Published as: *Curr Biol*. 2025 March 24; 35(6): R212–R213.

Concerns about ancient DNA data reported for Mengzi Ren, a Late Pleistocene individual from Southeast Asia

Daniel Tabin[1], **Nick Patterson**[1,2], **Matthew Mah**[2,3,4], **David Reich**[1,2,3,4]

[1]Department of Human Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA

[2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA

[3]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA

[4]Howard Hughes Medical Institute, Boston, Massachusetts 02115, USA

Abstract

Zhang et. al 2022 reported DNA sequences from an approximately 14-thousand-year-old skeleton excavated from Red Deer Cave: Mengzi Ren (MZR). MZR's data are the first reported from pre-Holocene Southeast Asia, with genetic affinities dissimilar to all previously published ancient DNA data. We find extremely high error rates and an abnormal error distribution in the published sequences of MZR. Even ignoring these issues, we fail to replicate key population genetic findings from Zhang et al. These results raise concerns regarding the paper's conclusions about population history and the usability of the published sequences.

MZR has an extraordinarily high rate of atypical errors

MZR's genome[1] has an extremely high rate of errors, and, particularly concerningly, these errors are not typical of those seen in typical ancient DNA. To illustrate the errors, Figure 1 and Data S1A shows the mismatch rate of both nuclear and mitochondrial DNA data to the reference sequence for MZR. There is an elevated rate of mismatch in the final nucleotides not only for cytosine deamination, which is expected in ancient DNA, but all substitution types (Data S1A, Data S1B, and Data S1C). Moreover, the error rate is often larger and the falloff as a function of distance from the end of sequence is slower on the 3' than on the 5' end of the sequence, which is not expected from known laboratory or bioinformatic processes. As an example of the impact, the 1st, 4th, 7th, and 8th highest number of sequences supporting non-consensus alleles in MZR's mitochondrial DNA are not even polymorphic in a diverse sample of 256 modern human mitochondrial genomes[2] (Data S1D). A plausible explanation is that these sites have high error rates in the MZR data. There is no biochemical

To whom correspondence should be addressed: D. Tabin (dtabin@g.harvard.edu) and D. Reich (reich@genetics.med.harvard.edu).

Declaration of Interests
The authors declare no competing interests.

reason to think that an error process should be restricted to mitochondrial DNA; it plausibly applies to the whole genome as well.

The published estimate of the contamination rate is unreliable

A powerful tool for verifying the authenticity of ancient DNA data is the mitochondrial genome, as it is not expected to be polymorphic in an uncontaminated individual, and it has a high copy number making it possible to determine a consensus sequence and thereby a rate of mismatch to the consensus. To analyze MZR's mitochondrial DNA data, we first used *samtools*[3] to remove fragments likely to be PCR duplicates, leaving an average coverage on mitochondrial DNA positions of 99. When we ran *ContamMix* on these data using standard settings, we estimate a 55–66% match rate to the consensus (95% confidence interval). This is in tension with the finding of the authors who inferred a much higher match rate to the mitochondrial consensus sequence of 94–100%. They obtained their estimate after restricting to a manually curated set of 14 positions which they chose because they recognized that the error rate in MZR's sequences is extraordinarily high and sought to address this challenge by focusing on a subset of positions minimally affected by such errors. However, the error process in MZR's data is so unusual that we are concerned it cannot be adequately addressed through filtering to a set of sites that are potentially less error prone. Moreover, trimming the sequences, either 8 bases per side, or 2 on the 5' end and 17 on the 3' end (Zhang et al.'s approach) does not address these issues as *ContamMix* estimates similar contamination in both cases.

The authors also report a direct estimate of autosomal contamination of 0.7% or less using the *AuthentiCT* software[4], which appears to be in tension with our findings. However, the published MZR sequences do not have the characteristics required for *AuthentiCT*. *AuthentiCT* compares the damage profiles on the two ends of single-stranded ancient DNA libraries constructed without damage repair. Rates are expected to rise on both the 5' and 3' ends by similar amounts, with a correlation expected for contamination. However, the published MZR sequences show their damage predominantly on the 3' end (Figure 1 and Data S1A). Whether this is due to a difference in biochemical processing compared to standard single-stranded libraries, or a bioinformatics issue, the conditions required by *AuthentiCT* are not met.

The inference of MZR's mitochondrial haplogroup is not reliable

The error rate in MZR's mitochondrial data is so high it is not even clear what the consensus mitochondrial haplogroup is, let alone what the haplogroup of any potential contaminant would be. The authors infer that the haplogroup of MZR is basal M9 which they interpret as evidence that MZR carried a deeply divergent and previously unsampled Asian lineage. However, the consensus does not have all the expected derived alleles for M9 and contains derived alleles associated with different lineages as well as private alleles. This results in a low quality-score of 0.78 according to *haplogrep*[5]. Twenty other haplogroups, including the non-M haplogroups N and L3, have comparable quality scores (0.76–0.78) for the same data. We also aligned the sequences not to the inferred human ancestral mitochondrial sequence[6], but instead to an M9 sequence constructed via a published M9a sequence[7]

from which we removed the mutations listed by Phylotree[8] to reach basal M9. The new haplogroup inferred by *haplogrep* is even more basal: L3. The instability of haplogroup inference and evidence of abnormal and extremely high errors can also be seen in manual analysis. The diagnostic position for haplogroup M9 (G4491A) is not strongly supported, while the other allele correlated to this haplogroup (T16362C) is a recurrent mutation in the hypervariable region and thus cannot be seen as reliable support for M9. When trimming the reads following the strategy of Zhang et al with 2 bases on the 5' and 17 bases on the 3' end, these issues persist; for example, the most likely haplogroup according to *haplogrep* becomes R9 (a haplogroup outside of M), and the haplogroup quality-score remains low.

Even if MZR's data were authentic, they would not support a key conclusion

Even if we were to assume that the data from MZR accurately represent the ancestry of the population from which this individual came, MZR's data do not in fact support one of its main conclusions that "there was an express northward expansion of AMHs starting in southern East Asia through the coastal line of China … eventually crossing the Bering Strait and reaching the Americas." This finding was premised on a symmetry f_4-statistic suggesting that Native Americans share alleles at an equal rate with Amur River Basin individuals from Northeast Asia from 19 kya[9] and MZR, thus implying no more affinity of Native Americans to Late Pleistocene Northeast Asians than to Late Pleistocene Southeast Asians. However, when we recompute the statistics more powerfully (using our 8.65 million SNP set), Native Americans[10] do in fact share significantly more ancestry with ~19kya Amur River Basin individuals than they do with MZR, as the symmetry statistic D(*MZR, AR19K; Ancient USA Anzick,Ancient Cameroon*) is Z = −3.6 standard errors below zero.

Discussion

The ancestry of the people of the Red Deer Caves is important. While the MZR dataset has sufficiently high error rate and contamination that analysis is challenging, high quality data from MZR or other Red Deer Cave people has the potential to provide significant insights into the deep history of eastern non-Africans.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Xiaoming Zhang and co-authors for collegial discussions, which informed the final manuscript, and Nadin Rohland for critical comments. We thank three anonymous reviewers of earlier versions of this manuscript, and particularly acknowledge reviewer #3 who carried out a manual reanalysis of the MZR mitochondrial sequences to evaluate the degree of support for the reported consensus sequence, an analysis that we reprised here after rechecking. This work was supported by NIH grant (HG012287), the John Templeton Foundation (grant 61220), and by the Howard Hughes Medical Institute.

References

1. Zhang X, Ji X, Li C, Yang T, Huang J, Zhao Y, Wu Y, Ma S, Pang Y, Huang Y, et al. (2022). A Late Pleistocene human genome from Southwest China. Curr Biol 32, 3095–3109 e3095. 10.1016/j.cub.2022.06.016. [PubMed: 35839766]
2. Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, and Paabo S (2013). DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci U S A 110, 2223–2227. 10.1073/pnas.1221359110. [PubMed: 23341637]
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. 10.1093/bioinformatics/btp352. [PubMed: 19505943]
4. Peyregne S, and Peter BM (2020). AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination. Genome Biol 21, 246. 10.1186/s13059-020-02123-y. [PubMed: 32933569]
5. Schonherr S, Weissensteiner H, Kronenberg F, and Forer L (2023). Haplogrep 3 - an interactive haplogroup classification and analysis platform. Nucleic Acids Res 51, W263–W268. 10.1093/nar/gkad284. [PubMed: 37070190]
6. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, et al. (1981). Sequence and organization of the human mitochondrial genome. Nature 290, 457–465. 10.1038/290457a0. [PubMed: 7219534]
7. Peng MS, Palanichamy MG, Yao YG, Mitra B, Cheng YT, Zhao M, Liu J, Wang HW, Pan H, Wang WZ, et al. (2011). Inland post-glacial dispersal in East Asia revealed by mitochondrial haplogroup M9a'b. BMC Biol 9, 2. 10.1186/1741-7007-9-2. [PubMed: 21219640]
8. van Oven M, and Kayser M (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30, E386–394. 10.1002/humu.20921. [PubMed: 18853457]
9. Mao X, Zhang H, Qiao S, Liu Y, Chang F, Xie P, Zhang M, Wang T, Li M, Cao P, et al. (2021). The deep population history of northern East Asia from the Late Pleistocene to the Holocene. Cell 184, 3256–3266 e3213. 10.1016/j.cell.2021.04.040. [PubMed: 34048699]
10. Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW Jr., Rasmussen S, Moltke I, Albrechtsen A, Doyle SM, et al. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature 506, 225–229. 10.1038/nature13025. [PubMed: 24522598]

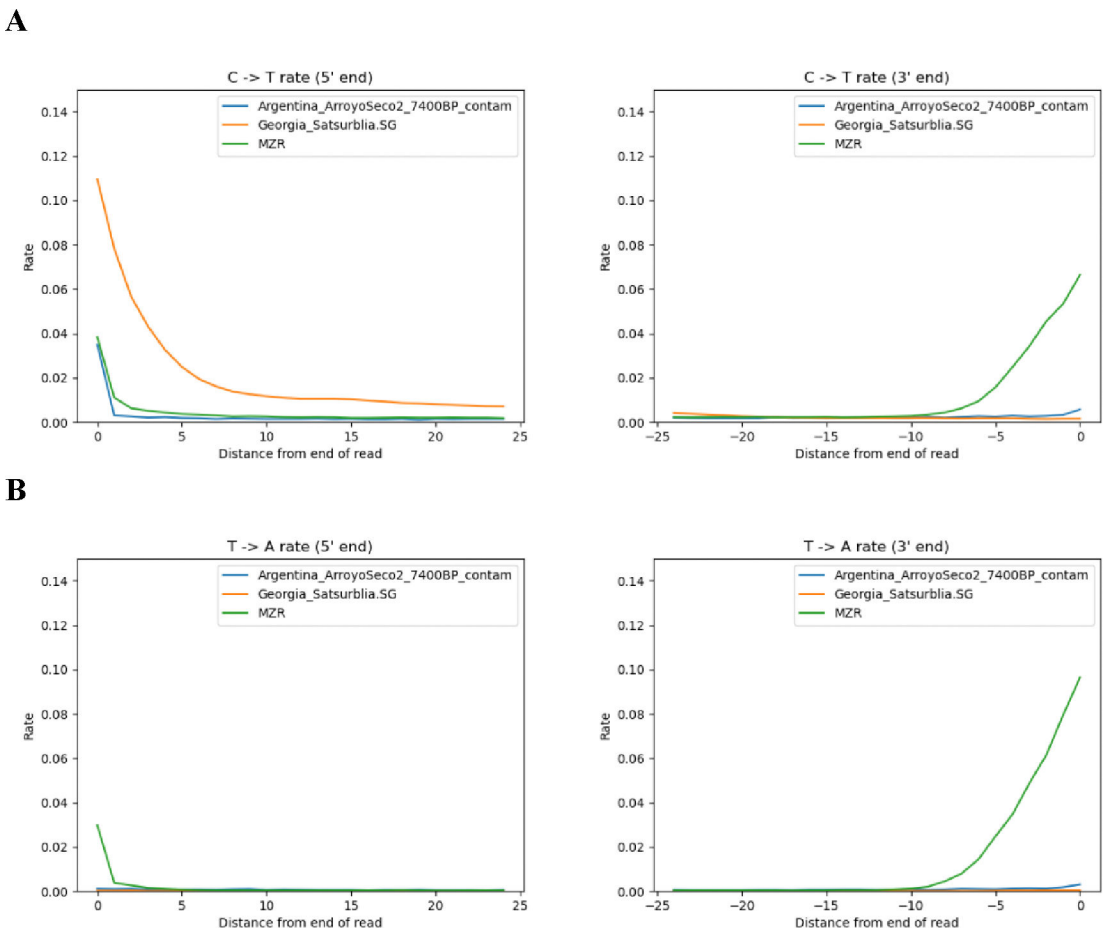

Figure 1: The MZR data show patterns of mismatch to the reference genome data unexpected for authentic ancient DNA data.

(A) We examine C>T mismatches: positions that are cytosines in the human reference genome sequence but are misread as a thymine. At the 5' end of sequences, we observe such characteristic signatures of ancient DNA at an elevated rate in MZR which superficially seems to be a signature of authenticity and indeed is also seen in two other ancient DNA datasets (a contaminated, low coverage ~7000-year-old ancient Argentinian, and a high-quality ~13000 year old ancient Georgian). However, at the 3' end we observe a stronger mismatch rate than at the 5' end, and at the 3' end we also observe a slower falloff in the rate of mismatch as function of distance from the terminus than is the case at the 5' end with substantially error rates in the final 8 base pairs; neither pattern is expected for known ancient DNA lab processes. (B) Elevated error rates, with stronger effects at the 3' end extending to the final 8 bases, are also seen at other substitution classes like T>A not expected to show elevated rate of mismatches in authentic ancient DNA data. Data S1A, S1B, and S1C shows all substitution classes revealing additional instances of patterns unexpected for authentic ancient DNA data.