An Efficient Pipeline for Ancient DNA Mapping and Recovery of Endogenous Ancient DNA from Whole-Genome Sequencing Data

Wenhao Xu[1], Yu Lin[2], Ke Zhao[1], Hai Li[2], Yin Tian[2], Jacob Ngatia[3], Yue Ma[3], Hua Guo[4], Xiao Guo[2], Yan Xu[3], Karsten Kristiansen[2], Tian Lan[2], and Xin Zhou[1]

[1]Chinese Academy of Sciences
[2]BGI-Shenzhen
[3]Northeast Forestry University
[4]Forest Inventory and Planning Institute of Jilin Province

April 28, 2020

Abstract

Ancient DNA research has developed rapidly over the past few decades due to the improvement 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 the raw sequencing data. This is often difficult due to the degradation of ancient DNA and high levels of contamination, especially homologous contamination. 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 the homologous contaminations, we attempted to recover reads based on the 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 the homologous contaminations to a very low proportion. Overall, these recommendations for ancient DNA mapping and separation of endogenous DNA in this study could facilitate future studies of ancient DNA.

1. Introduction

Ancient DNA research provides direct evidence to reconstruct the prehistoric biogeography and biodiversity, which can further help to explain long-standing questions in evolution, phylogeny, taxonomy, and adaptations (1-5). Ancient DNA research has developed rapidly over the past thirty years due to the improvement in PCR and next-generation sequencing (NGS) technologies. The first successful attempt to extract ancient DNA was made by Higuchi et al (1984), where muscle DNA of*Equus quagga* was extracted and DNA fragments of 228 bp were amplified (6, 7). With the advancement 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 fossilized excrements (7, 8). Studies on ancient DNA were previously restricted to mitochondrial DNA and extremely short nuclear DNA fragments (7, 9). 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(10). The first whole genome of the woolly mammoth was sequenced in 2008(11). Three Neanderthals' genomes were also sequenced in 2010, which revealed an extensive gene flow to modern humans (12). In 2012, the first high coverage genome (˜30X) of Denisovans was published(13). In 2015, Allentoft et al. (2015) sequenced 101 ancient humans at the whole genome level (14). At present, more than 1100 ancient human and hominine genomes (15) and more than 300 ancient animal genomes (2, 16, 17) have been sequenced and published.

Although great breakthroughs have been made in ancient DNA extraction, library preparation and bioinformatics, there still remain some challenges (18-21). Effective mapping and distinguishing of the present-day DNA contaminations from the endogenous ancient DNA is still complicated and difficult to perform, and needs 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 (22, 23). Ancient DNA is often degraded into very small fragments due to physical, chemical or biological factors during the long-term preservation in unfavorable conditions. These effects always leave valuable marks on ancient DNA to help us distinguish it from modern DNA, including the 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 the ancient DNA fragments, and the severely fragmented nature (4, 21). These unique characteristics of ancient DNA can be used to identify the true ancient DNA.

Bioinformatics methods have been developed for mapping and extracting endogenous ancient DNA from total ancient DNA (18, 21). In the mapping procedure for ancient DNA, the software BWA with the parameters set "*aln -l 1024 -n 0.03* " is usually applied to map ancient sequencing data against the reference genome(18). However, this process is time-consuming. A newly developed method like BWA *mem* with the*seed-reseed-extend* algorithm, provides new insights for mapping of ancient DNA(24). Meanwhile, Skoglund *et al* (21) developed PMDtools to separate genuine endogenous DNA from homologous contaminations. This method is effective in filtering modern human contaminated DNA from the ancient human DNA. However, it is difficult for the PMDtools to set an appropriate threshold value of PMDS, when the contamination rate 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 Illumina Hiseq platform from 6 samples (representing three species) to optimize the ancient DNA mapping, which 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 further explored a more universal and effective filtration pipeline to filter present-day contaminations based on the ancient DNA cytosine deamination, depurination and fragmentation using the simulated data. The final recommendations presented in this study enabled reduction of modern human DNA contamination to an extremely low level while maintaining a high rate of endogenous DNA. The mapping guidelines coupled with screening recommendations to control for modern DNA contamination could support future studies on ancient DNA.

2. Materials and Methods

2.1 Samples and Data Resource

We investigated previously sequenced whole-genome sequences from ancient animals in order to exclude the possibility of missing out any important data. 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*)(25-27), one ancient goat (28) and one ancient aurochs (29). 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 the 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 may lead to inaccurate evaluation. Therefore, we avoided using the real ancient sequencing data for analysis, but instead used the simulated ancient sequences with the same damage parameters as those of the real data. With simulated data, it is possible to clearly determine the true state of the genuine ancient DNA and effectively evaluate the mapping and separating method of the endogenous DNA. To obtain damage and fragmentation parameters from the real data, we used mapDamage2.0 (30) to calculate the frequency of C-to-T and G-to-A changes at the ends of DNA fragments, and the length distribution of the real ancient DNA data we collected. To simulate the real contaminations, we sequenced an ancient panda (˜ 100 years old, CNP0000732) and mapped the raw reads by blast using the nucleotide database (31) to obtain the real proportion of contaminated reads. This contamination consisted of more than twenty thousand modern species including all other possible contaminates (Figure S1). The real contamination data were added into simulated endogenous ancient DNA to test for the ancient genome mapping methods, and modern human DNA fragments were mixed with simulated ancient human DNA to explore the method for filtering the homologues contaminations. Finally, we used the software gargammel (perl gargammel.pl -n 1000000 –comp 0,cont_rate,endo_rate -f sizefreq.size -mapdamagee misincorporation.txt single/double -o data/simulation data/)(32) to simulate FASTQ files including one million reads of ancient DNA sequences for the above mentioned 6 ancient samples. Nine different contamination rates were simulated (20%, 40%, 60%, 80%, 90%, 95%, 99%, 99.5% and 99.9%).

2.3. Genome Mapping of Simulated Ancient DNA

We compared the BWA *aln* (Version: 0.7.17) and BWA *mem*(Version: 0.7.17) to explore a more effective mapping method for ancient DNA. Here,"bwa *aln -l 1024 -n* 0.03"(MS parameters) (18) was compared with BWA *mem* to seek a better mapping strategy for ancient DNA.

We defined the valid mapping hits as reads with endogenous ancient DNA tags (all simulated endogenous ancient DNA were tagged before mapping) and with a mapping quality higher than 30. For the BWA *mem*algorithm, the most important part is the seed-reseed-extend strategy which might be suitable for ancient DNA. As such, we also tested BWA*mem* with the minimum seed length (parameter -k: 9/14/19/24/29) and the maximum seed length without reseeding (parameter -r: 0.5/1/1.5/2/2.5) parameters. In order to evaluate the 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

To explore a more universal and effective pipeline to separate 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 the "DeamNum" (which represents the number of C-to-T or G-to-A mutations), one, two and three were tested. For the "DetectRange" (which represents the base number), five, ten and fifteen were tested. For the "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). We wrote a program using Python to simplify this pipeline (home page:*https://github.com/tianminglan/AncFil*). One can test more possible conditions by adjusting parameters "-DeamNum", "-DetectRange" and "-DoubleOrSingle" Secondly, given that there is a natural tendency of depurination at the 5' ends of ancient DNA fragments (33), we screened reads with an A or G at the position preceding the first base of the 5'end. 3Finally, the effects of the length of ancient DNA fragments on separating endogenous DNA was evaluated. Here, two criteria's 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 (21) were used to filter the homologous contaminations using the same data and evaluating criteria used to evaluate our recommended method above. Meanwhile, the "-threshold" is one of the most important parameters in PMDtools for adjusting the strictness of the filtration. To make a comprehensive 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

The average length of ancient DNA data that we collected ranged from 45bp to 58bp (Table 1). The age of samples 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 ends of DNA fragments ranging from 2% to 80 % (Figure. S2). We simulated a total of 54 ancient DNA data set containing the same length distribution and damage pattern as the real data set (Table 1). One million reads were finally simulated under each condition.

3.2 Comparing different Mapping Algorithms on Ancient DNA

Ancient DNA damage, especially C-to-T changes, can result in mis-mapping when ancient DNA fragments are mapped to the reference genome. The mapping methods and parameters used for modern DNA are not always suitable for ancient DNA. We compared BWA *aln* and BWA *mem*to explore a more effective mapping strategy based on the characteristics of ancient DNA damage.

The comparison was achieved by calculating the CRT, LRE and MT for each dataset, and performing Repeated Measurement Analysis of Variance. The average and median value of CRT, LRE and MT of the two algorithms with different contamination rates are shown in Table S3. The analysis showed no significant differences in CRT ($F = 1.42$, $P = 0.2870$) and LRE ($F = 0.44$, $P = 0.5344$). However, Significant differences were found in MT ($F = 41.57$, $P = 0.0013$) and BWA *aln* with the MS parameter taking 8.13-fold time than BWA*mem* by default (Table S4). We further evaluated the influence of different samples and different contaminations rates on ancient DNA mapping. As shown in Table S5 and Figure 1, the CRT maintained the same level in different samples. The mean values of LRE were stable between different contamination rates but not when contamination rate was close to 100%.

The seed-reseed-extend strategy is one of the most important aspects of the BWA *mem* algorithm. This strategy is mainly supported by two parameters including the minimum seed length (parameter -k) and the maximum seed length without reseeding (parameter -r). Additionally, the algorithm searches for internal seeds inside a seed longer than x bp (x=[-k] * [-r]). In this study, we tried to optimize these two parameters to further explore more effective mapping parameters for ancient DNA mapping.

We calculated the CRT, LRE and MT using different parameters –k (9/14/19/24/29) (Figure. 2) and performed Repeated Measurement Analysis of Variance Analysis to compare the results generated under different parameters. There were significant differences in CRT ($F = 644.61$, $P < 0.0001$), LRE ($F = 17.99$, $P = 0.0057$) and MT ($F = 146.75$, $P < 0.0001$) (Table S6). The significant differences of LRE were found between "-k 29" and the other values. The MT remarkably increased at "–k 9" and was significantly longer than all other tests. Moreover, "–k 14" consumed more running time than that of "-k 14" and "-k 29" (Table S7).

We evaluated the parameter –r (0.5/1/1.5/2/2.5) with the same method used to evaluate the parameter –k (Figure. 3). Significant differences were found in CRT ($F = 392.45$, $P < 0.0001$), LRE ($F = 45.11$, $P = 0.0010$) and MT ($F = 9.19$, $P = 0.0002$) (Table S8). A significant decrease in LRE was observed when the set values of "-r" were greater than 1.5. MT was significantly higher with "-r 0.5" than that of all other "-r" options. Similarly, there was a significant decrease in MT between "-r 1.0" and the other "-r" options (except "-r 0.5"), but no significant differences were found between other pairwise comparisons (Table S9).

3.3 Separation of endogenous DNA

The unique ancient DNA characteristics, especially the C-to-T and/or G-to-A changes at ends of DNA fragments help to improve the filtering of contaminated present-day DNA. Using these characteristics, we tried to decrease the homologous contamination rate to a very low level through screening of DNA fragments with C-to-T and G-to-A changes within the first or last "DetectRange" base pair at 3' and/or 5' ends (Figure.

4). The mean values of CRT and LRE are shown in the Table S10. No significant differences were found in CRT (F = 3.27, P = 0.1097) and LRE (F = 1.11, P= 0.3893) (Table S11).

We further tested the influence of parameter "DeamNum" on separating endogenous DNA. Significant differences were found in CRT (F = 26.01, P = 0.0011) and LRE (F = 24.03, P = 0.0152) (Table S12). We found that an increase of "DeamNum" could lower the CRT, but result in a higher LRE (Figure S3). We also calculated CRT and LRE to evaluate the influence of the parameter "DoubleOrSingle" on ancient DNA mapping (Figure. S4). Significant difference was found in the CRT (F = 44.97, P = 0.0068) but not in the LRE (F = 7.20, P = 0.0748) (Table S13). The result showed a nearly 10-fold decrease of 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 S10). The homologous contamination rate can be kept to an average of 0.92% by using the filtering strategy with "-DetectRange=15 -DeamNum=1 -DoubleOrSingle=or".

We further compared our method with the software PMDtools. These two methods were run in parallel using the same dataset, and the results generated by the PMDtools with different parameters are shown in the Table S14. To make the comparison fairer, the LRE of the tested dataset were kept similar in both our pipeline and the PMDtools. We found no significant differences in the CRT (Z = -1.171, P =0.241) between the two methods. However, the running time of our method was 6.48 times shorter than that of PMDtools, and the difference was significant (Difference=2.3mins, Z=-6.50, P=8.28E^{-11}). Although this comparison could not show the superiority of our method to PDMtools, we at least demonstrate a fast and reliable complement to the PDMtools.

Furthermore, 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 the depurination characteristic.

4. Discussion

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

BWA *aln* with the MS parameters has proved to be effective in ancient DNA mapping by disabling the seed function and decreasing the difference tolerance in mapping (18). However, we did not find significant differences in CRT (F = 1.42, P = 0.2870) between the Schubert's method and the BWA *mem* algorithm. It indicated that BWA *mem* with default parameters (BWA *mem* -k 19 -r 1.5) was able to perform ancient genome mapping as well as BWA*aln* with the MS parameter. Additionally, the seed-reseed-extend strategy in BWA *mem* can help to accelerate the mapping process, and it resulted in an 8.13-fold decrease of MT than the BWA *aln*algorithm. Therefore, BWA *mem* can reduce the contamination rate and improve the accuracy of ancient genome analysis while consuming a short time to run the mapping process. Additionally, the CRT was maintained at the same level among different samples, demonstrating the universal property on ancient genome mapping.

Soft clipping is one of most important issue to consider when using the BWA *mem* . To determine whether soft clipping could help decrease the contamination rate, we counted all soft clipped reads. The results showed that 7.9% mapped reads were soft clipped during the mapping and 6% 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. As to hard clipping, it's a special kind of soft clipping to mark the multiple mapping of a read. But only 0.0036% mapped reads showed damaged pattern. Therefore, the soft clipping can hardly make a big impact on the further filtering of endogenous DNA by using the deamination characteristic. In summary, the BWA *mem*performed equally well as the BWA *aln* with MS parameters when considering the contamination rate used in this study, but BWA*mem* costs less running time than that of the BWA *aln*method. If we took all the conditions into consideration, BWA *mem*performed better.

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

The parameters -k and -r were extremely important for the “seeding and reseeding” mapping stages in BWA *mem* (24). The different parameter values of -k and -r could significantly affect the CRT, LRE and MT, indicating that we can obtain ancient DNA mapping results with a lower contamination rate by optimizing the parameter values (Table S6, S8). The LER was highest at -k = 29, and it increased as the value of k increased. The value of LER 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 23.3 times larger than that between -k = 19 and -k = 9(Table S7). This means that the decrease of –k resulted to exponential reduction of loss of true endogenous ancient DNA. Therefore, it is of central importance to choose the appropriate value of -k to map against genome for ancient DNA. For the parameter –r, the LER reached the lowest level at -r=2.5, which means that it could avoid losing endogenous reads significantly (Table S9).

In addition, there were significant differences in running time when changing the –k and -r parameters (Table S6, S8). The running time significantly decreased from -k=9 and –r=0.5 to –k=19 and –r=1.5, but it was relatively stable and even slightly longer when –k and –r was larger than 19 and 1.5, respectively (Table S7, Table S9, Figure. 2, Figure. 3). The BWA *mem* algorithm only found the exact matches in a read while seeding, and this algorithm could trigger re-seeding with the maximal exact matches (MEMs) to reduce the loss of mis-mapping if MEMs are larger than [-k*-r] (24). The larger [-k*-r] meant fewer seeds, which can accelerate the mapping process. However, too long seed would also make seed mapping against genomes more difficult and eventually more time-consuming. We also found that the running time was more sensitive to the change of –k parameter rather than the –r (Table S7, Table S9, Figure. 2, Figure. 3), indicating that the running time was mainly influenced by the minimum seed length. The –r cannot affect the seeding for MEMs, but the –k can influence both seeding and reseeding procedures, which might be the possible reason for their influence on the running time. Based on these comparisons, we recommend use of –k=19 and –r=2.5 for BWA *mem* mapping of ancient DNA.

4.3 Improving the separation of endogenous DNA

Among all kinds of homologous contaminations, the present-day human DNA is the most frequent contamination in ancient human DNA, because it can be easily induced from the time sample are collected to the time DNA library preparation is performed. These homologous contaminations are extremely difficult to remove. In our testing, the proportion of homologous contamination that could be removed from the simulated raw data decreased with the increase in simulated contamination rates, and there was a significant negative correlation between them (R^2 =0.391, P =0.019). However, it was still possible to remove more than 99% homologous contaminations even when the simulated contamination rate reached 99% (Figure. 5). It is worth noting that the proportion of homologous contamination increased to 99.9% when the simulated contamination rate decreased to 95%. On average, 99.07% contamination could be removed using our recommended screening method (Table S15), which was lower than many other ancient DNA studies (26, 27). Besides, no significant differences were found in the endogenous DNA rate considering the different samples, different damage patterns and different contamination rates, demonstrating 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 explore more possibly 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 homologous contamination ratio to 2.25% which meant that this method may enable the recovery of more endogenous DNA (Table S16). Similar to the deamination screening, no differences were found in relation to sample ages, which was largely due to the weak correlation between depurination and samples ages. No significant correlation between samples ages and the extent of DNA fragmentation was recorded. DNA fragments are usually heavily degraded due to depurination shortly after death (34, 35). However, only 10%-40% of ancient DNA fragmentation is triggered by depurination, and other factors can also result in DNA fragmentations. As such, it is difficult to identify more endogenous ancient reads by screening the DNA length. However, we also offer this option in our python script to support the filtration by depurination and fragmentation.

5. Conclusion

In this study, we found that BWA *mem* with the parameter –k=19 and –r=2.5 is comparable to the BWA *aln* with MS parameters (18) when considering the recovery of ancient DNA, but has a significantly shorter running time than that of BWA *aln* with MS parameters. For the recovery of endogenous DNA from the ancient sequencing data with homologous contaminations, we recommend screening of reads with the parameter –DeamNum=1, –DetectRange=15 and –DoubleOrSingle=or, which will allow removal of 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 will benefit ancient DNA studies, especially for samples preserved under poor conditions.

References

1. Delsuc F, Kuch M, Gibb GC, Karpinski E, Hackenberger D, Szpak P, et al. Ancient Mitogenomes Reveal the Evolutionary History and Biogeography of Sloths. Curr Biol. 2019.

2. Palkopoulou E, Lipson M, Mallick S, Nielsen S, Rohland N, Baleka S, et al. A comprehensive genomic history of extinct and living elephants. Proc Natl Acad Sci U S A. 2018;115(11):E2566-E74.

3. Chang D, Knapp M, Enk J, Lippold S, Kircher M, Lister A, et al. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. Sci Rep. 2017;7:44585.

4. Stoneking M, Krause J. Learning about human population history from ancient and modern genomes. Nat Rev Genet. 2011;12(9):603-14.

5. Sikora M, Pitulko VV, Sousa VC, Allentoft ME, Vinner L, Rasmussen S, et al. The population history of northeastern Siberia since the Pleistocene. Nature. 2019.

6. Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC. DNA sequences from the quagga, an extinct member of the horse family. Nature. 1984;312(5991):282-4.

7. Kefi R. Ancient DNA investigations: A review on their significance in different research fields. International Journal of Modern Anthropology. 2011;1(4).

8. Rizzi E, Lari M, Gigli E, De Bellis G, Caramelli D. Ancient DNA studies: new perspectives on old samples. Genet Sel Evol. 2012;44:21.

9. Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci U S A. 2013;110(39):15758-63.

10. Hofreiter M, Paijmans JL, Goodchild H, Speller CF, Barlow A, Fortes GG, et al. The future of ancient DNA: Technical advances and conceptual shifts. Bioessays. 2015;37(3):284-93.

11. Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, et al. Sequencing the nuclear genome of the extinct woolly mammoth. Nature. 2008;456(7220):387-90.

12. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710-22.

13. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338(6104):222-6.

14. Allentoft ME, Sikora M, Sjogren KG, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522(7555):167-72.

15. Marciniak S, Perry GH. Harnessing ancient genomes to study the history of human adaptation. Nat Rev Genet. 2017;18(11):659-74.

16. Fages A, Hanghøj K, Khan N, Gaunitz C, Seguin-Orlando A, Leonardi M, et al. Tracking five millennia of horse management with extensive ancient genome time series. 2019;177(6):1419-35. e31.

17. MacHugh DE, Larson G, Orlando L. Taming the Past: Ancient DNA and the Study of Animal Domestication. Annu Rev Anim Biosci. 2017;5:329-51.

18. Schubert M, Ginolhac A, Lindgreen S, Thompson JF, Al-Rasheid KA, Willerslev E, et al. Improving ancient DNA read mapping against modern reference genomes. BMC Genomics. 2012;13:178.

19. Rohland N, Glocke I, Aximu-Petri A, Meyer M. Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. Nat Protoc. 2018;13(11):2447-61.

20. Gansauge MT, Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc. 2013;8(4):737-48.

21. Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Paabo S, Krause J, et al. 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. 2014;111(6):2229-34.

22. Green RE, Briggs AW, Krause J, Prufer K, Burbano HA, Siebauer M, et al. The Neandertal genome and ancient DNA authenticity. EMBO J. 2009;28(17):2494-502.

23. Richards MB, Sykes BC, Hedges REM. Authenticating DNA Extracted From Ancient Skeletal Remains. 1995;22(2):0-299.

24. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;1303.

25. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. The genetic history of Ice Age Europe. Nature. 2016;534(7606):200-5.

26. Schuenemann VJ, Peltzer A, Welte B, van Pelt WP, Molak M, Wang CC, et al. Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. Nat Commun. 2017;8:15694.

27. Sawyer S, Renaud G, Viola B, Hublin JJ, Gansauge MT, Shunkov MV, et al. Nuclear and mitochondrial DNA sequences from two Denisovan individuals. Proc Natl Acad Sci U S A. 2015;112(51):15696-700.

28. Daly KG, Maisano Delser P, Mullin VE, Scheu A, Mattiangeli V, Teasdale MD, et al. Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. Science. 2018;361(6397):85-8.

29. Park SD, Magee DA, McGettigan PA, Teasdale MD, Edwards CJ, Lohan AJ, et al. Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. Genome Biol. 2015;16:234.

30. Jonsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29(13):1682-4.

31. Lan TM, Lin Y, Njaramba-Ngatia J, Guo XS, Li RG, Li HM, et al. Improving Species Identification of Ancient Mammals Based on Next-Generation Sequencing Data. Genes (Basel). 2019;10(7).

32. Renaud G, Hanghoj K, Willerslev E, Orlando L. gargammel: a sequence simulator for ancient DNA. Bioinformatics. 2017;33(4):577-9.

33. Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prufer K, et al. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci U S A. 2007;104(37):14616-21.

34. Dabney J, Meyer M, Paabo S. Ancient DNA damage. Cold Spring Harbor perspectives in biology. 2013;5(7).

35. Sawyer S, Krause J, Guschanski K, Savolainen V, Paabo S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PLoS One. 2012;7(3):e34131.

Tables

*. BP: before present

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**	**Ave**
Homo sapiens	JK2911	2.7	(26)	2.35 E+06	1.37E+08	58.2
Homo sapiens	Villabruna	14	(25)	1.22E+07	6.70E+08	54.9
Homo sapiens	AfontovaCava 3	17	(25)	8.88E+05	5.15E+07	58.03
Homo sapiens	Denisova_8	>50	(27)	8.26E+05	3.69E+07	44.6
Bos primigenius	British aurochs	6.7	(29)	7.51E+07	3.48E+09	46.29
Capra aegagrus hircus	Direkli5	11.5	(28)	3.04E+07	1.40E+09	45.94

Figure Legend

Figure 1. Comparison of two mapping strategies. A: Comparison of CRT. B: Comparison of LRE. C: Comparison of MT. "Bwa mem" were performed using default parameter values. And, "Bwa aln" was used with MS parameters.

Figure 2. Comparison of BWA mem results with "-k" parameters. A: Comparison of CRT B: Comparison of LRE. C: Comparison of MT.

Figure 3. Comparison of "BWA mem" results with "-r" parameters. A: Comparison of CRT B: Comparison of LRE. C: Comparison of MT.

Figure 4. Comparison of deamination filtering with "-DetectRange" parameters. A: Comparison of the CRT after filtering. B: Comparison of the LRE after filtering.

Figure 5. The homologous contamination rate after filtering by use of the AncFil with the 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.

Figure S1. Species information of the top 10 simulated exogenous contaminations

Figure S2. the damage pattern distribution of 6 samples

Figure S3. Comparison of deamination filtering with "-DeamNum" parameters. A: Comparison of the CRT after filtering. B: Comparison of the LRE after filtering.

Figure S4. Comparison of deamination filtering with "-DoubleOrSingle" parameters. A: Comparison of the CRT after filtering. B: Comparison of the LRE after filtering.

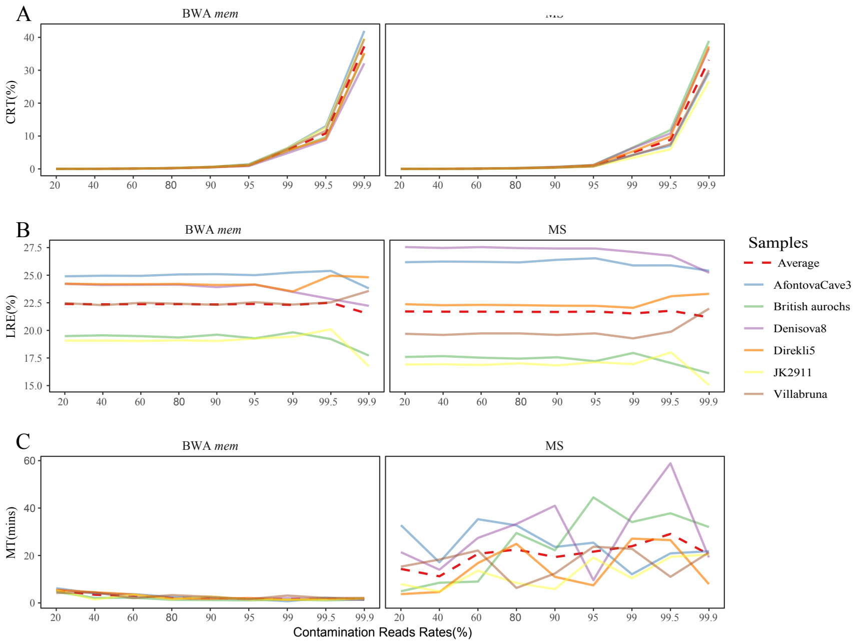
Data availability: Raw sequencing data of the ancient panda have been deposited to the CNSA (CNGB Nucleotide Sequence Archive) with accession number CNP0000732 (*https://db.cngb.org/cnsa/*).

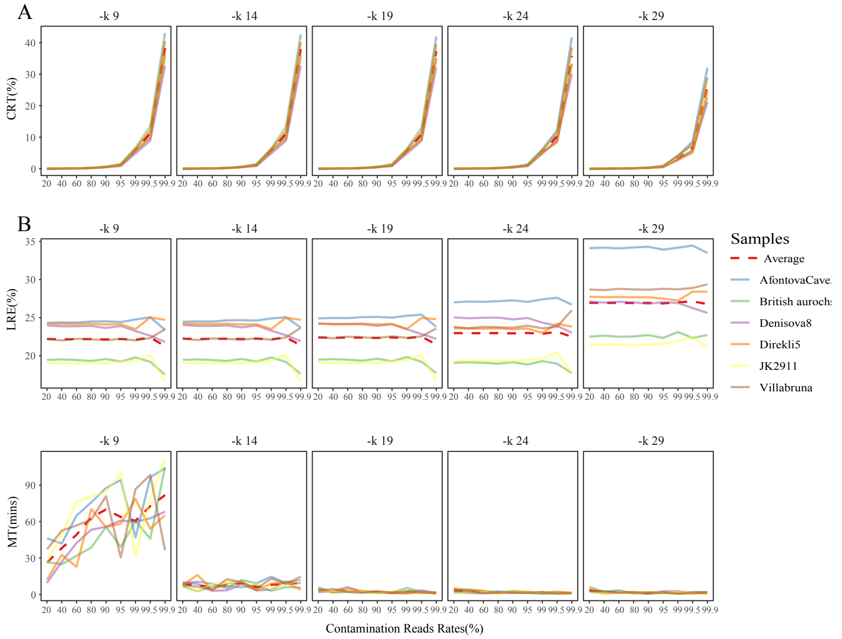
Conflicts of Interest: The authors declare no conflicts of interest.

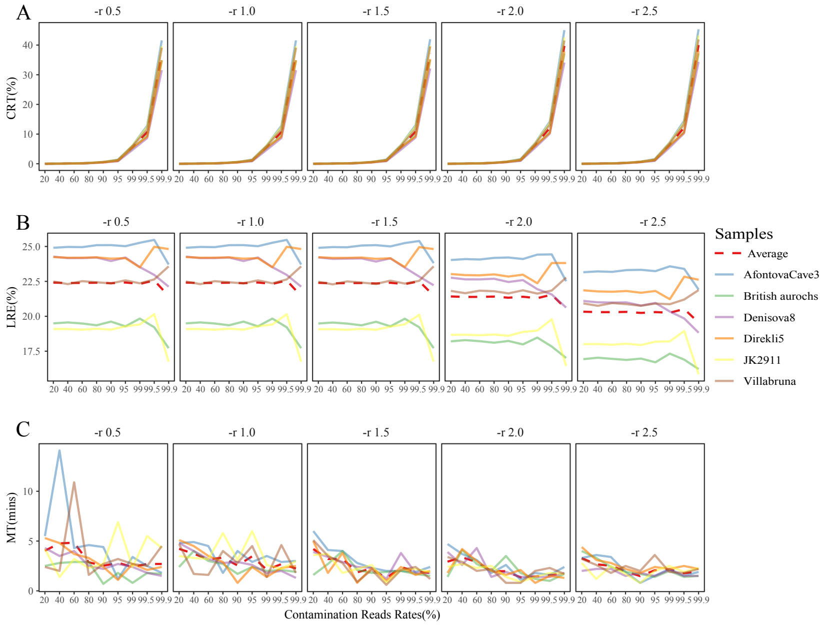
Author Contributions: conceptualization, T.M.L., X.Y.Z., Y.C.X., K.L.Z. and K.K.; methodology, T.M.L. and W.H.X.; software, W.H.X. and T.M.L.; validation, H.M.L., Y.L., Y.M., X.S.G. and J.N.N.; data curation, T.M.L. and Y.P.T.; investigation, T.M.L., W.H.X., H.B.G; writing–original draft preparation, T.M.L. and W.H.X., and J.N.N.; writing—review and editing, T.M.L. and J.N.N.; visualization, W.H.X and T.M.L.; supervision, T.M.L. and X.Y.Z..

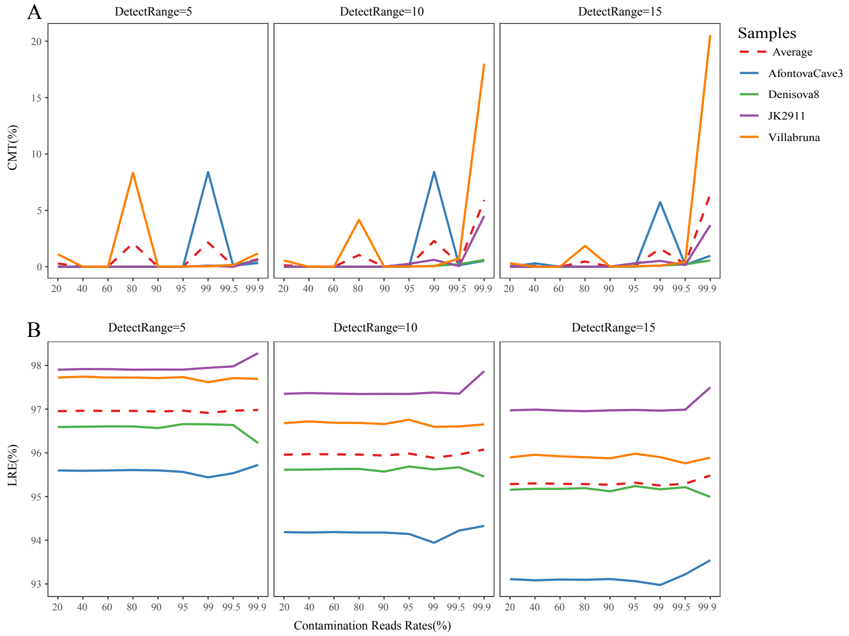
Acknowledgments:This study was supported by the Fundamental Research Funds for Central Universities (2572017PZ02), the Strategic Priority Research Program of the Chinese Academy of Sciences (grant No.

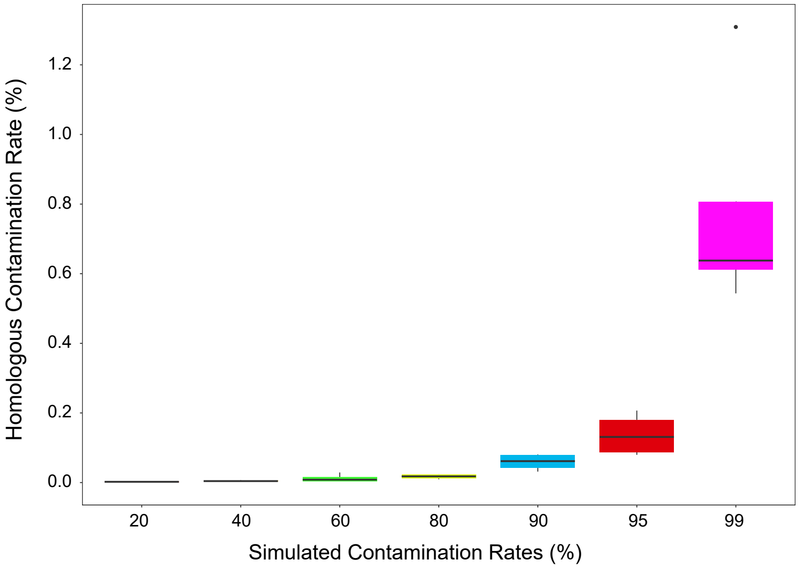
XDB 26000000, the Shenzhen Municipal Government of China (grant No. JCYJ20170817145132389) and the Guangdong Provincial Key Laboratory of Genome Read and Write (grant No. 2017B030301011). We thank the Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Guangdong, China. Finally, we are thankful to the China National GeneBank for producing the sequencing data. We thank Xiaohui Liu for giving suggestions on data analysis, Chenhui Liu and Dan Xiong for helping plot figures.

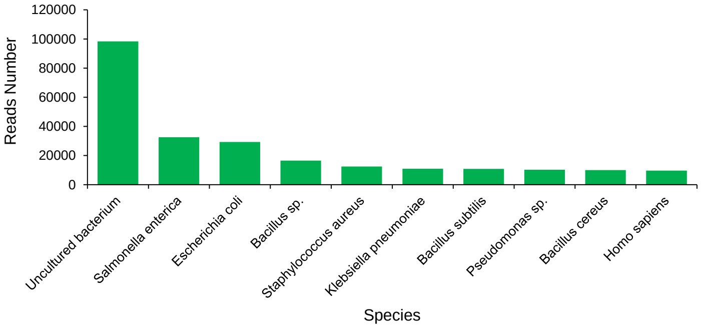


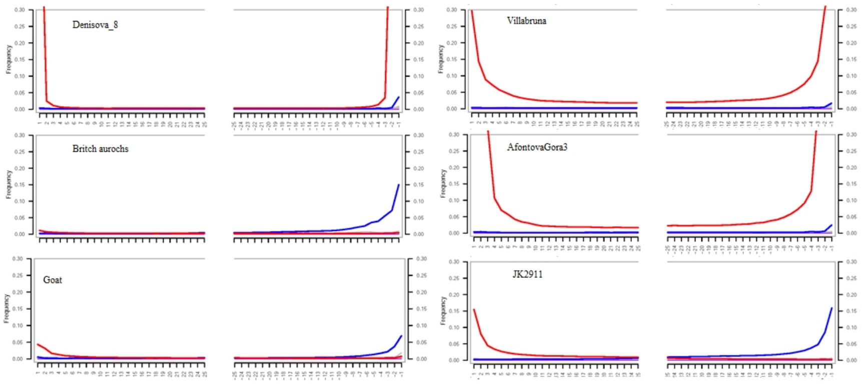


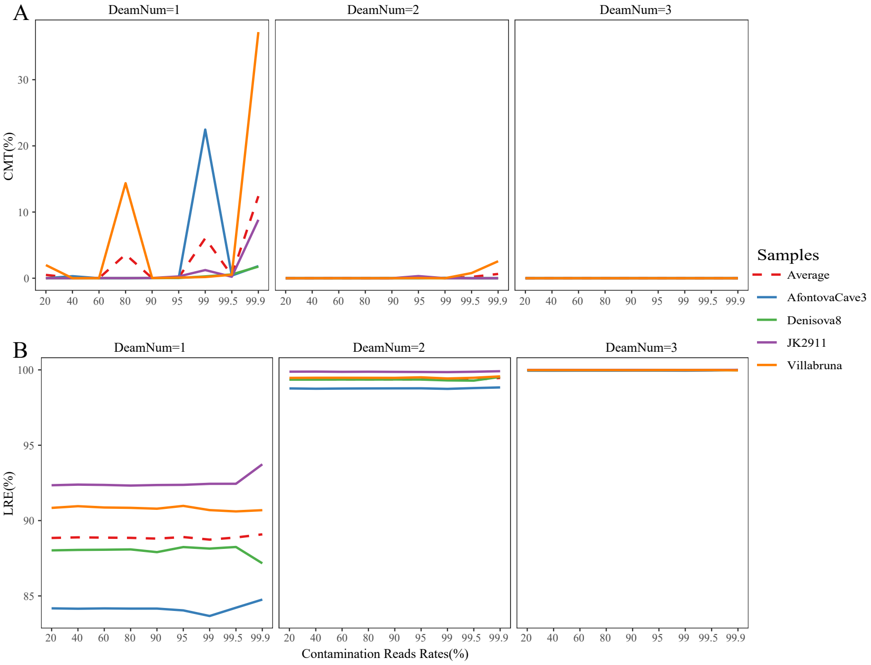

A
-r 0.5
-r 1.0
-r 1.5
-r 2.0
-r 2.5
CRT(%)
0
10
20
30
40
20 40 60 80 90 95 99 99.5 99.9
B
LRE(%)
17.5
20.0
22.5
25.0
Samples
Average
AfontovaCave3
British aurochs
Denisova8
Direkli5
JK2911
Villabruna
C
MT(mins)
0
5
10
Contamination Reads Rates(%)

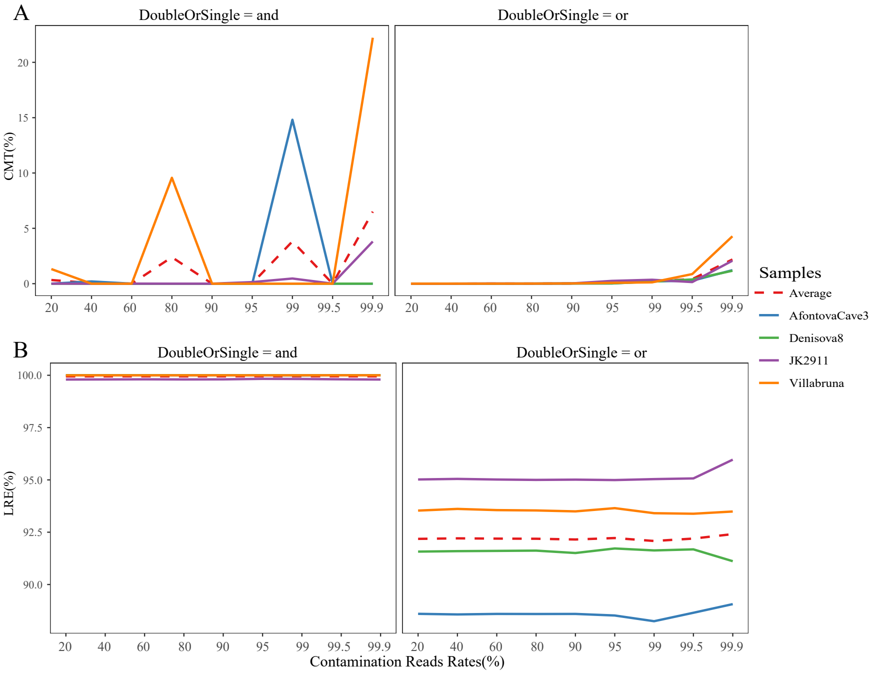

A
DetectRange=5
DetectRange=10
DetectRange=15
CMT(%)
0
5
10
15
20
Samples
Average
AfontovaCave3
Denisova8
JK2911
Villabruna
20 40 60 80 90 95 99 99.5 99.9
B
LRE(%)
93
94
95
96
97
98
Contamination Reads Rates(%)


Homologous Contamination Rate (%)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
20
40
60
80
90
95
99
Simulated Contamination Rates (%)


Reads Number
120000
100000
80000
60000
40000
20000
0
Uncultured bacterium
Salmonella enterica
Escherichia coli
Bacillus sp.
Staphylococcus aureus
Klebsiella pneumoniae
Bacillus subtilis
Pseudomonas sp.
Bacillus cereus
Homo sapiens
Species


Denisova_8
Villabruna
Britch aurochs
AfontovaGora3
Goat
JK2911
Frequency
0.30
0.25
0.20
0.15
0.10
0.05
0.00


A
DeamNum=1
DeamNum=2
DeamNum=3
CMT(%)
30
20
10
0
20
40
60
80
90
95
99
99.5
99.9
B
LRE(%)
100
95
90
85
Contamination Reads Rates(%)
Samples
Average
AfontovaCave3
Denisova8
JK2911
Villabruna


A
DoubleOrSingle = and
DoubleOrSingle = or
CMT(%)
0
5
10
15
20
20
40
60
80
90
95
99
99.5
99.9
Samples
Average
AfontovaCave3
Denisova8
JK2911
Villabruna
B
DoubleOrSingle = and
DoubleOrSingle = or
LRE(%)
90.0
92.5
95.0
97.5
100.0
Contamination Reads Rates(%)