Received: 16 April 2020 | Revised: 10 October 2020 | Accepted: 29 October 2020

DOI: 10.1002/ece3.7056

ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

An efficient pipeline for ancient DNA mapping and recovery of endogenous ancient DNA from whole-genome sequencing data

Wenhao Xu[1,2] | Yu Lin[3,4] | Keliang Zhao[1,5] | Haimeng Li[3,6] | Yinping Tian[3] | Jacob Njaramba Ngatia[7] | Yue Ma[7] | Sunil Kumar Sahu[3] | Huabing Guo[8] | Xiaosen Guo[3,9] | Yan Chun Xu[7] | Huan Liu[3,10] | Karsten Kristiansen[3,10] | Tianming Lan[3,10] | Xinying Zhou[1,5]

[1]Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China

[2]College of Informatics, Huazhong Agricultural University, Wuhan, China

[3]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen, China

[4]Guangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Shenzhen, China

[5]CAS Center for Excellence in Life and Paleoenvironment, Beijing, China

[6]School of Future Technology, University of Chinese Academy of Sciences, Beijing, China

[7]College of Wildlife Resources, Northeast Forestry University, Harbin, China

[8]Forest Inventory and Planning Institute of Jilin Province, Changchun, China

[9]Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen, China

[10]Department of Biology, Laboratory of Genomics and Molecular Biomedicine, University of Copenhagen, Copenhagen, Denmark

Correspondence
Xinying Zhou, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, No. 142, Xizhimenwai Street, Xicheng District, Beijing 100044, China.
Email: zhouxinying@ivpp.ac.cn

Tianming Lan, State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China.
Email: lantianming@genomics.cn

Funding information
Shenzhen Municipal Government of China, Grant/Award Number: JCYJ20170817145132389; Strategic Priority Research Program of the Chinese Academy of Sciences, Grant/Award Number: grant No. XDB 26000000; Guangdong Provincial Key Laboratory of Genome Read and Write, Grant/Award Number: grant No. 2017B030301011; Fundamental Research Funds for Central Universities, Grant/Award Number: DL10DA01

Abstract

Ancient DNA research has developed rapidly over the past few decades due to improvements in PCR and next-generation sequencing (NGS) technologies, but challenges still exist. One major challenge in relation to ancient DNA research is to recover genuine endogenous ancient DNA sequences from raw sequencing data. This is often difficult due to degradation of ancient DNA and high levels of contamination, especially homologous contamination that has extremely similar genetic background with that of the real ancient DNA. In this study, we collected whole-genome sequencing (WGS) data from 6 ancient samples to compare different mapping algorithms. To further explore more effective methods to separate endogenous DNA from homologous contaminations, we attempted to recover reads based on ancient DNA specific characteristics of deamination, depurination, and DNA fragmentation with different parameters. We propose a quick and improved pipeline for separating endogenous ancient DNA while simultaneously decreasing homologous contaminations to very low proportions. Our goal in this research was to develop useful recommendations

Xu and Lin are contributed equally.

for ancient DNA mapping and for separation of endogenous DNA to facilitate future studies of ancient DNA.

KEYWORDS
ancient DNA, *BWA mem*, deamination, DNA damage, genome mapping

1 | INTRODUCTION

Ancient DNA research provides direct evidence to reconstruct prehistoric biogeographies and biodiversities, which can further help to explain long-standing questions in evolution, phylogeny, taxonomy, and adaptations (Chang et al., 2017; Delsuc et al., 2019; Palkopoulou et al., 2018; Sikora et al., 2019; Stoneking & Krause, 2011). Ancient DNA research has developed rapidly over the past thirty years due to improvements in PCR and next-generation sequencing (NGS) technologies. The first successful attempt to extract ancient DNA was made by Higuchi et al. (1984), where DNA of *Equus quagga* was extracted from muscle and DNA fragments of 228 bp were amplified (Higuchi et al., 1984; Kefi, 2011). With advancements in biomolecular techniques, it is now possible to extract and amplify ancient DNA fragments from different ancient species and biological samples, including bones, teeth, soft tissue, fur, and subfossilized excrements (Kefi, 2011; Rizzi et al., 2012). Studies on ancient DNA were previously restricted to mitochondrial DNA and extremely short nuclear DNA fragments (Dabney, Knapp, et al., 2013; Kefi, 2011). However, the advent of NGS technology has enabled ancient DNA studies at the whole-genome level. Consequently, the number of ancient DNA studies has increased exponentially in the last decade (Hofreiter et al., 2015). The first whole genome of woolly mammoth was sequenced in 2008 (Miller et al., 2008). Three Neanderthal genomes were also sequenced in 2010, revealing extensive gene flow to modern humans (Green et al., 2010). In 2012, the first high coverage genome (~30×) of Denisovans was published (Meyer et al., 2012). In 2015, Allentoft et al. (2015 sequenced 101 ancient humans at the whole-genome level (Allentoft et al., 2015). At present, more than 1,100 ancient human and hominine genomes (Marciniak & Perry, 2017) and more than 300 ancient animal genomes (Fages et al., 2019; MacHugh et al., 2017; Palkopoulou et al., 2018) have been sequenced and published.

Although great breakthroughs have been made in ancient DNA extraction, library preparation and bioinformatics, some challenges remain (Gansauge & Meyer, 2013; Rohland et al., 2018; Schubert et al., 2012; Skoglund et al., 2014). Effective mapping and distinguishing of the present-day DNA contaminations from endogenous ancient DNA are still complicated and difficult to perform, and need to be improved for ancient DNA analysis. It is particularly difficult to filter the present-day human DNA contamination from ancient human or hominine DNA (Green et al., 2009; Richards et al., 1995). Ancient DNA is often degraded into very small fragments due to physical, chemical, or biological factors during a long-term preservation in unfavorable conditions. These effects always leave valuable marks on ancient DNA to help us distinguish it from modern DNA, including C-to-T changes at the ends of ancient DNA fragments induced by deamination, high proportion of purine bases at the first physical position preceding ancient DNA fragments, and the severely fragmented nature (Skoglund et al., 2014; Stoneking & Krause, 2011).

Bioinformatics methods have been developed for mapping and separating endogenous ancient DNA from total ancient DNA (Schubert et al., 2012; Skoglund et al., 2014). In the mapping procedure for ancient DNA, the software BWA (Li & Durbin, 2009) with parameters set *"aln -l 1,024 -n 0.03"* is usually applied to map ancient sequencing data against the reference genome (Schubert et al., 2012). However, this process is time-consuming. The newly developed method BWA *mem* with the *seed-reseed-extend* algorithm, provides improved efficiencies for mapping of ancient DNA (Li, 2013). Skoglund et al. (2014) developed PMDtools to separate genuine endogenous DNA from homologous contaminations. This method is effective in filtering modern human contaminated DNA from ancient human DNA. However, it is difficult for the PMDtools to set an appropriate threshold value of PMDS when contamination rates cannot be accurately evaluated. Besides, the power of PMDtools is further weakened for extremely young or old ancient samples.

In this study, we collected whole-genome sequencing data generated by the Illumina Hiseq platform from 6 samples (representing three species) to optimize ancient DNA mapping. This step is critical to improving the mapping rate of endogenous ancient DNA. Since optimization of ancient DNA mapping may not only require filtering of present-day contaminations from endogenous ancient DNA, we used our simulated data to further explored a more universal and effective filtration pipeline to filter present-day contaminations based on ancient DNA cytosine deamination, depurination and fragmentation. The final recommendations presented here enabled reduction of modern human DNA contamination to an extremely low level while maintaining a high rate of endogenous DNA. We sought to develop mapping guidelines that, when coupled with screening recommendations to control for modern DNA contamination, could increase the effectiveness of future studies of ancient DNA.

2 | MATERIALS AND METHODS

We used simulated ancient DNA data to find more effective strategies for mapping and separating endogenous ancient DNA from homologous contaminations. To make the design clearer, we drafted a flowchart to show our overall study design (Figure 1).

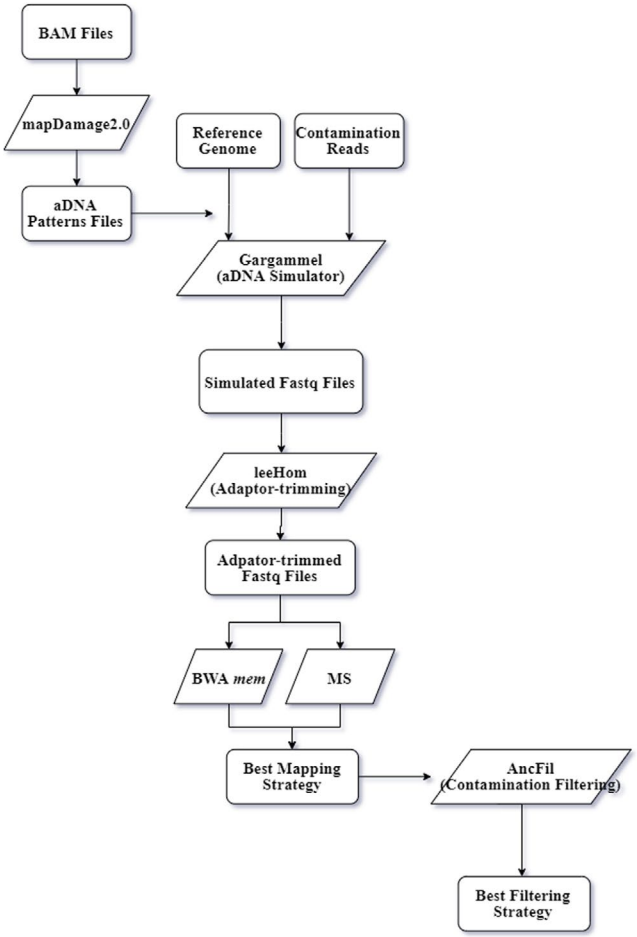

FIGURE 1 The experimental approach flow chart. Parallelogram box means the software we used

2.1 | Samples and data resource

We investigated previously sequenced whole-genome sequences from ancient animals. In total, we retrieved whole-genome sequencing (WGS) data from 6 ancient samples derived from different age groups of three species, namely four ancient humans (*Homo sapiens*) (Fu et al., 2016; Sawyer et al., 2015; Schuenemann et al., 2017), one ancient goat (Daly et al., 2018), and one ancient aurochs (Park et al., 2015). The BAM files were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/). The 6 samples were used to explore the methods for mapping and separating endogenous DNA (Table 1). The reference genomes for each species used for genome mapping are listed in Table S1.

2.2 | DNA damage analysis and ancient DNA simulation

Removing all contaminations present in real ancient data is often difficult and can lead to inaccurate evaluation. Therefore, we did not use real ancient sequencing data for analysis, but rather, we used simulated ancient sequences with the same damage parameters as those of the real data. With simulated data, it is possible to tag contamination and endogenous reads, which allows more reliable quantification of the effects. The most important thing for this ancient DNA simulation is to know the real state of the ancient DNA data we collected, especially investigating the real ancient DNA length distribution and the real proportion of deamination induced misincorporation (C-to-T and G-to-A) at ends of ancient DNA fragments. So we used mapDamage2.0 (Jonsson et al., 2013) to calculate the frequency of C-to-T and G-to-A changes at the ends of DNA fragments (misincorporation.txt) and length distribution (length_distribution.txt). To simulate real contaminations, we sequenced ancient DNA isolated from an ancient giant panda sample with ~100 years old (CNP0000732) by DIPSEQ-T1 platform, and then filtered adaptors using Trimmomatic software based on adaptor sequences (>Adapter/1:AAGTCGGAGGCCAAGCGGTCTTAGGAAGACAA;>Adapter/2:AAGTCGGATCGTAGCCATGTCGTTCTGTGAGCCAAGGAGTTG). Raw reads were then mapped by BLASTing raw reads to the nucleotide database (Lan et al., 2019) to obtain all contaminated reads. This contamination consisted of DNA from more than twenty thousand modern species, mostly consisting of bacteria, and the top 10 contaminant species have been listed in the Figure S1. We also mapped raw sequencing reads to the giant panda reference genome to identify endogenous DNA. However, the rate of endogenous DNA is extremely low (<0.001%) and therefore not included in further ancient DNA simulation of the giant panda. Real contamination data

TABLE 1 The description of samples and sequencing data used for simulating ancient DNA sequences

Species	Sample ID	Age (kyr BP)	Data sources	Reads number	Bases number	Average length of DNA fragments (bp)
Homo sapiens	JK2911	2.7	Schuenemann et al. (2017)	2.35E + 06	1.37E + 08	58.2
Homo sapiens	Villabruna	14	Fu et al. (2016)	1.22E + 07	6.70E + 08	54.9
Homo sapiens	AfontovaCava 3	17	Fu et al. (2016)	8.88E + 05	5.15E + 07	58.03
Homo sapiens	Denisova_8	>50	Sawyer et al. (2015)	8.26E + 05	3.69E + 07	44.6
Bos primigenius	British aurochs	6.7	Park et al. (2015)	7.51E + 07	3.48E + 09	46.29
Capra aegagrus hircus	Direkli5	11.5	Daly et al. (2018)	3.04E + 07	1.40E + 09	45.94

Abbreviation: BP, before present.

were then added into simulated endogenous ancient DNA to test ancient genome mapping methods. We also added modern human DNA fragments (hg38 reference genome) into simulated ancient human DNA to explore the method for filtering homologue contamination. Finally, we used gargammel (Renaud et al., 2017) (perl gargammel.pl -n 1,000,000 --comp 0,cont_rate,endo_rate -f length_distribution.txt -mapdamage misincorporation_distribution.txt single_strand/double_strand -o data/simulation data/) to simulate FASTQ files including one million reads of ancient DNA sequences for our six ancient samples. Parameters in gargammel were strictly set based on results calculated by mapDamage2.0, in order to simulate the real state of these six samples. Nine different contamination rates (cont_rate) were simulated (20%, 40%, 60%, 80%, 90%, 95%, 99%, 99.5%, and 99.9%) (Table S3). And the gargammel works as following: Step 1: Reference genome sequences were cut into different length fragments, which are in consistent with real ancient DNA data length distribution (provided by mapDamage2.0). Step 2: The reference reads are added with different DNA damage characteristics which are in consistent with DNA damage patterns of real ancient DNA data (also provided by mapDamage2.0). And, contamination reads will not be cut into different length fragments and added with DNA damage patterns as step 1 and step 2. Contamination reads can be also generated from the reference genome of some target species. Step 3: Gargammel will generate a Fastq file including simulated real ancient data and simulated contamination data. And the percentage of simulated read ancient data in the Fastq file is consistent with the parameter "--com" of gargammel (Figure 1).

2.3 | Genome mapping of simulated ancient DNA

Ancient DNA damage, especially C-to-T changes, can result in mismapping when ancient DNA fragments are mapped to reference genomes. Mapping methods and parameters used for modern DNA are not always suitable for ancient DNA (Schubert et al., 2012). We compared BWA *aln* and BWA *mem* to develop a more effective mapping strategy based on the characteristics of ancient DNA damage.

We used leeHom (Renaud et al., 2014) to trim adaptors and merge Illumina sequencing reads, and compared BWA *aln* (Version: 0.7.17) and BWA *mem* (Version: 0.7.17) to enhance mapping methods for ancient DNA. Here, "bwa *aln -l 1,024 -n* 0.03" (MS parameters) (Schubert et al., 2012) was compared with BWA *mem*. The valid mapping hits were defined as reads with endogenous ancient DNA tags (all simulated endogenous ancient DNA were tagged before mapping) and with a mapping quality higher than 30. Because it might be suitable for study of ancient DNA, the most important part of the BWA mem algorithm is the seed-reseed-extend strategy. When seeding, BWA will do exactly mapping by using part of the read length (19 bp in length when using the default parameter) on the reference genome based on FM-index algorithm. A DNA fragment in the read will be chosen as a seed when its length and number of successful matches meet thresholds the user set. Then, the seed will be used to extend both in reads and reference genome to find global match based on Smith-Waterman algorithm. This is mainly supported by two parameters including minimum seed length (parameter *-k*) and maximum seed length without reseeding (parameter *-r*). The parameter "*-k*" controls the seeding function; seeding can accelerate genome mapping. Additionally, the algorithm searches for internal seeds inside a seed longer than x bp (x=[*-k*] * [*-r*]). We tried to optimize these two parameters to further explore more efficient mapping parameters for ancient DNA mapping. We tested BWA *mem* with *-k* (9/14/19/24/29) and *-r* (0.5/1/1.5/2/2.5) parameters. To evaluate mapping effectiveness, we defined three main criteria: (1) CRT: the contamination rate after treatment (the number of mapped contamination reads/ the number of mapped reads); (2) LRE: the loss rate of endogenous DNA (the number of unmapped endogenous ancient reads/the number of endogenous ancient reads); (3) MT: the running time of mapping.

2.4 | Separating endogenous DNA from the contaminations

The unique ancient DNA characteristics, especially C-to-T and/or G-to-A changes at ends of DNA fragments help to improve filtering of contaminated present-day DNA. We wrote a program named AncFil using Python (home page: https://github.com/tianmingla n/AncFil) to explore a more universal and effective pipeline for separating endogenous ancient DNA from homologous contaminations. We first screened reads with at least "DeamNum" C-to-T or G-to-A mutations within the first or last "DetectRange" base pair at 3' and/or 5' ends ("DoubleOrSingle"). For "DeamNum" (the number of C-to-T or G-to-A mutations), we tested one, two and three. For "DetectRange" (the base number), we tested five, ten and fifteen, while for "DoubleOrSingle," either 3' or 5' end (parameter "or") and both ends (parameter "and") were included. We explored all 18 possible screening conditions by adjusting the parameter combinations ("DeamNum," "DetectRange," "DoubleOrSingle") (Table S2). One can test more possible conditions by adjusting parameters "-DeamNum," "-DetectRange," and "-DoubleOrSingle." Given that there is a natural tendency toward depurination at the 5' ends of ancient DNA fragments (Briggs et al., 2007), we screened reads with an A or G at the position preceding the first base of the 5' end. Finally, we evaluated the effect of fragment length of ancient DNA on the separation of endogenous DNA. Here, two criteria were used to evaluate this pipeline: (1) CRT: the contamination rate after treatment (the number of contamination reads after filtering/the number of reads after filtering); (2) LRE: the loss rate of true endogenous DNA (the number of filtered endogenous ancient reads/the number of endogenous ancient reads before filtering);

Finally, PMDtools (Skoglund et al., 2014) were used to filter homologous contaminations using the same data and evaluating criteria used to evaluate our recommended method above. Meanwhile, "-threshold" is one of the most important parameters in PMDtools for adjusting the strictness of the filtration. To make a complete comparison, we tested five threshold values (one, two, three, four, and five) to adjust the PMD scores by setting "-threshold."

3 | RESULTS

3.1 | Description of samples and the simulated data

We simulated a total of 90 ancient DNA datasets (Table S3) containing the same length distribution and damage patterns as the real dataset (Table 1). One million reads were finally simulated under each condition. The average length of ancient DNA data that was collected ranged from 45 bp to 58 bp (Table 1). The length distributions for most ancient DNA datasets ranged from 30 bp to 70 bp (Figure S2). The sample ages ranged from ~ 2.7 kyr BP (Before Present) to ~50 kyr BP, which provided a good basis to evaluate the influence of age on ancient DNA mapping and separation of homologous contaminations. The DNA damage analysis showed an obvious increase of deaminated substitutions with the frequency of C-to-T and G-to-A at the ends of DNA fragments ranging from 2% to 80% (Figure S3). Samples collected in our study included ancient DNA of different conditions, which enabled us to draw conclusions suitable for most ancient DNA samples.

3.2 | Comparing different mapping algorithms on ancient DNA

(a) CRT: the contamination rate after treatment (the number of mapped contamination reads/ the number of mapped reads); (b) LRE: the loss rate of endogenous DNA (the number of unmapped endogenous ancient reads/ the number of endogenous ancient reads); (c) MT: the running time of mapping.

The comparison was achieved by calculating the contamination rate after treatment (CRT), the loss rate of endogenous DNA (LRE), and the running time of mapping (MT) for each dataset, and performing Repeated Measurement Analysis of Variance. The average and median values of CRT, LRE, and MT of the two algorithms with different contamination rates are shown in Table S4. The analysis showed no significant differences in CRT ($F = 1.42$, $p = .2870$) and LRE ($F = 0.44$, $p = .5344$). However, significant differences were found in MT ($F = 41.57$, $p = .0013$) (Table S5) and BWA *aln* with the MS parameter requiring a multiple of 7.13 more times than BWA *mem* by default. We further evaluated the influence of different samples and different contamination rates on ancient DNA mapping. As shown in Table S6 and Figure 2, CRT levels were unchanged across different samples. The mean values of LRE were stable between contamination rates but not when the contamination rate was close to 100%.

We calculated CRT, LRE, and MT using different parameters *-k* (9/14/19/24/29) (Figure 3) and performed Repeated Measurement Analysis of Variance Analysis to compare the results generated under different parameters. There were significant differences in CRT ($F = 644.61$, $p < .0001$), LRE ($F = 17.99$, $p = .0057$), and MT ($F = 146.75$, $p < .0001$) (Table S7). LRE was highest at *-k* = 29, and it increased as the value of k increased. LRE value decreased by ~0.20% from *-k* = 19 to *-k* = 9; however, this decrease continued to ~4.66% from *-k* = 29 to *-k* = 19, which was 22.3 times larger than that between *-k* = 19 and *-k* = 9 (Table S8). In addition, the running time significantly decreased from *-k* = 9 to *-k* = 19, but was relatively stable and slightly longer when *-k* was larger than 19 (Table S8, Figure 3).

We evaluated the parameter *-r* (0.5/1/1.5/2/2.5) with the same method used to evaluate the parameter *-k* (Figure 4). Significant differences were found in CRT ($F = 392.45$, $p < .0001$), LRE ($F = 45.11$, $p = .0010$), and MT ($F = 9.19$, $p = .0002$) (Table S9). A significant decrease in LRE was recorded when the set values of "*-r*" were greater than 1.5 and LRE reached the lowest level at *-r* = 2.5 (Table S10). Furthermore, the running time significantly decreased from *-r* = 0.5 to *-r* = 1.5, but it was relatively stable and slightly longer when *-r* was greater than 1.5 (Table S10, Figure 4).

3.3 | Separation of endogenous DNA

Using unique ancient DNA characteristics, the homologous contamination rate was reduced to a very low level (Figure 5). The mean values of CRT and LRE are also shown in Table S11. No significant differences were found in CRT ($F = 3.27$, $p = .1097$) and LRE ($F = 1.11$, $p = .3893$) (Table S12).

When testing the influence of parameter "DeamNum" on separation of endogenous DNA, significant differences were found in CRT ($F = 26.01$, $p = .0011$) and LRE ($F = 24.03$, $p = .0152$) (Table S13). An increase in "DeamNum" resulted in lower values for CRT, but higher values for LRE (Figure S4). We also calculated CRT and LRE to evaluate the influence of the parameter "DoubleOrSingle" on ancient DNA mapping (Figure S5). Significant difference was found in values of CRT ($F = 44.97$, $p = .0068$) but not in LRE ($F = 7.20$, $p = .0748$) (Table S14). The result showed a decline of 90.27% in CRT when screening the reads with C-to-T or G-to-A on single end (-DoubleOrSingle = or) compared to screening on both 3' and 5' ends (-DoubleOrSingle = and) (Table S11). The homologous contamination rate was held to an average of 0.92% by using the filtering strategy with "-DetectRange = 15 -DeamNum = 1 -DoubleOrSingle = or."

We compared our method with PMDtools software. These two methods were run in parallel using the same dataset. The results generated by PMDtools with different parameters are shown in Table S15. To make the comparison fairer, LRE values of the tested dataset were kept similar in both our pipeline and PMDtools. CRT values did not differ ($Z = -1.171$, $p = .241$) between the two methods. However, the running time of our method was 15.43% of the runtime for PMDtools, and the difference was significant (Difference = 2.3mins, $Z = -6.50$, $p = 8.28E^{-11}$). Although this comparison does not show that our tool outperforms PMDtools, it does demonstrate a fast and reliable complement to PMDtools.

We tried to screen reads with G or A residues preceding the first base at 5' end. The average homologous contamination rate was 2.25% after filtering using depurination characteristic.

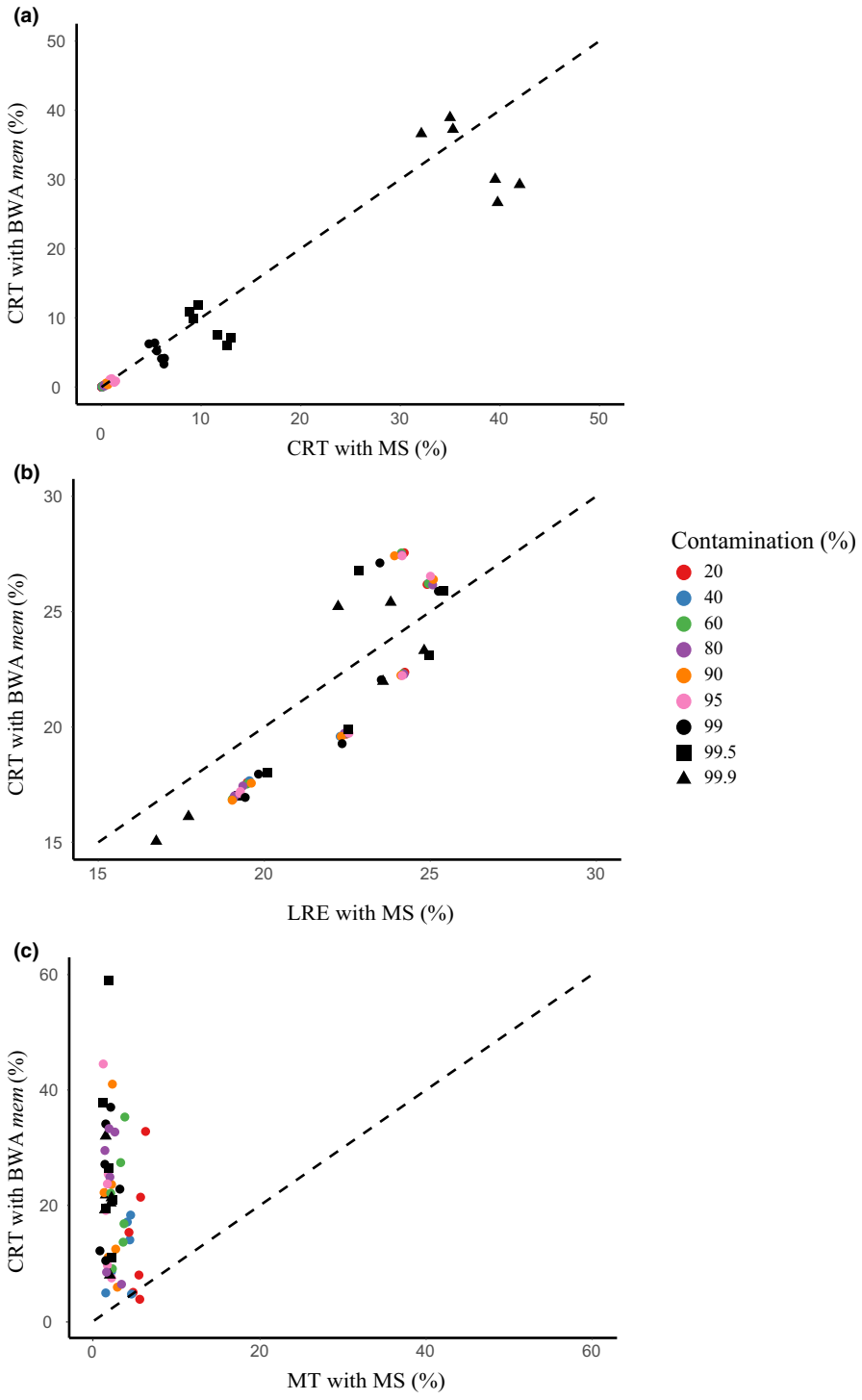

FIGURE 2 The differences between BWA *aln* with MS parameters and BWA *mem* with default parameters. (a) Comparison of CRT. (b) Comparison of LRE. (c) Comparison of MT

4 | DISCUSSION

4.1 | Comparing BWA *aln* and BWA *mem* to improve ancient DNA mapping

BWA *aln* uses backward search for exact matching and its seeding function allows differences in the first few tens of base pairs on a read to search inexact matching (Li & Durbin, 2009). This can accelerate genome mapping but it also increases the probability of incorrect alignments. Therefore, disabling the seed function often tends to be more effective in ancient DNA mapping (Schubert et al., 2012). BWA *mem*, however, uses a re-seeding strategy to increase correct alignments when no maximal exact matches (SMEMs) can be found (Li, 2013). This could compensate the shortcoming of BWA *aln* mentioned above. In our experiment, no significant difference was found in CRT (F = 1.42, p = .2870) between Schubert's method (Schubert et al., 2012) and the BWA *mem* algorithm. Consequently, the performance on ancient DNA mapping of BWA *mem* with default parameters (BWA *mem* -*k* 19 -*r* 1.5) was comparable to BWA *aln* with the MS parameters.

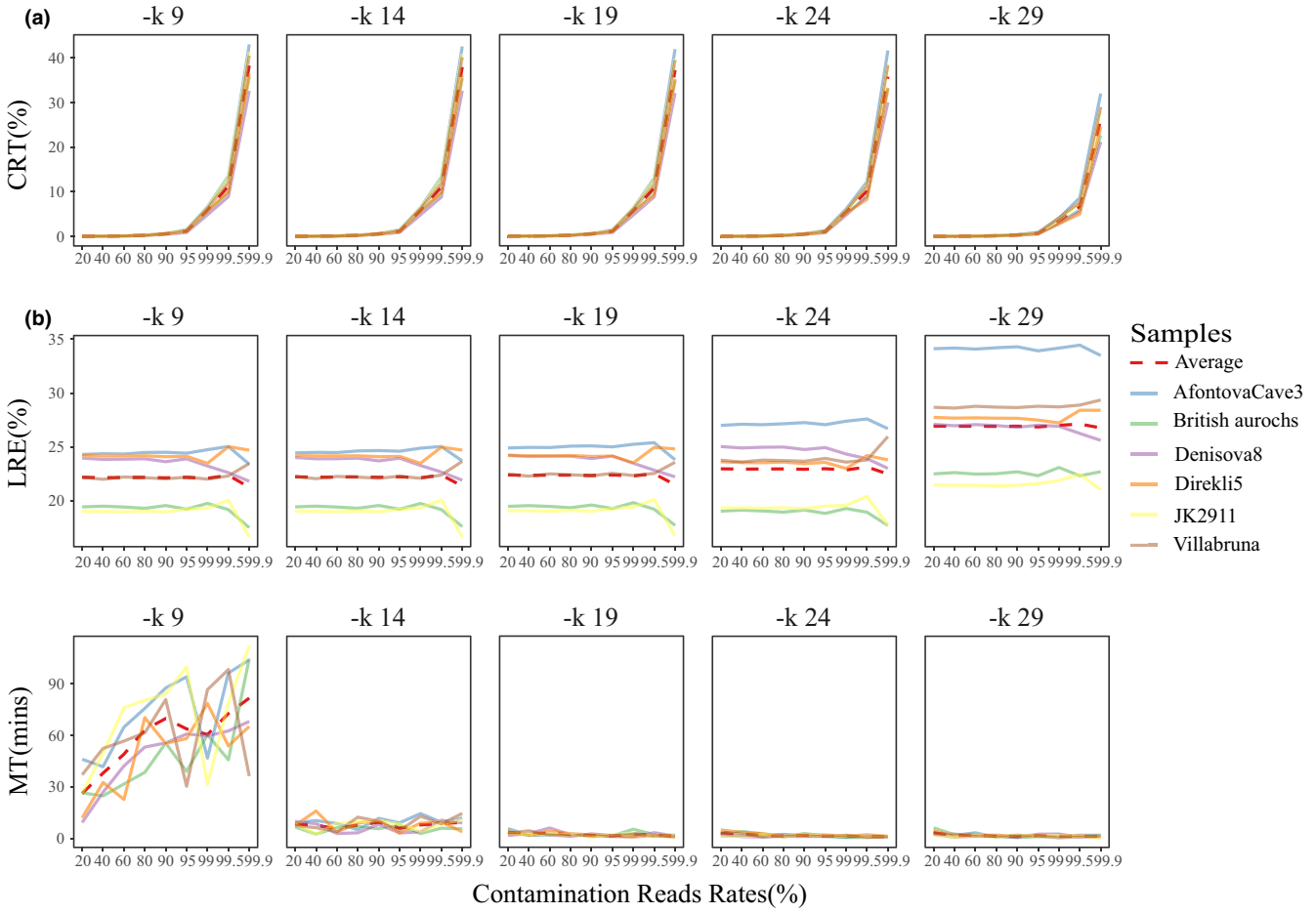

FIGURE 3 Comparison of BWA mem results with "-*k*" parameters. (a) Comparison of CRT. (b) Comparison of LRE. (c) Comparison of MT

Additionally, the seed-reseed-extend strategy in BWA *mem* can help to accelerate the mapping process (Li & Durbin, 2009), and it resulted in a 87.70% decrease of MT compared to the BWA *aln* algorithm. Therefore, BWA *mem* can improve the accuracy of ancient genome mapping in a shorter time than that required for analysis using BWA *aln*.

Soft clipping (Langmead & Salzberg, 2012) means that some nucleotides at either terminal of the reads can be omitted as determined by the mapping scoring scheme. And it's one of most important issues to consider when using BWA *mem*. In our study, 7.9% of mapped reads were soft clipped during mapping and 6% of soft-clipped reads contained C-to-T and/or G-to-A changes within soft-clipped regions. In other words, only ~0.47% (7.9%*0.6%) mapped reads with damaged patterns were soft clipped, which was a small proportion when considering the large number of damaged endogenous DNA. With regard to hard clipping (Langmead & Salzberg, 2012), this means that some nucleotides at either terminal of reads can be omitted as determined by the mapping scoring scheme but the omitted nucleotides do not exist in the fragment. This is a special kind of soft clipping to mark the multiple mapping of a read. But only 0.0036% of mapped reads showed damaged patterns. Therefore, soft clipping only slightly impacted the filtering of endogenous DNA by using deamination characteristics. In summary, BWA *mem* performed as well as BWA *aln* with MS parameters in this study, but BWA *mem* required less running time (87.70% time) than did the BWA *aln* method. Taking all results into account, BWA *mem* performed better than BWA *aln*.

4.2 | Exploring more accurate and effective mapping parameters of BWA *mem*

The parameters *-k* and *-r* are extremely important for the "seeding and reseeding" mapping stages in BWA *mem* (Li, 2013). The different parameter values of *-k* and *-r* could significantly affect CRT, LRE, and MT, indicating that we can obtain ancient DNA mapping results with a lower contamination rate by optimizing these parameter values (Tables S7, S9).

The BWA *mem* algorithm only found the maximal exact matches (SMEMs) in a read while seeding and this algorithm can trigger re-seeding with SMEMs to reduce the loss of mis-mapping if SMEMs are larger than [*-k*-r*] (Li, 2013). The large [*-k*-r*] values meant fewer re-seedings, which could accelerate the mapping process. This was consistent with the observation of runtime results. However, too long seeds could also make seed mapping against genomes more difficult and eventually more time-consuming.

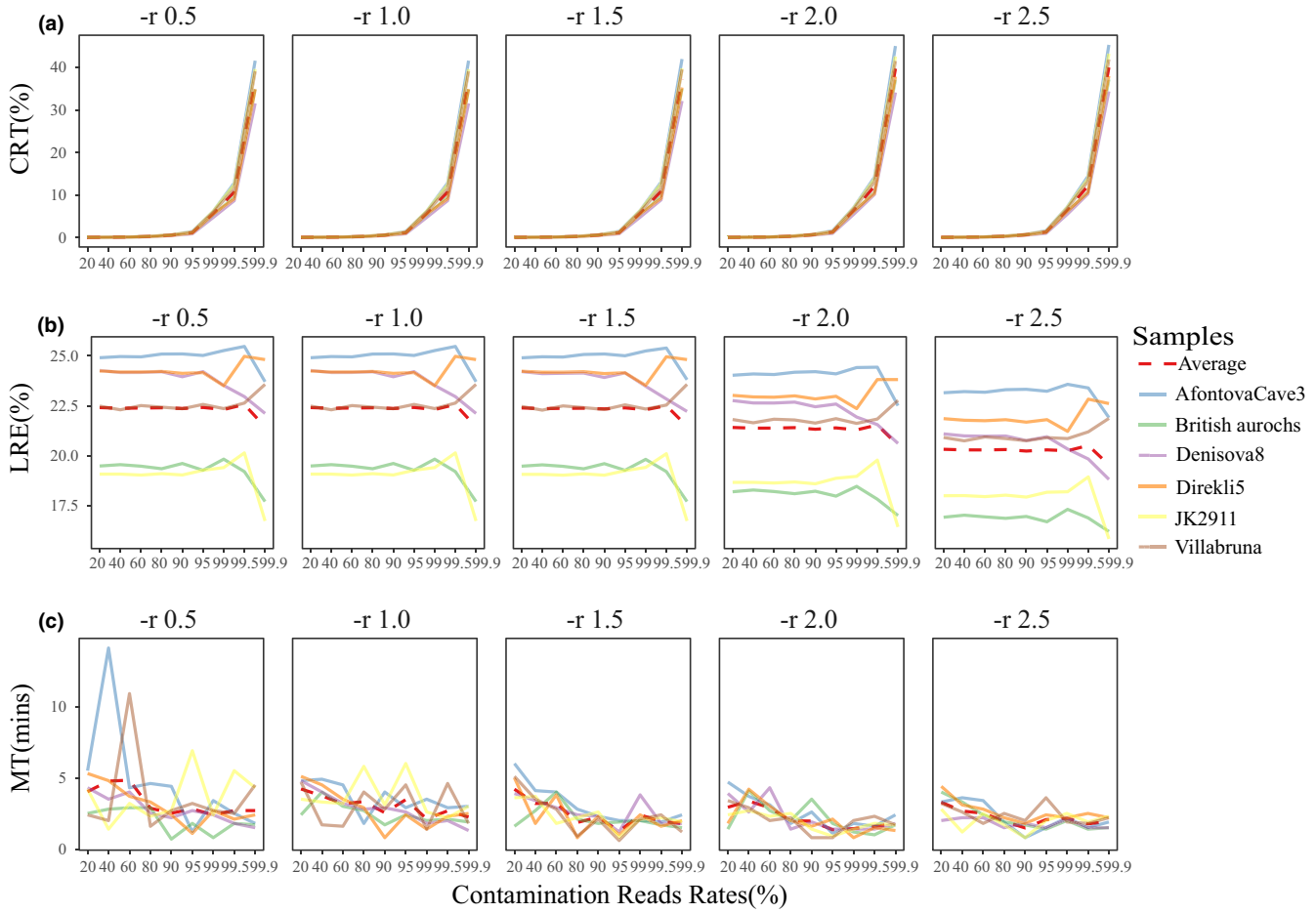

FIGURE 4 Comparison of "BWA mem" results with "-*r*" parameters. (a) Comparison of CRT. (b) Comparison of LRE. (c) Comparison of MT

We also found that running time was more sensitive to changes in -*k* parameter than in -*r* (Table S8, Table S10, Figure 3, Figure 4), indicating that running time was mainly influenced by minimum seed length. The -*r* cannot affect seeding for SMEMs, but -*k* can influence both seeding and reseeding procedures (Li, 2013), which might be reason for their differing influence on running time. Finally, it took minimum runtime to save endogenous ancient DNA reads as much as possible when using "BWA mem -*k* = 19 -*r* = 2.5" for mapping of ancient DNA.

4.3 | Improving the separation of endogenous DNA

Among all kinds of homologous contaminations, the present-day human DNA is a very common contamination in ancient human DNA. This is because contaminations can easily be induced from the time samples are collected to the time DNA library preparation is performed. These homologous contaminations are extremely difficult to remove (Skoglund et al., 2014).

In our testing, the proportion of homologous contamination that could be removed from the simulated raw data decreased with increase in simulated contamination rates, and there was a significant negative correlation between them (R^2 = 0.391, p = .019). However, it remained possible to remove > 99% of homologous contaminations even when the simulated contamination rate reached 99% (Figure 6). It was notable that 99.9% homologous contamination was removed when the simulated contamination rate was only 95%. On average, 99.07% of contamination could be removed using our recommended screening method (Table S16), which was lower than that reported by many other ancient DNA studies (Sawyer et al., 2015; Schuenemann et al., 2017). No significant differences were found in endogenous DNA rates considering the different samples, different damage patterns, and different contamination rates: This demonstrated the universal property of our recommended method. Using the remaining endogenous ancient reads, we summarized a best combination with DeamNum = 1, DetectRange = 15, and DoubleOrSingle = or.

To test a potentially more effective filtering strategy, we further screened reads with G or A residues preceding the first base at 5' end of the DNA fragments. This depurination screening decreased the homologous contamination rate to 2.25% (the initial contamination rates were from 20% to 99.9%), which meant that this method enables recovery of more endogenous DNA (Table S17). Similar to deamination screening, filtering effect showed no difference in relation to sample ages, which was largely due to weak correlation between depurination and samples ages. Sample age and the extent

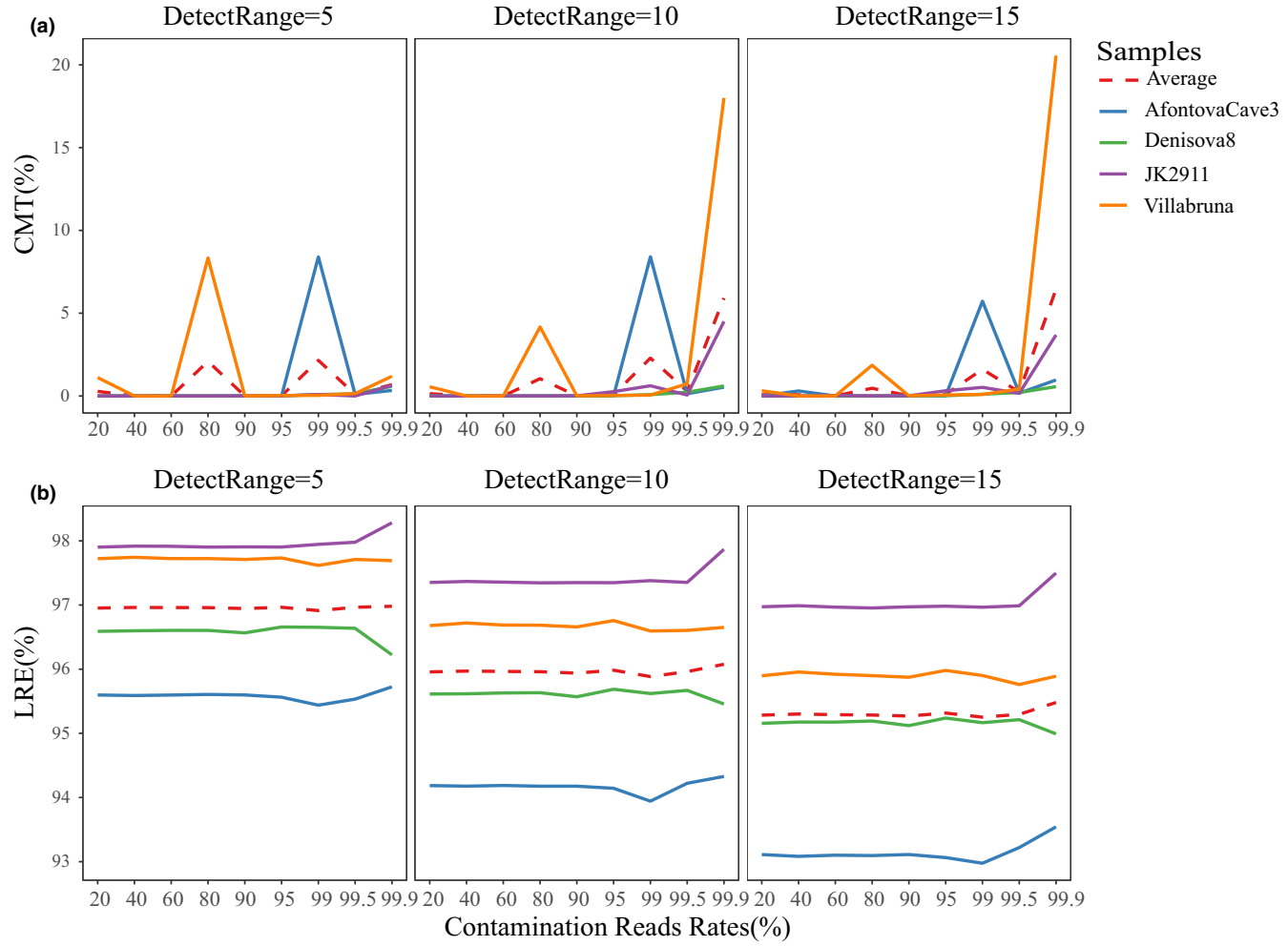

FIGURE 5 Comparison of deamination filtering with "-DetectRange" parameters. (a) Comparison of CRT after filtering. (b) Comparison of LRE after filtering

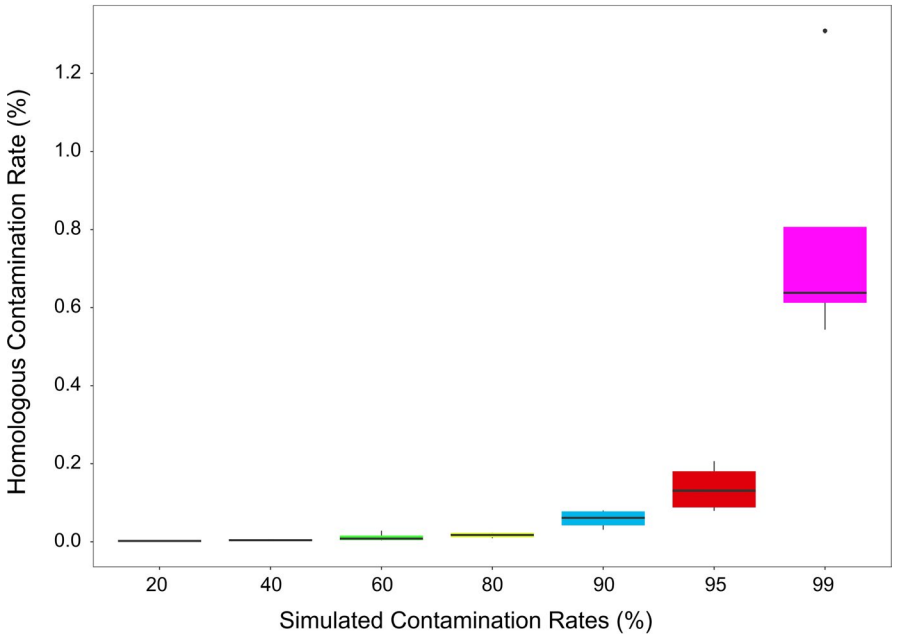

FIGURE 6 The homologous contamination rate after filtering by use of the AncFil with parameters –DeamNum = 15 –DetectRange = 1 –DoubleOrSingle = or. X-axis means the rate of homologous contamination which was added in simulation data. Y-axis means the rate of homologous contamination which was remaining after filtering

of DNA fragmentation were not significantly correlated. DNA fragments are usually heavily degraded due to depurination shortly after death (Dabney, Meyer et al., 2013; Sawyer et al., 2012). However, only 10%–40% of ancient DNA fragmentation is triggered by depurination although other factors can also result in DNA fragmentation. As such, it is difficult to identify more endogenous ancient reads by screening the DNA length. However, this has also been provided in our python script to support filtration by depurination and fragmentation as week filtering options (not recommended).

5 | CONCLUSION

We found that BWA *mem* with the parameters *-k* = 19 and *-r* = 2.5 was comparable to BWA *aln* with MS parameters (Schubert et al., 2012) when considering the recovery of ancient DNA, but had a significantly shorter running time than did BWA *aln* with MS parameters. For the recovery of endogenous DNA from ancient sequencing data with homologous contaminations, we recommend screening of reads with parameters: –DeamNum = 1, –DetectRange = 15, and –DoubleOrSingle = or, which could remove more than 99% of homologous DNA contaminations from the raw contaminated sequencing data. Overall, these recommendations for ancient DNA mapping and separation of endogenous DNA can benefit ancient DNA studies, especially for samples preserved under poor conditions.

ACKNOWLEDGMENTS

This study was supported by the Fundamental Research Funds for Central Universities (DL10DA01), the Strategic Priority Research Program of the Chinese Academy of Sciences (grant No. XDB 26000000, the Shenzhen Municipal Government of China (grant No. JCYJ20170817145132389) and the Guangdong Provincial Key Laboratory of Genome Read and Write (grant No. 2017B030301011). We thank the Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Guangdong, China. Finally, we are thankful to the China National GenBank for producing the sequencing data. We thank Thomas D. Dahmer for helpful discussion and comments on the manuscript, Xiaohui Liu for giving suggestions on data analysis, Chenhui Liu and Dan Xiong for helping plot figures.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTION

Wenhao Xu: Investigation (equal); Methodology (equal); Software (lead); Visualization (equal); Writing-original draft (lead). **Yu Lin:** Validation (lead). **Keliang Zhao:** Conceptualization (equal). **Haimeng Li:** Validation (equal). **Yinpin Tian:** Data curation (equal). **Jacob Njaramba Ngatia:** Validation (equal); Writing-original draft (equal); Writing-review & editing (equal). **Yue Ma:** Validation (equal). **Sunil Kumar Sahu:** Writing-review & editing (equal). **Huabing Guo:** Investigation (equal). **Xiaosen Guo:** Validation (equal). **Yanchun Xu:** Conceptualization (equal). **Huan Liu:** Conceptualization (equal). **Karsten Kristiansen:** Conceptualization (equal). **Tianming Lan:** Conceptualization (equal); Data curation (equal); Investigation (equal); Methodology (equal); Software (equal); Supervision (equal); Visualization (equal); Writing-original draft (equal); Writing-review & editing (equal). **Xinying Zhou:** Conceptualization (equal); Supervision (equal).

DATA AVAILABILITY STATEMENT

Raw sequencing data of the ancient panda have been deposited to the CNSA (CNGB Nucleotide Sequence Archive) with accession number CNP0000732 (https://db.cngb.org/cnsa/).

ORCID

Wenhao Xu https://orcid.org/0000-0002-3919-1030

REFERENCES

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlström, T., Vinner, L., Malaspinas, A.-S., Margaryan, A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Casa, P. D., Dąbrowski, P., … Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*(7555), 167–172. https://doi.org/10.1038/nature14507

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prufer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., & Paabo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, *104*(37), 14616–14621. https://doi.org/10.1073/pnas.0704665104

Chang, D., Knapp, M., Enk, J., Lippold, S., Kircher, M., Lister, A., MacPhee, R. D. E., Widga, C., Czechowski, P., Sommer, R., Hodges, E., Stümpel, N., Barnes, I., Dalén, L., Derevianko, A., Germonpré, M., Hillebrand-Voiculescu, A., Constantin, S., Kuznetsova, T., … Shapiro, B. (2017). The evolutionary and phylogeographic history of woolly mammoths: A comprehensive mitogenomic analysis. *Scientific Reports*, *7*(1), 44585. https://doi.org/10.1038/srep44585

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., Garcia, N., Paabo, S., Arsuaga, J.-L., & Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*, *110*(39), 15758–15763. https://doi.org/10.1073/pnas.1314445110

Dabney, J., Meyer, M., & Paabo, S. (2013). Ancient DNA Damage. *Cold Spring Harbor Perspectives in Biology*, *5*(7), a012567. https://doi.org/10.1101/cshperspect.a012567

Daly, K. G., Maisano Delser, P., Mullin, V. E., Scheu, A., Mattiangeli, V., Teasdale, M. D., Hare, A. J., Burger, J., Verdugo, M. P., Collins, M. J., Kehati, R., Erek, C. M., Bar-Oz, G., Pompanon, F., Cumer, T., Çakırlar, C., Mohaseb, A. F., Decruyenaere, D., Davoudi, H., … Bradley, D. G. (2018). Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. *Science*, *361*(6397), 85–88. https://doi.org/10.1126/science.aas9411

Delsuc, F., Kuch, M., Gibb, G. C., Karpinski, E., Hackenberger, D., Szpak, P., Martínez, J. G., Mead, J. I., McDonald, H. G., MacPhee, R. D. E., Billet, G., Hautier, L., & Poinar, H. N. (2019). Ancient mitogenomes reveal the evolutionary history and biogeography of sloths. *Current Biology*, *29*(12), 2031–2042. https://doi.org/10.1016/j.cub.2019.05.043

Fages, A., Hanghøj, K., Khan, N., Gaunitz, C., Seguin-Orlando, A., Leonardi, M., Constantz, C. M. C., Gamba, C., Al-Rasheid, K. A. S., Albizuri, S., Alfarhan, A. H., Allentoft, M., Alquraishi, S., Anthony, D., Baimukhanov, N., Barrett, J. H., Bayarsaikhan, J., Benecke, N., Bernáldez-Sánchez, E., … Orlando, L. (2019). Tracking five millennia

of horse management with extensive ancient genome time series. *Cell*, *177*(6), 1419–1435.

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A., Nickel, B., Peltzer, A., Rohland, N., Slon, V., Talamo, S., Lazaridis, I., Lipson, M., Mathieson, I., Schiffels, S., … Reich, D. (2016). The genetic history of Ice Age Europe. *Nature*, *534*(7606), 200–205. https://doi.org/10.1038/nature17993

Gansauge, M. T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, *8*(4), 737–748. https://doi.org/10.1038/nprot.2013.038

Green, R. E., Briggs, A. W., Krause, J., Prüfer, K., Burbano, H. A., Siebauer, M., Lachmann, M., & Pääbo, S. (2009). The Neandertal genome and ancient DNA authenticity. *The EMBO Journal*, *28*(17), 2494–2502. https://doi.org/10.1038/emboj.2009.222

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., … Paabo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., & Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature*, *312*(5991), 282–284. https://doi.org/10.1038/312282a0

Hofreiter, M., Paijmans, J. L. A., Goodchild, H., Speller, C. F., Barlow, A., Fortes, G. G., Thomas, J. A., Ludwig, A., & Collins, M. J. (2015). The future of ancient DNA: Technical advances and conceptual shifts. *BioEssays*, *37*(3), 284–293. https://doi.org/10.1002/bies.201400160

Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., & Orlando, L. (2013). mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, *29*(13), 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Kefi, R. (2011). Ancient DNA investigations: A review on their significance in different research fields. *International Journal of Modern Anthropology*, *1*(4), 2. https://doi.org/10.4314/ijma.v1i4.4

Lan, T., Lin, Y. U., Njaramba-Ngatia, J., Guo, X., Li, R., Li, H., Kumar-Sahu, S., Wang, X., Yang, X., Guo, H., Xu, W., Kristiansen, K., Liu, H., & Xu, Y. (2019). Improving Species Identification of Ancient Mammals Based on Next-Generation Sequencing Data. *Genes*, *10*(7), 509. https://doi.org/10.3390/genes10070509

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

MacHugh, D. E., Larson, G., & Orlando, L. (2017). Taming the Past: Ancient DNA and the Study of Animal Domestication. *Annual Review of Animal Biosciences*, *5*(1), 329–351. https://doi.org/10.1146/annurev-animal-022516-022747

Marciniak, S., & Perry, G. H. (2017). Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics*, *18*(11), 659–674. https://doi.org/10.1038/nrg.2017.65

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prufer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Paabo, S. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, *338*(6104), 222–226. https://doi.org/10.1126/science.1224344

Miller, W., Drautz, D. I., Ratan, A., Pusey, B., Qi, J. I., Lesk, A. M., Tomsho, L. P., Packard, M. D., Zhao, F., Sher, A., Tikhonov, A., Raney, B., Patterson, N., Lindblad-Toh, K., Lander, E. S., Knight, J. R., Irzyk, G. P., Fredrikson, K. M., Harkins, T. T., … Schuster, S. C. (2008). Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, *456*(7220), 387–390. https://doi.org/10.1038/nature07446

Palkopoulou, E., Lipson, M., Mallick, S., Nielsen, S., Rohland, N., Baleka, S., Karpinski, E., Ivancevic, A. M., To, T.-H., Kortschak, R. D., Raison, J. M., Qu, Z., Chin, T.-J., Alt, K. W., Claesson, S., Dalén, L., MacPhee, R. D. E., Meller, H., Roca, A. L., … Reich, D. (2018). A comprehensive genomic history of extinct and living elephants. *Proceedings of the National Academy of Sciences*, *115*(11), E2566–E2574. https://doi.org/10.1073/pnas.1720554115

Park, S. D. E., Magee, D. A., McGettigan, P. A., Teasdale, M. D., Edwards, C. J., Lohan, A. J., Murphy, A., Braud, M., Donoghue, M. T., Liu, Y., Chamberlain, A. T., Rue-Albrecht, K., Schroeder, S., Spillane, C., Tai, S., Bradley, D. G., Sonstegard, T. S., Loftus, B. J., & MacHugh, D. E. (2015). Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. *Genome Biology*, *16*(1), 234. https://doi.org/10.1186/s13059-015-0790-2

Renaud, G., Hanghoj, K., Willerslev, E., & Orlando, L. (2017). gargammel: A sequence simulator for ancient DNA. *Bioinformatics*, *33*(4), 577–579.

Renaud, G., Stenzel, U., & Kelso, J. (2014). leeHom: Adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Research*, *42*(18), e141. https://doi.org/10.1093/nar/gku699

Richards, M. B., Sykes, B. C., & Hedges, R. E. M. (1995). Authenticating DNA extracted from ancient skeletal remains. *Journal of Archaeological Science*, *22*(2), 291–299. https://doi.org/10.1006/jasc.1995.0031

Rizzi, E., Lari, M., Gigli, E., De Bellis, G., & Caramelli, D. (2012). Ancient DNA studies: New perspectives on old samples. *Genetics Selection Evolution*, *44*(1), 21. https://doi.org/10.1186/1297-9686-44-21

Rohland, N., Glocke, I., Aximu-Petri, A., & Meyer, M. (2018). Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. *Nature Protocols*, *13*(11), 2447–2461. https://doi.org/10.1038/s41596-018-0050-5

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., & Paabo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One*, *7*(3), e34131. https://doi.org/10.1371/journal.pone.0034131

Sawyer, S., Renaud, G., Viola, B., Hublin, J.-J., Gansauge, M.-T., Shunkov, M. V., Derevianko, A. P., Prüfer, K., Kelso, J., & Pääbo, S. (2015). Nuclear and mitochondrial DNA sequences from two Denisovan individuals. *Proceedings of the National Academy of Sciences*, *112*(51), 15696–15700. https://doi.org/10.1073/pnas.1519905112

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., AL-Rasheid, K. A. S., Willerslev, E., Krogh, A., & Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, *13*(1), 178. https://doi.org/10.1186/1471-2164-13-178

Schuenemann, V. J., Peltzer, A., Welte, B., van Pelt, W. P., Molak, M., Wang, C.-C., Furtwängler, A., Urban, C., Reiter, E., Nieselt, K., Teßmann, B., Francken, M., Harvati, K., Haak, W., Schiffels, S., & Krause, J. (2017). Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nature Communications*, *8*(1), 15694. https://doi.org/10.1038/ncomms15694

Sikora, M., Pitulko, V. V., Sousa, V. C., Allentoft, M. E., Vinner, L., Rasmussen, S., Margaryan, A., de Barros Damgaard, P., de la Fuente, C., Renaud, G., Yang, M. A., Fu, Q., Dupanloup, I., Giampoudakis, K., Nogués-Bravo, D., Rahbek, C., Kroonen, G., Peyrot, M., McColl, H., … Willerslev, E. (2019). The population history of northeastern Siberia since the Pleistocene. *Nature*, *570*(7760), 182–188. https://doi.org/10.1038/s41586-019-1279-z

Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., & Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, *111*(6), 2229–2234. https://doi.org/10.1073/pnas.1318934111

Stoneking, M., & Krause, J. (2011). Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics*, *12*(9), 603–614. https://doi.org/10.1038/nrg3029

SUPPORTING INFORMATION

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

How to cite this article: Xu W, Lin Y, Zhao K, et al. An efficient pipeline for ancient DNA mapping and recovery of endogenous ancient DNA from whole-genome sequencing data. *Ecol Evol*. 2021;11:390–401. https://doi.org/10.1002/ece3.7056