q_{i,k}, \quad (1)$$

where $q_{i,k}$ represents the membership proportion of individual i in mismatch profile (cluster) $k \in 1, \dots, K$.

We further assume that, given $z_{ij} = k$,

$$\Pr(x_{ij,l} = m | z_{ij} = k) = f_{k,l}(m), \quad (2)$$

where $m \in \{1, \dots, M_l\}$, and $f_{k,l}(m)$ denotes the relative frequency of m at feature l in cluster k . We follow (Shiraishi

Fig. 1. Illustration of the aRchaic grades of membership and mismatch profiles. (a) The features of a mismatch modeled by aRchaic. (b) A depiction of an ancient DNA sample that has 80% of its reads assigned to cluster 1 and 20% of its reads assigned to cluster 2. Each cluster is defined by a *mismatch profile* showing the enrichment of the mismatch type, bases flanking the mismatch, the distance of the mismatch from the nearest end of the read, and the base immediately 5' to the strand-break. To produce a mismatch profile for a cluster, mismatch features are aggregated across reads assigned to the cluster, and their frequencies are represented by an *EDLogo* plot (Dey et al., 2017a). In the *EDLogo* plot, the frequencies are scaled against a background frequency computed from 1000 Genomes Project Consortium (2012)

et al., 2015) in assuming independence among features within each cluster.

Putting this all together, and assuming independence of observations yields the likelihood:

$$L(q, f; x) = \prod_{i,j,l} \sum_k f_{k,l}(x_{i,j,l}) q_{i,k}. \quad (3)$$

We fit this model, and estimate the individual parameters (q) and cluster parameters (f) by maximum likelihood using an accelerated EM algorithm. We use the same EM updates as in Equations 2–4 in (Shiraishi et al., 2015), and we add first-order quasi-Newton acceleration to improve convergence (Alexander et al., 2009; Lange, 1995; Taddy, 2012).

For each cluster k , we visualize the cluster parameters f_k using an *EDLogo* plot (Dey et al., 2017a) (see Fig. 1). The *EDLogo* plot allows one to visualize both enrichment and depletion of mismatch features scaled against a reference frequency. In our application, the reference frequency was computed by averaging the proportion of the five aRchaic features across individuals in the 1000 Genomes Project. This allows us to compute an enrichment score which effectively compares mismatch profiles in our samples against that of modern individuals from 1000 Genomes Project Consortium

(2012). We use a STRUCTURE plot (Rosenberg et al., 2002) to visualize the estimates of $q_{i,k}$ for each sample.

3 Results

We demonstrate the utility of aRchaic using three case-studies.

3.1 aRchaic clustering of modern and ancient individuals

We applied aRchaic to a combined dataset of 52 ancient samples from four recent studies (Gamba et al., 2014; Lazaridis et al., 2016; Lipson et al., 2017; Skoglund et al., 2014a) and 60 modern samples from 1000 Genomes Project Consortium (2012) ($n=50$) and 10 individuals from HGDP (Cann et al., 2002) individuals sequenced by Meyer et al. (2012). Two of the aDNA studies used partial-UDG treated libraries, which removes most—but not all—of the C-to-T deamination (Rohland et al., 2015).

Figure 2 shows results from aRchaic with $K=3$ (see Supplementary Fig. S2 for $K=4, 5, 6$). To give a sense of computational requirements, these results took approximately 23 min to generate on a single modern compute node. The results clearly highlight

Fig. 2. aRChaic clearly distinguishes between modern, ancient (UDG) and ancient (non-UDG) samples. aRChaic is applied with $K=3$ to a collection of ancient individuals from four studies (Gamba *et al.*, 2014; Lazaridis *et al.*, 2016; Lipson *et al.*, 2017; Skoglund *et al.*, 2014a) along with modern individuals randomly sampled from the 1000 Genomes Project and 10 individuals from the Human Genome Diversity Panel (1000 Genomes Project Consortium, 2012; Cann *et al.*, 2002; Meyer *et al.*, 2012). Modern samples have high membership in the red cluster. The EDLogo representation of this cluster does not show strong enrichment against a modern background. The ancient (non-UDG) samples are representative of the blue cluster. The EDLogo plot for the blue cluster shows a strong enrichment in C-to-T mismatches at the end of reads, a depletion of guanine in the right flanking base, and a depletion of cytosine at the 5' strand-break. The ancient (UDG) samples have partial membership both in the red cluster and in the gold cluster. The EDLogo plot for the gold cluster is enriched in C-to-T mismatches at the terminal ends of the reads, shows an enrichment of guanine at the right flanking base, and a depletion of thymine one base 5' upstream of the strand break

differences between modern, ancient (UDG) and ancient (non-UDG) samples. The modern samples show very strong membership in a single cluster (red). As expected, this 'modern' cluster shows only modest enrichment in its mismatch type, flanking base composition and mismatch location, relative to the modern background.

Ancient (non-UDG) samples show high membership in a second cluster (blue). This cluster is characterized by a very strong enrichment of C-to-T mismatches at the ends of the reads, which is a typical sign of DNA damage (Rohland *et al.*, 2015), and this enrichment is accompanied by a depletion of guanine just 3' to the mismatch. We see this depletion in guanine because the blue cluster is driven by mismatches at cytosine sites which seldom precedes a guanine because CpG sites occur less frequently than expected (Shen *et al.*, 1994).

The ancient UDG-treated individuals show high membership in the third (orange) cluster, and partial membership in the first (red) cluster. The membership in the red cluster presumably reflects the fact that the UDG-treatment repairs much of the damage in these samples, making them look more 'modern' in their mismatch profiles. The orange cluster is characterized by enrichment of C-to-T mismatches very close to the ends of reads, with a strong enrichment

of guanine at the right flanking base. That is, an enrichment of CpG-to-TpG mismatches at the ends of reads. This may be explained by the fact that when a methylated cytosine undergoes deamination it becomes thymine (in contrast to unmethylated cytosines, which deaminate to uracil) and these thymines are not repaired by the UDG-treatment (Duncan and Miller, 1980). Furthermore, we see a depletion of thymine 5' upstream of the strand-break, which consequently is manifested as a depletion of thymine at the left flanking base.

3.2 The effects of contamination on inferred grades of membership

We next sought to examine the effects of exogenous modern contamination on inferred grades of memberships in ancient samples.

We performed an *in-silico* experiment to artificially contaminate ancient samples with modern data from the 1000 Genomes Project (1000 Genomes Project Consortium, 2012). We selected one BAM file from an ancient sample [K01 from (Gamba *et al.*, 2014)], and split its reads into 10 equal subsets. We then contaminated each subset with reads from a distinct modern individual from the 1000

Fig. 3. Estimated grades of membership reflect levels of contamination. (a) Reads from one ancient individual [KO1 from [Gamba et al. \(2014\)](#)] were split into 10 equally sized groups. Reads were added from a distinct individual in the 1000 Genomes Project (1000 Genomes Project Consortium, 2012) to each group (S1–S10) at varying levels of percentages (indicating levels of contamination). (b) We applied aRchaic with $K=2$ on a combined dataset comprised of these 10 contaminated groups of reads (S1–S10) along with 40 other modern individuals from 1000 Genomes. (c) The grades of membership in cluster 1 (‘modern’ cluster) were plotted as a function of the percentage of contamination before (red curve) and after (green curve) applying the correction factor discussed in the last paragraph of Section (3.2). (d) We repeated the same experiment as described in panel a, except we discarded all reads with no mismatches or greater than one mismatch. The grades of membership in cluster 1 were plotted as a function of the ‘mismatch contamination rate’ which is defined as the proportion of mismatches that originate from a contaminated read. (e) Each group (S1–S10) was further sub-sampled to 10 000 reads, and aRchaic was applied with $K=2$ to the new subsampled groups and the same 40 modern individuals as in panel b

Genomes Project, varying the contamination level from 0 to 100% (Fig. 3a). This results in 10 samples (S1–S10) representing 10 contaminated ancient samples with known levels of modern contamination.

We applied aRchaic with $K=2$ on the contaminated samples (S1–S10) plus 40 other modern individuals (randomly sampled from the 1000 Genomes Project; Fig. 3b). Modern individuals showed high grades of membership in one cluster (red). Fully contaminated individuals nearly showed full membership in the red cluster, while uncontaminated ancient samples showed essentially no membership in this cluster. For samples in between, membership in the red cluster increased approximately monotonically with the level of contamination (Fig. 3c). We obtained similar results even with only 10000 randomly-sampled reads for each sample (Fig. 3e) implying that these results are robust to low sequencing depth.

We find that the grades of membership in the red cluster (representing moderns) are consistently less than the proportion of contaminated reads (Fig. 3c red curve). This pattern can be partly explained by the fact that the vast majority of reads from a DNA library contain no damage. To elaborate, only a fraction of

contaminated DNA from a modern source have mismatches, and this fraction will typically be less than for ancient DNA. Thus, the estimated proportion of mismatches arising from a ‘modern DNA’ cluster should generally be expected to be less than the contamination fraction of the library. This implies aRchaic will typically provide a lower bound on contamination rates.

In this experiment, if we define the ‘mismatch contamination rate’ as proportion of mismatches that originate from a contaminating read, we see that the aRchaic grades of membership are approximately linearly proportional to the mismatch contamination rate, and only slightly underestimate the true proportion of contaminated reads (Fig. 3d). This encourages the use of a possible correction factor to infer a true read-level contamination rate. The fraction of a DNA library with mismatches can be measured ($\hat{f}$, e.g. $\hat{f} = 0.3$), and the fraction of modern DNA reads with mismatches to reference ($\hat{d}$, e.g. $\hat{d} = 0.09$) can be approximated using modern samples. With those values a simple moment estimator of the read level contamination rate would be: $\hat{c} = f(d/q + f - d)^{-1}$, where q is the fraction of membership in the ‘modern’ cluster. Applying this estimator to the grade of membership in the modern cluster in Figure 3c (red curve) provides values that are closer to the true contamination rates (Fig. 3c green curve). However, a complication is that high quality aDNA may generate mismatches that cluster with ‘modern’ mismatch profiles even without contamination. This effect is not accounted for by this simple estimator, and is difficult to predict.

3.3 aRchaic can identify both DNA damage and technical artifacts

As a final case study, we compiled data from 25 modern and 25 ancient Native Americans from the Northwest Coast of North America ([Lindo et al., 2016](#)). This dataset offers us an opportunity to apply aRchaic on modern and ancient DNA samples collected from the same population and sequenced in the same laboratory. In these data, the first two positions from the 5′ end of each read had been removed by the original authors in an attempt to mitigate effects of DNA damage. Despite this, aRchaic, when applied with $K=2$ to all 50 samples, clearly distinguishes between modern and ancient individuals (Supplementary Fig. S3).

When we applied aRchaic to just modern samples we were surprised to find two clear clusters (Fig. 4a). These clusters turned out to reflect the fact that the modern samples had been processed using two different library preparation kits, Nextera & TruSeq ([Lindo et al., 2016](#)). Samples prepared with the TruSeq kit showed nearly full membership in one cluster (beige), and those prepared with the Nextera kit showed nearly full membership in a second cluster (pink). These clusters show only small differences in mismatch patterns, but the pink cluster is characterized by a strong excess of mismatches at position 12, apparently an artifact introduced by the Nextera preparation (Supplementary Fig. S5).

We also applied aRchaic with $K=2$ to just the ancient samples (see Fig. 4b). Unlike the moderns, this yielded a continuous gradient of memberships in the two clusters (cyan and gold). One cluster (gold) is dominated by the strong enrichment of C-to-T mismatches relative to the modern background that is typical of ancient samples. The other cluster (cyan) is enriched primarily in C-to-A mismatches, possibly representing another type of damage, or other artifacts, in the ancient DNA. Interestingly, the individual with highest membership in this cyan cluster was much older than all the others (≈ 6000 years BP; all other samples are ≈ 2000 years BP).

Fig. 4. DNA damage and library preparation techniques drive grades of membership. **(a)** We applied aRChaic with $K=2$ to 25 modern samples from Lindo *et al.* (2016). The samples prepared with the TruSeq kit show nearly full membership in the pink cluster. Samples prepared with the Nextera kit show partial membership in the pink cluster and the tan cluster. The tan cluster shows a blip at the 12th position from the end of the read. **(b)** We applied aRChaic with $K=2$ to 25 ancient samples from Lindo *et al.* (2016). The two clusters show an enrichment of C-to-T mismatches at the ends of reads and an enrichment of purines at the 5' strand-break

4 Discussion

We developed a method (aRChaic) for clustering and visualization of samples based on DNA mismatch patterns. Our method is based on a Grade-of-Membership (GoM) model, which generalizes the concept of clustering to allow samples to have membership in multiple clusters. We provide a visual representation of the grades of membership using a 'Structure plot' (Rosenberg *et al.*, 2002) and visualization of the mismatch profiles (or clusters) with an EDLogo plot (Dey *et al.*, 2017a).

In GoM models, the choice of the number of clusters K (or mismatch profiles) is a contentious issue. In our analyses we selected values of K that highlight interpretable structure in the data. We emphasize that there will typically be no single 'true' K , and that examining results with different K can often provide additional insights (Dey *et al.*, 2017b). For example, Figure 2 shows results for $K=3$, but higher values of K reveal additional structure within each ancient subgroup (Supplementary Fig. S2). Similarly, Supplementary Figure S3 shows results for $K=2$ where the model fails to distinguish between Nextera and TruSeq modern samples, but analysis with $K=6$ does pick up this difference (Supplementary Fig. S4). More generally, aRChaic is useful for detecting batch effects as Figure 4 and Supplementary Figure S4 suggest.

A key challenge in analyzing ancient DNA is that data are often contaminated with exogenous modern DNA. Several approaches have been suggested to estimate the amount of contamination. One approach is to compute the rate of polymorphism across the X chromosome in males (Korneliusen *et al.*, 2014; Rasmussen *et al.*,

2011), where the presence of polymorphism would suggest contamination because males have only one X chromosome. Another approach is to quantify the contribution of a panel of modern mitochondrial haplotypes to the ancient DNA (Fu *et al.*, 2014; Renaud *et al.*, 2015). Additionally, much like in aRChaic, previous methods also leverage autosomal data (Racimo *et al.*, 2016) and deamination patterns to estimate contamination (Meyer *et al.*, 2014; Renaud *et al.*, 2015). Although aRChaic does not provide explicit estimates of contamination levels, in some settings (e.g. Fig. 3) the inferred grades of membership can reflect relative levels of contamination even with low sequencing depth, and may be a useful complement these other methods. Some caution is necessary though. Under conditions where an ancient sample has undergone a noticeable amount of DNA damage, we underestimate the proportion of contamination (Fig. 3). On the other hand, when ancient mismatches are indistinguishable from modern, the proportion in modern cluster of a sample might be greater than the proportion of contamination (as in the case of an ancient sample with little to no damage). Additionally, the true effects of contamination may sometimes be difficult to parse out owing to other underlying biases, like library protocol differences as in the (Fu *et al.*, 2016) data (see Supplementary Fig. S6). In such a case, we recommend adjusting for these biases before running the aRChaic model. For example, in case of the (Fu *et al.*, 2016), we separately ran the aRChaic models, with $K=2$, on two groups of aDNA samples with appreciable sample size—*dsnoUDG* (double stranded not UDG treated) and *ssUDG* (single stranded UDG treated) samples (see Supplementary Fig. S7). For the *dsnoUDG* samples, aRChaic largely separated out the moderns as a separate cluster, and the aRChaic grades of membership of the aDNA samples in the 'modern' cluster were weakly proportional to the mitochondrial estimates reported in Fu *et al.* (2016). For the *ssUDG* samples, however, the aRChaic clusters did not completely separate out the moderns from the aDNA samples and the grades of membership were no longer informative of contamination. We encourage users to keep these caveats in mind, and suggest that in practice aRChaic will be most useful for flagging potentially contaminated samples as those that have an above average clustering with modern samples.

Here we have chosen to model features at the level of mismatches, which have been shown to be informative of DNA damage in previous studies (Briggs *et al.*, 2007; Ginolhac *et al.*, 2011; Jónsson *et al.*, 2013). Alternatively, one could formulate a model at the level of reads. For example, the method PMDtools computes a score for every read representing the probability that the read is damaged (Skoglund *et al.*, 2014b). This method models mismatches along the read; additionally, one can incorporate indels and fragment length along with mismatches. One reason we chose not to model these extra features was to reduce the feature and computational complexity of our method. Furthermore, these extra features may not actually be driven by DNA damage.

Methods described here are available in an open-source R software package at www.github.com/kkdey/aRChaic.

Acknowledgements

We thank John Lindo for insightful discussions on ancient DNA and DNA damage in particular. We thank the Metlakatla and Lax Kw'alaams First Nations for access to data. The communities are located in the Prince Rupert Harbour region of British Columbia. We thank Anna Gosling, Pontus Skoglund, David Witonsky, Choongwon Jeong, Anna Di Rienzo and members of the Stephens and Novembre lab for helpful discussions.

Funding

This work was supported by National Institute of Health funding [U01CA198933 to H.A., M.S. and J.N.], [HG002585 to M.S.]; and the National Science Foundation Graduate Research Fellowship to H.A.

Conflict of Interest: none declared.

References

- 1000 Genomes Project Consortium, (2012) An Integrated Map of Genetic Variation from 1,092 Human Genomes. *Nature*, **491**, 56.
- Alexander,D.H. et al. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–1664.
- Allentoft,M.E. et al. (2015) Population Genomics of Bronze Age Eurasia. *Nature*, **522**, 167–172.
- Blei,D.M. et al. (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.
- Briggs,A.W. et al. (2007) Patterns of damage in genomic DNA sequences from a neandertal. *Proc. Natl. Acad. Sci. USA*, **104**, 14616–14621.
- Cann,H. et al. (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.
- Dey,K.K. et al. (2017a) A new sequence logo plot to highlight enrichment and depletion. *bioRxiv*, 226597.
- Dey,K.K. et al. (2017b) Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet.*, **13**, e1006599.
- Duncan,B.K. and Miller,J.H. (1980) Mutagenic deamination of cytosine residues in DNA. *Nature*, **287**, 560.
- Erosheva,E.A. (2006). *Latent Class Representation of the Grade of Membership Model*. Seattle: University of Washington.
- Fu,Q. et al. (2014) Genome sequence of a 45,000-year-old modern human from Western Siberia. *Nature*, **514**, 445–449.
- Fu,Q. et al. (2016) The genetic history of ice age Europe. *Nature*, **534**, 200.
- Gamba,C. et al. (2014) Genome flux and stasis in a five millennium transect of European Prehistory. *Nat. Commun.*, **5**, 5257.
- Ginolhac,A. et al. (2011) mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics*, **27**, 2153–2155.
- Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.
- Jónsson,H. et al. (2013) mapDamage2.0: fast approximate bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, **29**, 1682–1684.
- Korneliusson,T.S. et al. (2014) ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, **15**, 356.
- Lange,K. (1995) A Quasi-Newton acceleration of the EM algorithm. *Stat. Sin.*, **5**, 1–18.
- Lazaridis,I. et al. (2016) Genomic insights into the origin of farming in the ancient near east. *Nature*, **536**, 419.
- Lindo,J. et al. (2016) A time transect of exomes from a native american population before and after European Contact. *Nat. Commun.*, **7**, 13175.
- Lipson,M. et al. (2017) Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature*, **551**, 368.
- Malmström,H. et al. (2007) More on contamination: the use of asymmetric molecular behavior to identify authentic ancient human DNA. *Mol. Biol. Evol.*, **24**, 998–1004.
- Mathieson,I. et al. (2015) Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, **528**, 499–503.
- Mathieson,I. et al. (2018) The genomic history of southeastern Europe. *Nature*, **555**, 197–203.
- Meyer,M. et al. (2012) A high-coverage genome sequence from an Archaic Denisovan individual. *Science*, **338**, 222–226.
- Meyer,M. et al. (2014) A mitochondrial genome sequence of a Hominin from Sima De Los Huesos. *Nature*, **505**, 403.
- Olalde,I. et al. (2018) The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature*, **555**, 190–196.
- Pritchard,J.K. et al. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
- Prüfer,K. et al. (2014) The complete genome sequence of a neandertal from the Altai Mountains. *Nature*, **505**, 43.
- Racimo,F. et al. (2016) Joint estimation of contamination, error and demography for nuclear DNA from ancient humans. *PLoS Genet.*, **12**, e1005972.
- Rasmussen,M. et al. (2011) An aboriginal Australian genome reveals separate human dispersals into Asia. *Science*, **334**, 94–98.
- Renaud,G. et al. (2015) Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.*, **16**, 224.
- Rohland,N. et al. (2015) Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Phil. Trans. R. Soc. B*, **370**, 20130624.
- Rosenberg,N.A. et al. (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.
- Sawyer,S. et al. (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One*, **7**, e34131.
- Shapiro,B. and Hofreiter,M. (2014) A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science*, **343**, 1236573.
- Shen,J.-C. et al. (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.*, **22**, 972–976.
- Shiraishi,Y. et al. (2015) A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet.*, **11**, e1005657.
- Skoglund,P. et al. (2014a) Genomic diversity and admixture differs for stone-age Scandinavian foragers and farmers. *Science*, **344**, 747–750.
- Skoglund,P. et al. (2014b) Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl. Acad. Sci. USA*, **111**, 2229–2234.
- Taddy,M. (2012) On estimation and selection for topic models. In: *International Conference on Artificial Intelligence and Statistics*, pp. 1184–1193.