

Addendum to Ancient DNA data from Mengzi Ren, a Late Pleistocene individual from Southeast Asia, cannot be reliably used in population genetic analysis

Daniel Tabin,¹ Nick Patterson,^{1,2} Matthew Mah^{2,3,4} and 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; To whom correspondence should be addressed: D. Tabin (dtabin@g.harvard.edu) and D. Reich (reich@genetics.med.harvard.edu)

1 **In addition to the issues pointed out in Tabin et al¹, the MZR data from Zhang et. al**
2 **2022² are suggestive of high levels of contamination from a source similar to modern**
3 **Han Chinese, the majority population in the country where MZR was sequenced.**
4 **In fact, MZR can be modeled entirely as Han-related ancestry and noise. These**
5 **results raise further concerns about the veracity of the MZR data and thus the**
6 **paper's historical conclusions.**

7 One of Zhang et al.'s most surprising claims is that Northeast Asians including Han
8 Chinese share more alleles with the DNA sequences obtained from the remains of the
9 MZR individual sampled from Late Pleistocene Southeast Asia than they do with later
10 ancient individuals from early Holocene Southeast Asia². We replicate this finding with
11 symmetry statistics³ of the form $f_4(\text{MZR}, \text{Early Holocene Southeast Asian}^{4-8}; \text{Han}^{9,10},$
12 $\text{Mbuti}^{9,10})$ (Supplementary Data 1), of which many are significantly positive (and none
13 are significantly negative), using a dataset of 8.65 million single nucleotide
14 polymorphisms (SNPs) which we ascertained to increase statistical power¹. However, we
15 were concerned that the signal of MZR affinity to Northeast Asians might not reflect the

true ancestral heritage of MZR, and instead might be indicative of contamination. That would be especially problematic, because the contamination would also produce a false affinity towards Native Americans. This could falsely generate some of the findings of Zhang et al².

MZR's nuclear DNA are consistent with being largely Han Chinese contamination

We were concerned that the signal of MZR affinity to Northeast Asians might not reflect the true ancestral heritage of MZR, and instead might be an artifact of contamination from modern people with Northeast Asian ancestry like Han, the majority population in China where MZR was sampled and sequenced.

The data in fact point to very high rates of Han Chinese-related contamination. When we measure shared drift of MZR and other ancient Southeast Asians with diverse modern Eurasians using the statistic $f_3(\text{Ancient Southeast Asian}; \text{Modern Eurasian}, \text{Yoruba})$, MZR shares the most drift with Han and populations closely related to Han (Supplementary Data 6). While the ~8000 year old Ancient Southeast Asian individual from Liangdao also shares an affinity to Han, it lacks the affinity to Northeast Asians such as Korean and Daur that MZR (and Han) have. The ~12000 year old Longlin, the ~8000 year old Hoabinhian, the ~7000 year old Sulawesi hunter-gatherer, and the ~40000 year old Tianyuan individual all share less drift with Han^{4,5,7}. The Longlin, Hoabinhian and Sulawesi individuals also all share more drift with Australasian populations such as Onge, Papuans, and Australians.

To understand the genetic affinities of MZR's genome, we used the software *qpAdm*¹¹, attempting to model MZR as a mixture of Han and French (representing plausible contaminants), and a dummy population consisting of heterozygous genotypes at each SNP to represent the high error rate we have shown exists in MZR¹ (we hypothesize that

40 sequencing error might randomly switch each genotype at each SNP to the alternate
41 allele, hence the idea to model error as a 50-50% frequency at each SNP). *qpAdm*
42 requires using a set of “Right” reference populations that are differentially related to the
43 different ancestry sources, making it possible to tease them apart, and for this purpose we
44 used Yoruba, Papuans, Onge, Mala, Karitiana, Japanese, Georgian, and Ami. Due to the
45 fact that MZR had most of its unusual errors on the ends of its sequences (Figure 1), we
46 carried out analyses with three different trimming schemes. First, we analyze the full
47 sequences (untrimmed). Second, we analyzed sequence trimmed 8 bases on both ends.
48 Finally, we trimmed 2 bases off the 5’ end and 17 bases off the 3’ end of sequences,
49 following the procedure of Zhang et al². All three trimming frameworks provided
50 qualitatively identical results with quantitatively similar numbers.

51 Focusing here on the results based on applying *qpAdm* to the sequences with the last 8
52 bases trimmed, we find that MZR can be fit well as having $90.1 \pm 0.9\%$ Han Chinese-
53 related DNA with the remaining $\sim 9.9\%$ coming from the dummy population capturing
54 the unusually high sequencing error in the data. This model fits even when moving the
55 French to the outgroup populations indicating no evidence of European contamination
56 (however, both the Han and noise sources are required). This type of model does not fit
57 for other ancient samples including a Taiwanese individual from Liangdao (using the
58 individual with lower coverage than MZR, thus showing that the fit of MZR is not an
59 artifact of low coverage which increases the chance of a model passing), Longlin, the
60 hunter gatherer from Sulawesi, the Hoabinhian from Laos, or Tianyuan. A non-fitting
61 model is what is expected if an ancient individual has ancestry from a lineage that is
62 phylogenetically more closely related to Papuans, Onge, Ami, or any other Right
63 population than to either Han or French (Table 1). In contrast, a fit is consistent with all
64 data being Han-related contamination plus error.

Table 1: *qpAdm* results when modeling ancient Southeast Asians as mixtures of Han Chinese, French, and a dummy population meant to simulate the effects of sequencing error. When modeling ancient Southeast Asian samples as mixtures of Han, French, and a dummy population used to represent sequencing error, with many plausible modern representatives of Southeast Asian ancestry in the right population set, all targets other than MZR fail to be properly modeled. Unlike other published ancient Southeast Asians, MZR is modeled as 90.1% derived from modern Han Chinese and 9.9% error and fails to be properly modeled when Han Chinese is removed from the sources. This is an effect specific to Han, as when French is removed, MZR continues to produce a fit. These results are consistent with MZR's data being almost entirely Han contamination, and unlike all the other ancient individuals, provides no evidence of genuine ancient Southeast Asian ancestry.

Partial Bam Source	Fake hets	French	Han	Fake hets error	French error	Han error	pval	Fake Het Z	French Z	Han Z
Longlin	0.053	0.111	0.836	0.02	0.029	0.024	0.0000	2.65	3.83	34.83
Tianyuan	0.106	0.256	0.638	0.021	0.03	0.023	0.0000	5.05	8.53	27.74
Indonesia Sulawesi HG	0.239	0.202	0.559	0.026	0.042	0.032	0.0000	9.19	4.81	17.47
Laos Hòabinhian	0.145	0.295	0.56	0.016	0.024	0.017	0.0000	9.06	12.29	32.94
Taiwan 8kya HG (tai001)	0.049	0.11	0.841	0.029	0.038	0.027	0.0073	1.69	2.89	31.15
MZR	0.079	0.037	0.883	0.014	0.017	0.012	0.1602	5.64	2.18	73.58
Longlin	0.089 X		0.911	0.017 X		0.017	0.0000	5.24 X		53.59
Tianyuan	0.227 X		0.773	0.017 X		0.017	0.0000	13.35 X		45.47
Indonesia Sulawesi HG	0.32 X		0.68	0.022 X		0.022	0.0000	14.55 X		30.91
Laos Hòabinhian	0.28 X		0.72	0.012 X		0.012	0.0000	23.33 X		60.00
Taiwan 8kya HG (tai001)	0.094 X		0.906	0.02 X		0.02	0.0007	4.70 X		45.30
MZR	0.099 X		0.901	0.009 X		0.009	0.0470	11.00 X		100.11
Longlin	0.077	0.923 X		0.03	0.03 X		0.0000	2.57	30.77 X	
Tianyuan	0.118	0.882 X		0.028	0.028 X		0.0000	4.21	31.50 X	
Indonesia Sulawesi HG	0.226	0.774 X		0.034	0.034 X		0.0000	6.65	22.76 X	
Laos Hòabinhian	0.063	0.937 X		0.023	0.023 X		0.0000	2.74	40.74 X	
Taiwan 8kya HG (tai001)	-0.228	1.228 X		0.032	0.032 X		0.0000	-7.13	38.38 X	
MZR	0.051	0.949 X		0.037	0.037 X		0.0000	1.38	25.65 X	
Longlin	1 X		X	0 X		X	0.0000 X	X		X
Tianyuan	1 X		X	0 X		X	0.0000 X	X		X
Indonesia Sulawesi HG	1 X		X	0 X		X	0.0000 X	X		X
Laos Hòabinhian	1 X		X	0 X		X	0.0000 X	X		X
Taiwan 8kya HG (tai001)	1 X		X	0 X		X	0.0000 X	X		X
MZR	1 X		X	0 X		X	0.0000 X	X		X

65 **MZR's high error rates makes separating out genuine sequences nearly impossible**

66 In order to extract potentially genuine MZR sequences, we divided the MZR data based
67 on the rate of DNA damage on the sequences. We used *PMDtools*¹² to infer if sequences
68 are not likely to be damaged and thus likely to be authentic. Given the three groups of
69 PMD values sequences tend to fall into (Figure 2), we split the sequences into three parts.
70 The sequences in the “negative” group with PMD scores < 0 have the lowest likelihood
71 of being authentic ancient DNA. The “low” group includes sequences with PMD scores
72 between 0 and 2.7 consisting of sequences that have characteristic ancient DNA damage
73 but not in the most diagnostic positions. The “high” group sequences have the highest

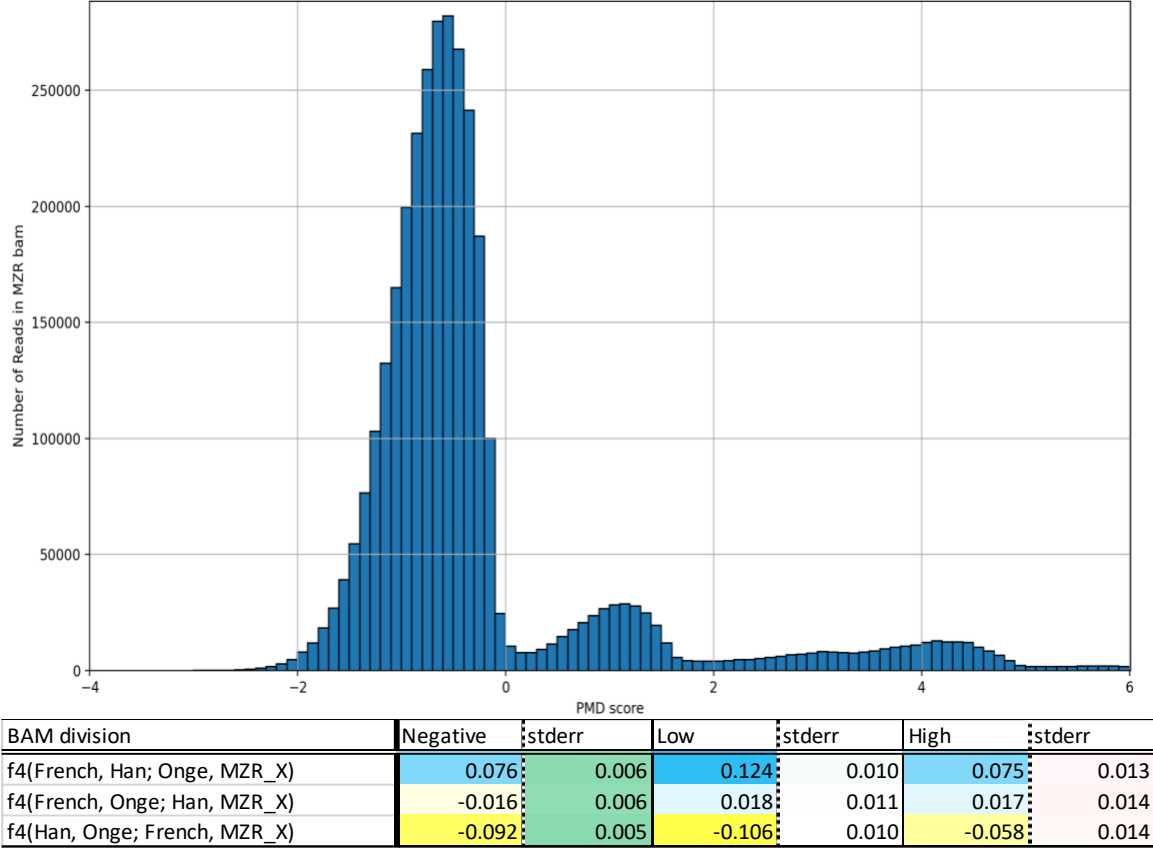
74 PMD scores >2.7 with the strongest evidence of characteristic ancient DNA damage. We
 75 computed f_4 -statistics such as $f_4(\text{MZR div 1}, \text{MZR div 2}; \text{Han}, \text{French})$ using the
 76 *qp4diff* program in ADMIXTOOLS³ as the difference between two statistics:
 77 $f_4(\text{MZR div 1}, \text{Yoruba}; \text{Han}, \text{French})$ and $f_4(\text{MZR div 2}, \text{Yoruba}; \text{Han}, \text{French})$.
 78 Because *qp4diff* does not require restriction to positions covered by sequences in two
 79 subdivisions, which is only a tiny fraction of sites for the low coverage MZR data, we can
 80 use much more of the data and obtain a more precise estimate of the statistic. We find
 81 that sequences with moderate evidence of damage have significant excess attraction to
 82 Han when compared both to the sequences with the strongest evidence of damage ($Z >$
 83 6.89), and the sequences with the least evidence of damage ($Z > 9.23$). A full comparison
 84 of the MZR divisions to Han and other Eurasians can be found in Supplementary Data 6.

85 These results are not consistent with analyzable population genetic data, both due to
 86 MZR's relative affinities to Han and French differing based on damage-score, and the
 87 fact that this differential affinity is not typical of contaminated data. While the sequences
 88 with the least evidence of damage are expected to show affinity to Han if MZR is
 89 contaminated, it is surprising that the least damaged sequences also show affinity to Han.

90 A potential explanation for this is the strange mismatch patterns on the ends of MZR
 91 sequencess¹. Whatever processes generated the errors—which appear across mutation
 92 classes and occur more on the 3' end than 5' end—also affected C->T and G->A
 93 mismatches, which are used to identify genuine ancient DNA through damage restriction.
 94 This means that it is not possible to separate the sequences into contaminated and
 95 uncontaminated bins reliably, making it difficult to extract “genuine” MZR sequences
 96 from the presumably contaminated data.

Figure 2: MZR subdivisions show differential relatedness toward French, Onge, and Han.

(a) A histogram of sequences from the MZR bam by PMD-score per sequences. MZR bams are binned into 0.1 PMD score difference bins. Three modes are observed: one with negative damage, one with low but positive damage, and one with high damage. (b) Computing the three unique f_4 statistics consisting of Han, Onge, French, and an MZR subdivision produces different population genetics results depending on the subdivision used.



Discussion

The MZR data have specific affinity to modern Han Chinese and other related populations. This is *a priori* surprising given the date and location of the MZR individual. This affinity is so strong that MZR can be well modeled as entirely Han with added noise. Neither of these traits are shared with other ancient Southeast Asians and both raise additional concerns regarding the reliability of the MZR data. Contamination seems more plausible than a population with Han-like ancestry existing in Yunnan province 14 thousand years ago.

105 **List of Supplementary materials**

106

107 **Supplementary Data 1:** f-statistics of the form $f_4(\text{MZR, Ancient Southeast Asia; Han,}$
108 Outgroup)

109

110 **Supplementary Data 2:** Table of f_3 -outgroup showing shared drift between ancient
111 Southeast Asians and modern Eurasians

112

113 **Supplementary Data 3:** Table of f_4 -differences comparing MZR subdivisions to Han
114 and other modern Eurasians

115 **Acknowledgements**

116 We thank Xiaoming Zhang and co-authors for collegial discussions, which informed the
117 final manuscript. We thank three anonymous reviewers of earlier versions of this
118 manuscript. This research was supported by NIH grant (HG012287), the John Templeton
119 Foundation (grant 61220), and by the Howard Hughes Medical Institute.

120

121 **Declaration of Interests**

122 The authors declare no competing interests.

References

1. Tabin, D., Patterson, N., Mah, M., and Reich, D. (2025). Concerns about ancient DNA sequences reported from a Late Pleistocene individual from Southeast Asia. *Current Biology* 35, R212-R213. 10.1016/j.cub.2024.10.012.
2. 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.
3. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. *Genetics* 192, 1065-1093. 10.1534/genetics.112.145037.
4. McColll, H., Racimo, F., Vinner, L., Demeter, F., Gakuhari, T., Moreno-Mayar, J.V., van Driem, G., Gram Wilken, U., Seguin-Orlando, A., de la Fuente Castro, C., et al. (2018). The prehistoric peopling of Southeast Asia. *Science* 361, 88-92. 10.1126/science.aat3628.
5. Wang, T., Wang, W., Xie, G., Li, Z., Fan, X., Yang, Q., Wu, X., Cao, P., Liu, Y., Yang, R., et al. (2021). Human population history at the crossroads of East and Southeast Asia since 11,000 years ago. *Cell* 184, 3829-3841 e3821. 10.1016/j.cell.2021.05.018.
6. Larena, M., Sanchez-Quinto, F., Sjodin, P., McKenna, J., Ebeo, C., Reyes, R., Casel, O., Huang, J.Y., Hagada, K.P., Guilay, D., et al. (2021). Multiple migrations to the Philippines during the last 50,000 years. *Proc Natl Acad Sci U S A* 118. 10.1073/pnas.2026132118.
7. Yang, M.A., Fan, X., Sun, B., Chen, C., Lang, J., Ko, Y.C., Tsang, C.H., Chiu, H., Wang, T., Bao, Q., et al. (2020). Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* 369, 282-288. 10.1126/science.aba0909.
8. Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietruszewsky, M., Pryce, T.O., Willis, A., Matsumura, H., Buckley, H., et al. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* 361, 92-95. 10.1126/science.aat3188.
9. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538, 201-206. 10.1038/nature18964.
10. Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505, 43-49. 10.1038/nature12886.
11. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207-211. 10.1038/nature14317.
12. Skoglund, P., Northoff, B.H., Shunkov, M.V., Derevianko, A.P., Paabo, S., Krause, J., and Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc Natl Acad Sci U S A* 111, 2229-2234. 10.1073/pnas.1318934111.