Article

Enamel proteins from six *Homo erectus* specimens across China

https://doi.org/10.1038/s41586-026-10478-8

Received: 21 October 2025

Accepted: 1 April 2026

Published online: 13 May 2026

Open access

Check for updates

Qiaomei Fu[1,2,3]✉, Zhongyou Wu[1,2], E. Andrew Bennett[1], Song Xing[1,3], Qiang Ji[4], Zhe Dong[5], Huiyun Rao[1], Xuejun Gu[6], Yizhao Dang[7], Jun Xing[8], Kai Zhou[1] & Xiaotian Feng[1]

Homo erectus remains have been found in Africa, Eurasia and Southeast Asia[1–3], dating back around two million years; however, owing to their age and state of preservation, obtaining informative molecular data from them has proved challenging. Here we successfully extracted and analysed ancient enamel proteins from five male and one female Middle Pleistocene *H. erectus* specimens from approximately 0.4 million years ago, from the Zhoukoudian, Hexian and Sunjiadong sites. All specimens from all three sites share two amino acid variants. Of these, A253G in AMBN is previously unknown and has not been identified in other human lineages, including *H. erectus* from Dmanisi (Georgia), *Homo antecessor* from Atapuerca (Spain), Denisovans, Neanderthals and modern humans. The other variant, AMBN(M273V), has previously been identified in Denisovans, and our evidence now indicates it may have been introduced through populations related to these Middle Pleistocene *H. erectus*. The regions in the Denisovan genome attributed to super-archaic introgression, some of which later passed to modern humans, are likely to have originated from *H. erectus*. Late Middle Pleistocene *H. erectus* may have coexisted with Denisovans in parts of East Asia, where these interactions are presumed to have occurred.

Broadly defined *H. erectus* (*H. erectus sensu lato*) was widely distributed across three continents and was geologically long-lived[1–3], lasting nearly 2 million years (Myr). *Homo erectus* is thought to have spread from East Africa into western Asia (1.8 Myr ago (Ma), Dmanisi, Georgia), marking the earliest generally accepted evidence of the genus *Homo* outside of Africa[4,5]. The earliest *H. erectus* specimen found in Europe is from the TE7 level of the Sima del Elefante site (Sierra de Atapuerca, Spain)[6], dated between 1.4 Ma and 1.1 Ma. After dispersing to China and Indonesia, *H. erectus* persisted in these regions for a considerable geological duration. In China, the record of *H. erectus* dates back to about 2.1–1.6 Ma (refs. 7–9) and disappears roughly 0.4–0.3 Ma (ref. 10). In Java, Indonesia, the fossils of *H. erectus* range from 1.5 Ma to as recent as 0.1 Ma. Overall, *H. erectus* has had an important role in human evolution. Understanding the molecular characteristics of this lineage is essential to help us to better understand hominin biology throughout the genus *Homo*. Nevertheless, the only molecular data previously recovered from *H. erectus* comprise peptides extracted from a 1.77-Myr-old tooth originating from Dmanisi in Georgia, West Asia. However, these sequences lack any single amino acid polymorphisms (SAPs) that can distinguish *H. erectus* from other human lineages[11].

The *H. erectus* specimen represented by the Zhoukoudian fossils dates from approximately 0.78 to 0.3 Ma during the Middle Pleistocene in northern China[12,13]. It is frequently used as the primary reference for the overall morphological characteristics of all *H. erectus* or Asian *H. erectus*[14]. This group has had a crucial role in research on the origin and evolution of *H. erectus*. Studies of the Zhoukoudian fossils have settled debates over whether the Indonesian Java fossils are apes or humans, confirming the systematic position of *H. erectus* within the human lineage[14,15]. In addition to the Zhoukoudian fossils, several other Middle Pleistocene *H. erectus* fossils have been discovered in China, including remains from Hexian, dated to about 0.41 Ma (ref. 16), from Nanjing, dated to at least 0.58 Ma (ref. 17), and from Sunjiadong[18], dated around 0.4 Ma (ref. 19). Nanjing and Sunjiadong fossils show similar morphology to the Zhoukoudian fossils, whereas Hexian remains are morphologically closer to Indonesian Java fossils than to Zhoukoudian fossils[14,15,18,20]. Additionally, some affinities are shared with a Middle–Late Pleistocene mandible from the Penghu Channel in Taiwan, which is purported to be a Denisovan[21]. Retrieving proteins from more than one of these fossils would help to determine whether they share common molecular characteristics that can distinguish them from others and provide insights into Middle Pleistocene *H. erectus* genetic material.

The fossils used in the protein experiments and analyses of this study are from Zhoukoudian, Hexian and Sunjiadong sites (Fig. 1 and Supplementary Note 1). Although the human fossils excavated at Zhoukoudian[22] in Beijing, northern China between 1927 and 1937 were lost, some human fossils were discovered at this site between 1949 and

[1]New Cornerstone Science Laboratory, Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China. [2]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing, China. [3]Key Research Base on Paleolithic Human Evolution and Paleogenetics (IVPP), SACH, Beijing, China. [4]Hebei International Joint Research Center for Paleoanthropology, Hebei GEO University, Shijiazhuang, China. [5]Institute of Anhui Provincial Archaeology and Cultural Relics, Hefei, China. [6]Research Center for Origin of Heluo Civilization, Archaeology Institute of Luoyang, Luoyang, China. [7]Luanchuan County Culture, Radio, Television and Tourism Bureau, Luoyang, China. [8]Hexian Culture, Tourism and Sports Bureau, Ma'anshan, China. ✉e-mail: fuqiaomei@ivpp.ac.cn

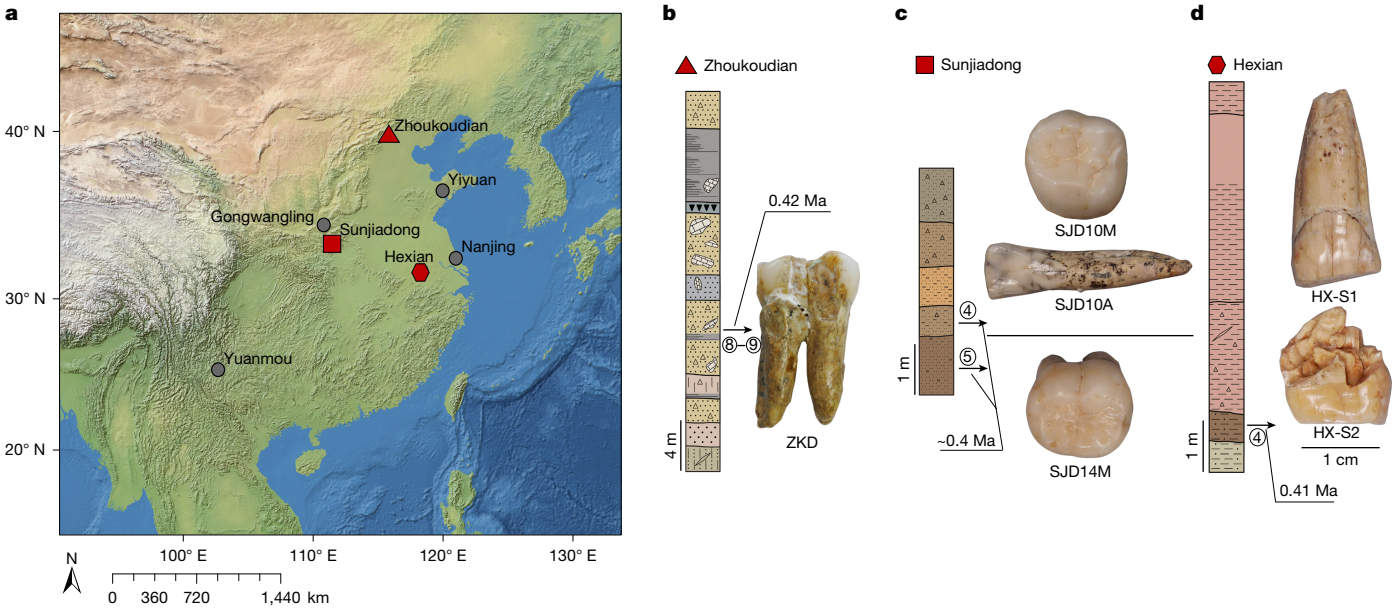


Fig. 1 | The geographic locations and samples of the Middle Pleistocene *H. erectus* sites used in this study. a, The geographic locations of Pleistocene *H. erectus* sites in China, with sites in the present study highlighted in red. Background raster: Natural Earth, Natural Earth II with Shaded Relief and Water (1:50 m), v3.2.0; public domain (https://www.naturalearthdata.com/downloads/10m-natural-earth-2/10m-natural-earth-ii-with-shaded-relief-and-water/). **b**, Stratigraphy of the Zhoukoudian site[22]. The Zhoukoudian tooth, ZKD (PA69) is from layers 8–9. **c**, Stratigraphy of the Sunjiadong site[18]. The Sunjiadong 12SJD1#10-42 (SJD10M) and Sunjiadong 12SJD1#10-45 (SJD10A) teeth are from layer 4, and Sunjiadong 112SJD1#14-22 (SJD14M) is from layer 5. **d**, Stratigraphy of the Hexian site[26]. Hexian sample 1 (HX-S1) and Hexian sample 2 (HX-S2) are from layer 4 (ref. 16).

1951, and again in 1966. The human tooth (PA69)[23] used in this study is from layers 8–9 (0.42 Ma (refs. 24,25)), found between 1949 and 1951. We also studied two human teeth from the Hexian site[26] in Anhui Province, southern China (0.41 Ma (ref. 16)) and three teeth from Sunjiadong (around 0.4 Ma (ref. 19)) in Henan Province, northern China[18]. Two of the teeth (12SJD1#10-42 and 12SJD1#10-45) at Sunjiadong are from layer 4, and the other tooth (12SJD1#14-22) is from layer 5 (ref. 18). Altogether, the dating of all 6 samples from these 3 sites yielded an approximate age of 0.4 Myr, making them roughly contemporary and representing individuals from both northern and southern China. Additionally, we included a tooth from an earlier-reported Denisovan skull (minimum age 0.15 Myr) from Harbin, China[27,28].

Preservation of animal fossil proteins

By first examining protein preservation in animal fossils from Zhoukoudian, Hexian and Sunjiadong, we gained initial insights into the feasibility of the analysis without damaging the ancient human fossils. Tests using matrix-assisted laser desorption/ionization-time-of-flight (MALDI-TOF) mass spectrometry revealed no evidence of ancient proteins in animal dentin or bones from 8 samples at Zhoukoudian, 40 samples from Hexian, and 37 samples from Sunjiadong (Supplementary Data 1). However, enamel from 3, 14 and 5 animal teeth from these respective sites did show protein signals (Supplementary Data 1). The successful retrieval of proteomic data from animal enamel suggests that extracting enamel proteins from human teeth would be feasible.

The endogenous enamel proteome

Ancient enamel proteins were successfully extracted using minimally invasive sampling techniques from six *H. erectus* teeth from Zhoukdoudian, Hexian and Sunjiadong sites, and a tooth from Harbin, China (Methods). The database used to search for endogenous peptide sequences included enamel proteins from translated genomes and published proteomes of present-day humans and other hominids[11,29–31] with spectrum data from liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysed using MaxQuant[32], PEAKS Online[33], and pFind[34] (Methods, Supplementary Note 2). These approaches generated between 4,097 and 22,107 peptide–spectrum matches (PSMs), and between 652 and 3,476 peptides (Extended Data Table 1 and Supplementary Tables 1 and 2) across 7 specimens. Peptides matching multiple genes were considered non-unique and excluded from further analysis.

We then evaluated the endogeneity of these enamel proteomes based on several criteria (Supplementary Note 3): (1) the protein and amino acid composition profiles across all seven human samples were similar and consistent with other published ancient enamel proteins[35], with minimal exogenous contamination (Fig. 2a, Extended Data Table 2 and Supplementary Tables 3 and 4); (2) peptide lengths and levels of peptide bond hydrolysis were typical of other ancient enamel proteins[11,35] (Extended Data Table 1, Supplementary Figs. 1–3 and Supplementary Table 5); and (3) elevated deamidation levels (Fig. 2b and Supplementary Tables 6 and 7) suggest more extensive protein degradation and help to authenticate the ancient proteome[11,36]. These criteria resulted in 650–3,457 peptides recovered from 6–11 enamel-related proteins (for example, AMEL, AHSG, ALB, AMBN, AMTN, COL17A1, ENAM, KLK4, MMP20, ODAM and SERPINC1) for the 7 samples (Extended Data Table 1 and Supplementary Table 1). The PSM coverage of proteins is shown in Fig. 3a and Supplementary Fig. 4. In total, the consensus sequences of endogenous proteins from each sample were constructed by retaining the overlapping amino acid coverage in PEAKS Online, pFind and MaxQuant (Methods and Supplementary Note 4), which covered 269 to 903 amino acid positions across the 6 *H. erectus* samples and the Harbin individual.

Sex determination

The male-specific marker amelogenin Y isoform (AMELY) is a gene located on the Y chromosome that is involved in tooth enamel development, and its presence is used to determine male sex[37]. However,

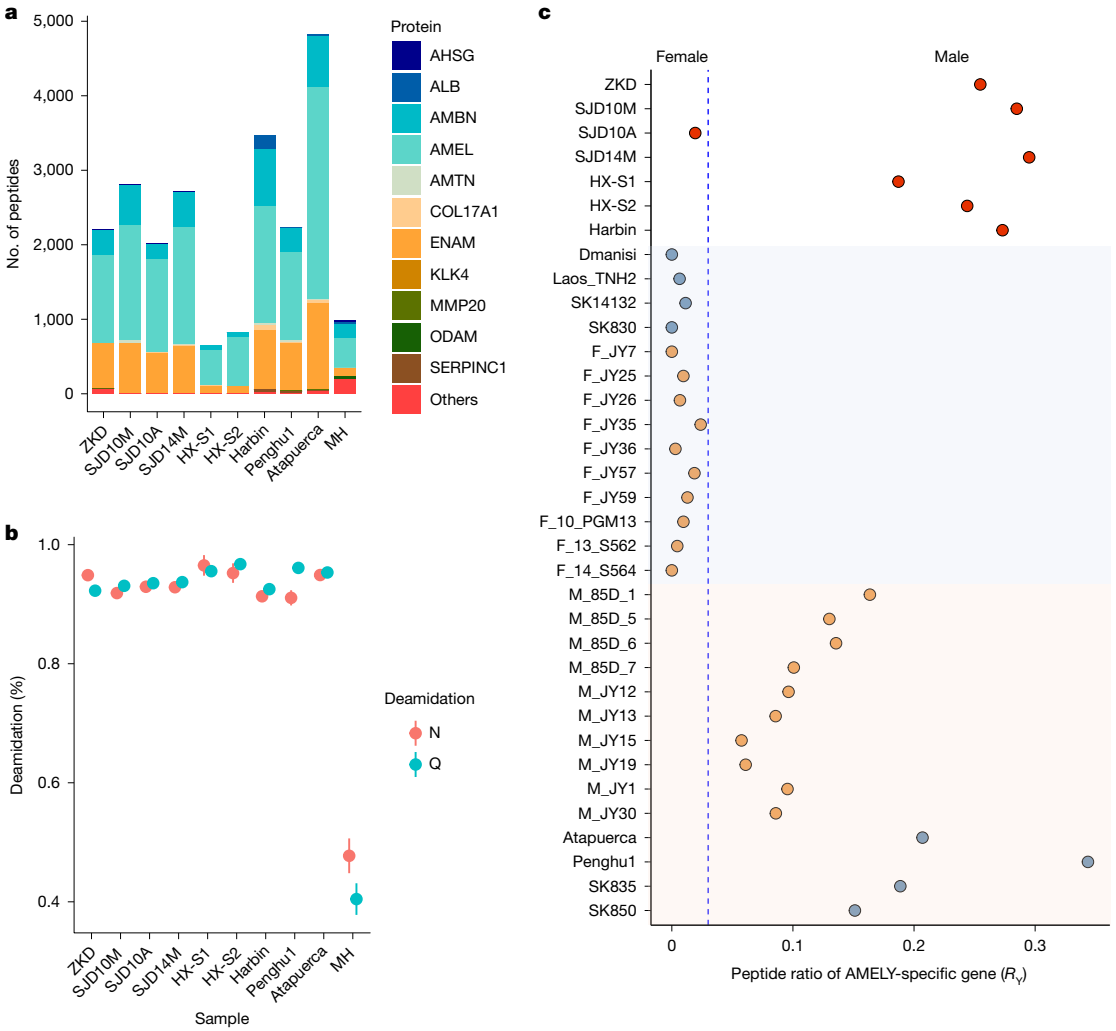


Fig. 2 | Ancient enamel protein preservation and sex identification of six Middle Pleistocene *H. erectus* specimens and the Harbin individual. **a**, Number of endogenous peptides after filtering. **b**, Overall deamidation rates of N and Q residues in the seven new samples and published samples (Penghu 1 and Atapuerca_ATD6–92). MH represents the enamel results of modern humans[43]. **c**, The observed ratio (R_Y) of peptides covering the AMELY-specific position relative to all peptides from both AMELY and AMELX in seven samples of this study (red circles), 4 *Paranthropus* specimens and 4 known archaic humans (light blue circles), and 16 ancient and 4 present-day modern humans of known sex (orange circles)[38,39] labelled as F_(female) or M_(male).

owing to the restricted database for enamel and the high sensitivity of the searching software, spurious detection of AMELY-specific peptides may result in the false-positive assignment of samples as male[38,39]. This outcome is not unexpected, as previous research has demonstrated cross-species proteomic effects resulting from search database bias[40,41]. To address this issue, we used rigorous filtering criteria to ensure the reliability of all identified SAPs in the endogenous proteome (details below), including those used to determine sex in amelogenin.

To further validate the sex determination, we established protSexInferer[42], a method based on the ratio of peptide counts covering the AMELY-specific positions relative to all peptides from both AMELY and AMELX (R_Y). Using 16 ancient and 4 present-day modern humans[38,39], along with 4 *Paranthropus* specimens (SK14132, SK830, SK835 and SK850)[2] and 4 archaic hominins (Penghu 1 (ref. 35), Dmanisi (D4163) and Atapuerca (ATD6-92)[11] and Laos_TNH2[43]) with confirmed sex, samples can be classified as male if R_Y exceeds 0.058, and female if R_Y is below 0.024 (Fig. 2c). Using this approach, the Zhoukoudian sample, the two Sunjiadong, two Hexian samples and the Harbin sample with R_Y from 0.187–0.295 were identified as males (Fig. 2c). The remaining Sunjiadong SJD10A (12SJD1#10-45, an adult), with R_Y of 0.019, was classified as female (Fig. 2c). The sex determination results were confirmed by the intensity method described in ref. 31 (Extended Data Fig. 1).

SAPs

To explore the relationships between these specimens and other humans and primates, we identified peptide counts with amino acid substitutions that are specific to modern humans, Neanderthals, Denisovans, chimpanzees, gorillas and orangutans across all nine endogenous proteins. These required at least two peptides per amino acid allele, derived from different mass spectrometry instruments or laboratories. For all SAPs recovered from the endogenous proteomes, the tandem mass (MS/MS) spectra linked to the relevant PSMs were manually inspected to confirm the presence of typical fragmentation patterns of high-energy collision dissociation, the fragmentation method used in our mass spectrometry. Furthermore, at least three out of four y/b ions bracketing the SAP site were required to ensure high-quality confirmation. SAPs at the C or N terminus of the peptide were accepted if one related PSM contained a single y or b ion or more, as these could only produce one y ion and one b ion

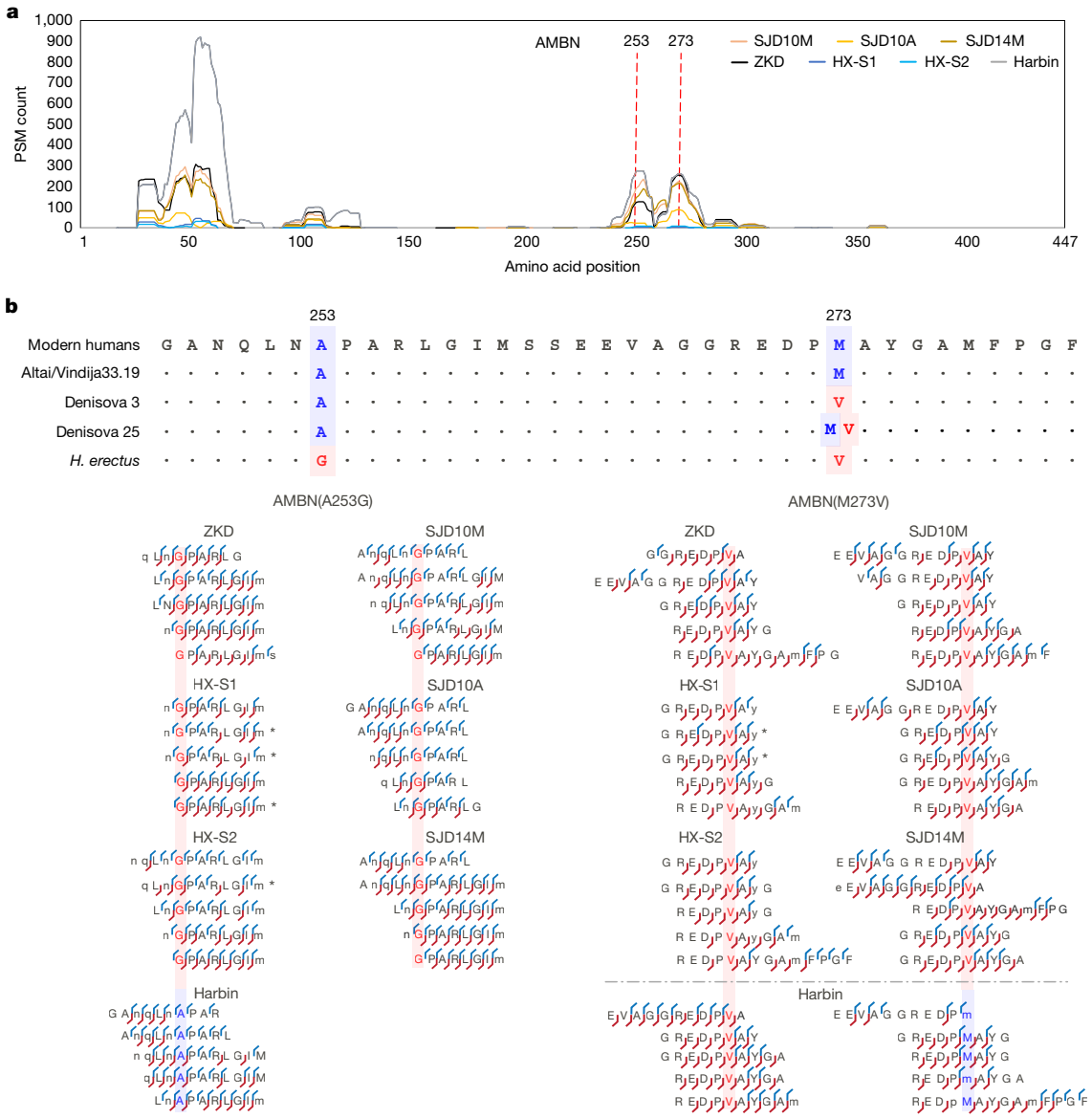


Fig. 3 | PSM coverage and supporting peptides for SAP AMBN(A253G) and AMBN(M273V) in the six *H. erectus* samples and the Harbin individual. **a**, PSM coverage of AMBN. **b**, The alignments of multiple high coverage human sequences, and the small subset of supporting peptides and fragment ions for the SAPs at AMBN 253 and AMBN 273 in the new samples. e, N-terminal pyroglutamic acid (pyro-Glu) from E; m, oxidated M in ZKD and SJD10M, SJD10A, SJD14M, or dioxidated M in HX-S1 and HX-S2; n, deamidated N; q, deamidated Q (not N-terminal Q) or pyro-Glu from Q (N-terminal Q); y, dichlorinated Y in HX-S1, or chlorinated Y in HX-S2. Peptides marked with an asterisk were from pFind, and the others were from PEAKS. The alignment presents a small subset of the supporting peptides.

at most. All SAPs were confirmed in this way using PEAKS Online, pFind and MaxQuant (Methods), which resulted in two confidently identified SAPs.

A new SAP in the ameloblastin protein (AMBN) (A253>G), was identified in all 6 human samples from Zhoukoudian, Hexian, and Sunjiadong, with PSM counts of 124, 3, 9, 191, 24 and 145, and peptide counts of 21, 2, 4, 39, 9 and 35 supporting each sample, respectively (Table 1 and Supplementary Note 5). The highest-quality peptide for this SAP in each of the six samples has at least three y/b ions (Fig. 3b and Extended Data Figs. 2 and 4). Comparative samples with available data covering this position—including mammals, all primates, *Homo antecessor* from Atapuerca[11], *H. erectus* from Dmanisi[11], Denisovan 3, Penghu 1 (ref. 35), all Neandertals and Human Genome Diversity Project (HGDP) modern humans—nearly all carry AMBN A253, with a few mammals having T (horse), S (pig), V (lesser mole-rats) or missing (bovines) residues. Thus, AMBN(A253G), identified in all six human samples from three distinct sites across northern and southern China, representing vastly different geographic regions, has not been previously detected in any other primate species, suggesting that it is likely to be a specific variant for these Middle Pleistocene *H. erectus*. This mutation causes these six Middle Pleistocene *H. erectus* samples from Zhoukoudian, Hexian and Sunjiadong to cluster together in the phylogenetic tree (Supplementary Note 6) (posterior probability 100%; Extended Data Fig. 3). These results also indicate that the Hexian teeth can be attributed to *H. erectus*, and are not Denisovan as has been previously proposed based on morphology[21].

Another SAP, AMBN(M273V), is also present in all six Middle Pleistocene *H. erectus* samples (Table 1, Fig. 3 and Extended Data Fig. 4). The corresponding single-nucleotide polymorphism (SNP allele (rs564905233, A>G)) is found in Denisovans and appears to have contributed to modern humans through Denisovan introgression. It exhibits a frequency of 21% in the Philippines, 1.17% in India, 0.71% in Papua

Table 1 | The details of SAPs within *Homo* identified in six Middle Pleistocene *H. erectus* specimens and other individuals

SAP assignment				East Asian *H. erectus* (new)	Denisovan-related	Confirmation		
Protein (UniProt accession)				AMBN (Q9NP70)				
Position				253	273	PEAKS	pFind	MaxQuant
Data	Species	Sample	Age (Myr)					
New	*H. erectus*	ZKD	0.42	G[a] (124;21)[b]	V[a] (252;52)	√	√	√
New	*H. erectus*	SJD10M	0.4	G[a] (191;39)	V[a] (219;75)	√	√	√
New	*H. erectus*	SJD10A	0.4	G[a] (24;9)	V[a] (87;46)	√	√	√
New	*H. erectus*	SJD14M	0.4	G[a] (145;35)	V[a] (211;69)	√	√	√
New	*H. erectus*	HX-S1	0.41	G[a] (3;2)	V[a] (4;3)	√	√	M273V only (1;1)
New	*H. erectus*	HX-S2	0.41	G[a] (9;4)	V[a] (8;6)	√	√	M273V only (1;1)
Published[11]	*H. erectus*	D4163	1.77	A	M	√	×	√
Published[11]	*H. antecessor*	ATD6-92	0.772–0.949	A	M	√	×	√
Published[43]	Possible Denisovan	TNH-2	0.131–0.164	A	M	√	×	√
New	Denisovan	Harbin	0.146–0.309	A (276;58)	V[a] (211;79) M (49;38)	√	√	√
Published[35]	Denisovan	Penghu 1	0.01–0.07 or 0.13–0.19	A	V[a]	√	√	√
Preprint[47]	Denisovan	Denisova 25	0.2	A	M/V[a]	Genome		
Published[52]	Denisovan	Denisova 3	0.072	A	V[a]	Genome		
Published[45]	Neanderthal	Altai	0.122	A	M	Genome		
Published[53]	Neanderthal	Vindija33.19	0.052	A	M	Genome		
Published[54]	Early modern human	Ust'-Ishim	0.045	A	M	Genome		
Database[30]	Modern human	HGDP	Modern	A	M 99.6%, V[a] 0.4%[c]	Genome		
Database[55]	Modern human	gnomAD	Modern	A (>99.999%)	M 99.993%, V[a] 0.007%[c]	Genome and exome		
Database[56]	Hominids	Great apes	Modern	A	M	Genome, proteome and so on		

Note that peptides shorter than 8 amino acids were discarded in SAP confirmation.
[a]Derived amino acid.
[b]The first number in the parentheses is the number of PSMs, and the second number is the number of peptides in the PEAKS result.
[c]Most individuals with AMBN(M273V) are in Southeast Asia and Oceania Islands, which have high Denisovan introgression.

New Guinea, and is absent from most other modern human populations[44]. Notably, genomic studies reveal that Denisovans received 0.5–8% gene flow from a hominin whose ancestors diverged more than 1 Ma from the common lineage ancestral to Neanderthals, Denisovans and modern humans[45], and about 15% of these 'super-archaic' DNA regions introgressed from Denisovans into Asian and Oceanian individuals[46]. This situation is similar to what we observe with AMBN(M273V). Consequently, this variant is not exclusive to Denisovans but appears to have been introduced into them through a population linked to Middle Pleistocene *H. erectus* from Zhoukoudian, Hexian, and Sunjiadong (Fig. 4).

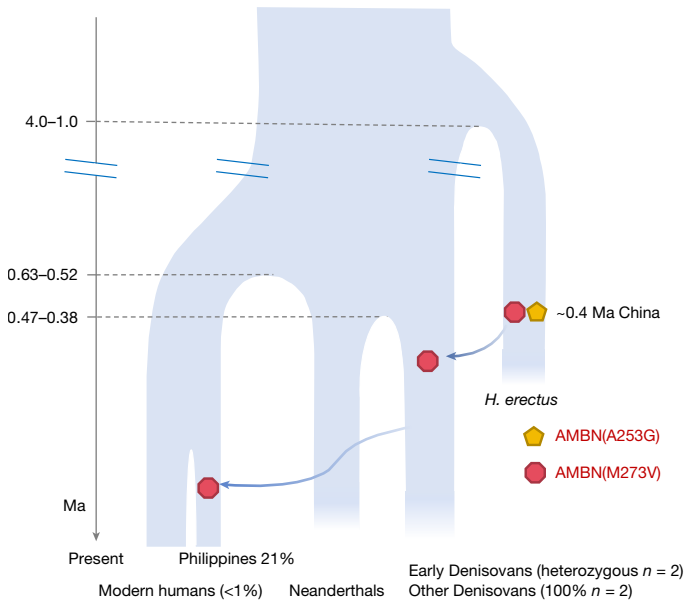


Fig. 4 | A possible model of gene flow related to AMBN(M273V) among *H. erectus* associated with the populations of Zhoukoudian, Hexian and Sunjiadong, Denisovans and modern humans. Neanderthal introgressions are not displayed.

Additionally, the DNA region containing rs564905233, which corresponds to the protein at position 273 of AMBN, shows greater sequence divergence between Africans and Denisova 3 than between Africans and the Altai Neanderthal, as identified through sliding window analysis (Methods, Extended Data Fig. 5). This supports the idea that this region, including rs564905233, may originate from a more diverged super-archaic group. In addition, although the AMBN(M273V) variant is homozygous in the more recent Denisovans, Denisova 3 and Penghu 1, it is heterozygous in both known 'earlier' Denisovans, Harbin (less than 0.15 Ma) and Denisova 25 (around 0.2 Ma)[47] (Table 1, Fig. 3b and Extended Data Fig. 4), indicating that early Denisovans had a lower allele frequency of the derived allele. Furthermore, the morphology of Zhoukoudian and Hexian is distinct from that of the Harbin cranium[48]. Therefore, because of differences in age, morphology and known genetic variants, Zhoukoudian, Hexian and Sunjiadong can be excluded from the early Denisovans.

Discussion and conclusions

Recent research has demonstrated the potential of ancient protein analysis to uncover human evolutionary history[11,35,36,49,50]. However, genetic data about *H. erectus*, a key hominin and the first to travel to

and inhabit geographically diverse regions, are scarce. The enamel proteins obtained from six Middle Pleistocene *H. erectus* fossils from Zhoukoudian, Hexian and Sunjiadong—three well-known sites spanning northern and southern China—permit a rare insight into the genetic makeup of *H. erectus*.

Morphological variability among these three sites, especially the Hexian samples, which differ from those of Zhoukoudian and Sunjiadong, prompted discussion about whether they are more closely related to *H. erectus* from Indonesia or possibly Denisovans[14,15,18,20,21]. Protein evidence clearly shows that the two Hexian specimens possess the newly identified AMBN(A253G) mutation, as do the *H. erectus* specimens from Zhoukoudian and Sunjiadong, and that this mutation is not found in Denisova 3, Penghu 1 or Harbin, placing them within *H. erectus*. Further palaeoproteomic research is needed to explore the molecular diversity within *H. erectus*.

The AMBN(A253G) and AMBN(M273V) variants are potentially specific to the populations to which these six Middle Pleistocene *H. erectus* specimens belonged. Notably, AMBN(A253G) has not been observed in any humans or primates before, making this finding highly notable. These *H. erectus* specimens come from three sites across East Asia, all dating to around 0.4 Ma, and Denisovans diverged from Neanderthals around 0.38–0.47 Ma (ref. 45). The known range of Denisovans includes Siberia, the Tibetan Plateau, Harbin in northeast Asia and the Penghu Channel in Southeast Asia, and these three *H. erectus* sites are from both northern and southern China. Their shared habitats create opportunities for interactions between populations related to these Middle Pleistocene *H. erectus* individuals and Denisovans. The identification of the derived AMBN(M273V) variant in all East Asian *H. erectus* specimens of this study and its occurrence in Denisovans, makes *H. erectus* a candidate source for the super-archaic introgressed DNA identified in the Denisova 3 genome. This variant is likely to originate from populations related to Late Middle Pleistocene *H. erectus* that interacted with Denisovans during the periods when these groups coexisted[51] in East Asia. Further research on *H. erectus*, including molecular data across different periods and regions, will help to clarify their microevolution, population diversity and interactions with Denisovans.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10478-8.

1. Antón, S. C. Natural history of *Homo erectus*. *Am. J. Biol. Anthropol.* **122**, S126–S170 (2003).
2. Herries, A. I. R. et al. Contemporaneity of *Australopithecus*, *Paranthropus*, and early *Homo erectus* in South Africa. *Science* **368**, eaaw7293 (2020).
3. Mussi, M. et al. Early *Homo erectus* lived at high altitudes and produced both Oldowan and Acheulean tools. *Science* **382**, 713–718 (2023).
4. Ferring, R. et al. Earliest human occupations at Dmanisi (Georgian Caucasus) dated to 1.85–1.78 Ma. *Proc. Natl Acad. Sci. USA* **108**, 10432–10436 (2011).
5. Lordkipanidze, D. et al. A complete skull from Dmanisi, Georgia, and the evolutionary biology of early *Homo*. *Science* **342**, 326–331 (2013).
6. Huguet, R. et al. The earliest human face of Western Europe. *Nature* **640**, 707–713 (2025).
7. Kamberov, Y. G. et al. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152**, 691–702 (2013).
8. Zhu, Z. Y. et al. New dating of the *Homo erectus* cranium from Lantian (Gongwangling), China. *J. Hum. Evol.* **78**, 144–157 (2015).
9. Zhu, Z. et al. Hominin occupation of the Chinese Loess Plateau since about 2.1 million years ago. *Nature* **559**, 608–612 (2018).
10. Liu, W., Wu, X. & Xing, S. The morphological evidence for the regional continuity and diversity of Middle Pleistocene human evolution in China. *Acta Archaeol. Sin.* **38**, 473–490 (2019).
11. Welker, F. et al. The dental proteome of *Homo* antecessor. *Nature* **580**, 235–238 (2020).
12. Shen, G., Gao, X., Gao, B. & Granger, D. E. Age of Zhoukoudian *Homo erectus* determined with 26Al/10Be burial dating. *Nature* **458**, 198–200 (2009).
13. Grün, R. et al. ESR analysis of teeth from the palaeoanthropological site of Zhoukoudian, China. *J. Hum. Evol.* **32**, 83–91 (1997).
14. Antón, S. C. Evolutionary significance of cranial variation in Asian *Homo erectus*. *Am. J. Phys. Anthropol.* **118**, 301–323 (2002).
15. Liu, W. & Zhang, Y. The cranial metric diversity of Chinese *Homo erectus*. *Acta Anthropologica Sinica* **24**, 121–136 (2005).
16. Grun, R. et al. ESR and U-series analyses of teeth from the palaeoanthropological site of Hexian, Anhui Province, China. *J. Hum. Evol.* **34**, 555–564 (1998).
17. Zhao, J.-X., Hu, K., Collerson, K. D. & Xu, H.-K. Thermal ionization mass spectrometry U-series dating of a hominid site near Nanjing, China. *Geology* **29**, 27–30 (2001).
18. Zhao, L. et al. Middle Pleistocene hominins from the Sunjiadong site, Luanchuan county, Henan Province, Central China. *Acta Anthropol. Sin.* **37**, 192–205 (2018).
19. Zhang, Y., Zhang, S., Gu, X. & Li, X. A study on the perforated bone from Sunjiadong site in Henan Province. *Acta Anthropol. Sin.* **36**, 457–464 (2017).
20. Liu, W. et al. A mandible from the Middle Pleistocene Hexian site and its significance in relation to the variability of Asian *Homo erectus*. *Am. J. Phys. Anthropol.* **162**, 715–731 (2017).
21. Kaifu, Y. Archaic hominin populations in Asia before the arrival of modern humans: their phylogeny and implications for the "Southern Denisovans". *Curr. Anthropol* **58**, S418–S433 (2017).
22. Jia, L. P. Report on excavations at the Sinanthropus site in 1958. *Palaeovertebrata Paleoanthropoligica* **1**, 21–26 (1959).
23. Woo, J.-K. & Chia, L. New discoveries of *Sinanthropus pekinensis* in Choukoutien. *Acta Palaeontol. Sin.* **2**, 267–288 (1954).
24. Huang, P.-H. et al. Study of ESR dating for burying age of the first skull of Pekinig Man and chronological scale of the cave deposit in Zhoukoudian site Loc. 1. *Acta Anthropol. Sin.* **10**, 107–115 (1991).
25. Pei, J. in *Multi-Disciplinary Study of the Peking Man Site at Zhoukoudian* (eds Wu, R. K. et al.) 258–260 (Science Press, 1985).
26. Huang, W., Fang, D. & Ye, Y. Preliminary studyon the fossil hominid skull and fauna of Hexian, Anhui. *Vertabrata Palasiatica* **20**, 248–256 (1982).
27. Ji, Q., Wu, W., Ji, Y., Li, Q. & Ni, X. Late Middle Pleistocene Harbin cranium represents a new *Homo* species. *Innovation* **2**, 100132 (2021).
28. Ni, X. et al. Massive cranium from Harbin in northeastern China establishes a new Middle Pleistocene human lineage. *Innovation* **2**, 100130 (2021).
29. Patramanis, I., Ramos-Madrigal, J., Cappellini, E. & Racimo, F. PaleoProPhyler: a reproducible pipeline for phylogenetic inference using ancient proteins. *Peer Commun. J.* **3**, e112 (2023).
30. Bergström, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
31. Madupe, P. P. et al. Enamel proteins reveal biological sex and genetic variability in southern African *Paranthropus*. *Science* **388**, 969–973 (2025).
32. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
33. Xin, L. et al. A streamlined platform for analyzing tera-scale DDA and DIA mass spectrometry data enables highly sensitive immunopeptidomics. *Nat. Commun.* **13**, 3108 (2022).
34. Chi, H. et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059–1061 (2018).
35. Tsutaya, T. et al. A male Denisovan mandible from Pleistocene Taiwan. *Science* **388**, 176–180 (2025).
36. Fu, Q. et al. The proteome of the late Middle Pleistocene Harbin individual. *Science* **389**, eadu9677 (2025).
37. Stewart, N. A., Gerlach, R. F., Gowland, R. L., Gron, K. J. & Montgomery, J. Sex determination of human remains from peptides in tooth enamel. *Proc. Natl Acad. Sci. USA* **114**, 13649–13654 (2017).
38. Parker, G. J. et al. Sex estimation using sexually dimorphic amelogenin protein fragments in human enamel. *J. Archaeol. Sci.* **101**, 169–180 (2019).
39. Lugli, F. et al. Enamel peptides reveal the sex of the Late Antique 'Lovers of Modena'. *Sci. Rep.* **9**, 13130 (2019).
40. Welker, F. Elucidation of cross-species proteomic effects in human and hominin bone proteome identification through a bioinformatics experiment. *BMC Evol. Biol.* **18**, 23 (2018).
41. Presslee, S. et al. Palaeoproteomics resolves sloth relationships. *Nat. Ecol. Evol.* **3**, 1121–1130 (2019).
42. Bai, F., Wu, Z., Xing, S. & Fu, Q. Rapid and robust sex determination from ancient enamel proteomes using protSexInferer. *J. Genet. Genomics* https://doi.org/10.1016/j.jgg.2026.04.012 (2026).
43. Demeter, F. et al. A Middle Pleistocene Denisovan molar from the Annamite Chain of northern Laos. *Nat. Commun.* **13**, 2557 (2022).
44. Wall, J. D. et al. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
45. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
46. Hubisz, M. J., Williams, A. L. & Siepel, A. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS Genet.* **16**, e1008895 (2020).
47. Peyrégne, S. et al. A high-coverage genome from a 200,000-year-old Denisovan. Preprint at *bioRxiv* https://doi.org/10.1101/2025.10.20.683404 (2025).
48. Feng, X. et al. The phylogenetic position of the Yunxian cranium elucidates the origin of *Homo longi* and the Denisovans. *Science* **389**, 1320–1324 (2025).
49. Xia, H. et al. Middle and Late Pleistocene Denisovan subsistence at Baishiya Karst Cave. *Nature* **632**, 108–113 (2024).
50. Chen, F. et al. A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature* **569**, 409–412 (2019).
51. Sawafuji, R., Tsutaya, T., Takahata, N., Pedersen, M. W. & Ishida, H. East and Southeast Asian hominin dispersal and evolution: a review. *Q. Sci. Rev.* **333**, 108669 (2024).

52. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
53. Prüfer, K. et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
54. Fu, Q. et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
55. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
56. Patramanis, I., Ramos Madrigal, J., Cappellini, E. & Racimo, F. Hominid Palaeoproteomic Reference Dataset (1.0.1) [Data set]. *Zenodo* https://doi.org/10.5281/zenodo.7728060 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

Methods

Protein extraction

Animal fossils. Powder was drilled from the dentin (Supplementary Data 1), then mixed with 1.5 ml of 0.6 M HCl for decalcification. The precipitate was washed with ultrapure water at least 3 times until the pH reached approximately 7. Then, 200 µl of 50 mM NH_4HCO_3 was added, and the mixture was incubated at 65 °C for 3 h to extract soluble proteins. The supernatant containing the proteins was then transferred into a new centrifuge tube, and 1 µg of trypsin (Promega) was added. The mixture was incubated at 37 °C for 18 h. A 2.5% trifluoroacetic acid (TFA) solution (final concentration 0.1%) was added to stop the reaction. Subsequently, the trypsinized peptides were desalted and purified using C18 ZipTips[57]. For enamel samples, the powder was mixed with 1 ml of 5% HCl[37] for decalcification, and the acid was replaced daily until the reaction ceased. The acid solution containing dissolved enamel peptides was concentrated using a vacuum concentrator, and the peptides were then desalted and purified using C18 ZipTips[58].

Peptides were eluted into a solution of 80% acetonitrile with 0.1% TFA for MALDI-TOF mass spectrometry analysis. All sample preparations were performed in the dedicated clean room at the Molecular Paleontology Laboratory, IVPP of the Chinese Academy of Sciences in Beijing.

Hominin fossils. An acid etching method was used to extract protein from tooth enamel, modified from the process described in ref. 37. Disposable toothbrushes were used to remove surface contaminants from a small area of the enamel for etching. At the same time, the remaining teeth were wrapped with parafilm to prevent contact with any liquids. Before etching, the small enamel area was initially washed with 3% H_2O_2 for 30 s, followed by a rinse with ultrapure water. Approximately 100 µl of 5% (v/v) HCl was placed in the cap of a 1.5-ml microcentrifuge tube. A 2-min etch was performed by immersing the etching region in the HCl solution, and the initial etch solution was discarded. A second etch, lasting 15 min, was carried out in the cap of another separate microcentrifuge tube, and the etch solution was retained. This second etch was repeated, and the etch solutions were combined. After etching, the etched area was treated with 100 µl of 50 mM ammonium bicarbonate solution for 1 min to neutralize the acid. It was then rinsed with ultrapure water for 30 s and dried. The combined etch solution was then desalted using C18 ZipTips (Thermo Fisher Scientific) and eluted into a solution of 0.1% TFA and 80% acetonitrile (ACN). The peptide mixture (50 µl) was further divided into three aliquots; one aliquot was composed of 16 µl, among which 3 µl was used for the MALDI-TOF mass spectrometry test and 13 µl was retained as backup in our laboratory; two aliquots (each composed of 17 µl) were dried for LC–MS/MS analysis in two independent laboratories. All sample preparation for the experiment was conducted in the dedicated clean room at the Molecular Paleontology Laboratory, IVPP of the Chinese Academy of Sciences in Beijing.

MALDI-TOF mass spectrometry analysis

The peptide mixture was analysed on a Bruker autoflex maX MALDI-TOF mass spectrometer. In detail, 1 µl of peptide mixture was spotted onto a MTP384 Bruker ground-steel MALDI target plate, and 1 µl of α-cyano-4-hydroxycinnamic acid matrix solution (1% in 50% ACN/0.1% TFA (v/v/v)) was added on top. They were mixed, dried, and analysed on the mass spectrometer with a *m/z* range of 700–3,500. Each sample was analysed in triplicate. The raw data files were processed by mMass (v5.5.0)[59].

LC–MS/MS analysis

For each hominin fossil, the eluted peptides from the enamel extraction were analysed under DDA mode in triplicate, including two runs on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) at Capital Medical University, Beijing, and one run on an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific) at Fudan University. Both devices were coupled to an Easy nLC 1200 HPLC system (Thermo Fisher Scientific).

For the Orbitrap Fusion Lumos at Capital Medical University, the peptides were initially loaded onto a 100 µm internal diameter × 2 cm trap column and then separated on a 150 µm internal diameter × 15 cm analytical column. Both columns were packed in-house using 3 µm reversed-phase silica (Reprosil-Pur C18 AQ, Dr. Maisch). The peptides were eluted using a 120 min linear gradient programme (0–8 min, 7–11% B; 8–96 min, 11–28% B; 96–108 min, 28–40% B; 108–113 min, 40–90% B; 113–120 min, 90% B) at a flow rate of 500 nl min^{-1}. Buffer A was 0.1% formic acid in water, and buffer B was 80% acetonitrile and 0.1% formic acid. The MS1 data were acquired across 375–1,400 *m/z*, with a resolution of 120k at *m/z* 200, a 250% AGC target, and a maximum injection time of 50 ms. The MS2 scans were performed with a resolution of 15k at *m/z* 200, an AGC target of 100%, a 35% normalized collision energy, and a maximum injection time of 22 ms.

For the Orbitrap Exploris 480 at Fudan University, Shanghai, the peptides were separated on a 75 µm internal diameter × 25 cm analytical column, which was packed in-house using reversed-phase silica of 1.9 µm (Reprosil-Pur C18 AQ, Dr. Maisch). Buffer A was 0.1% formic acid in water, and buffer B was 80% acetonitrile and 0.1% formic acid. An 80 min gradient was used with the following profile: 5–8% B, 2 min, at a flow rate of 200 nl min^{-1}; 8–44% B, 38 min, 200 nl min^{-1}; 44–70% B, 8 min, 200 nl min^{-1}; 70–100% B, 2 min, 200 nl min^{-1}; 100% B, 10 min, 200 nl min^{-1}; 100–5% B, 2 min, 200 nl min^{-1}; 5% B, 2 min, 300 nl min^{-1}; 5–100% B, 6 min, 300 nl min^{-1}; 100% B, 10 min, 300 nl min^{-1}. Full mass spectrometry scans were acquired for the first 65 min, after which the column was washed and re-equilibrated for 15 min without data acquisition. The full mass spectrometry data acquisition was conducted across the range of *m/z* 350–1,600, with a resolution of 60k at *m/z* 200. The AGC target was set to 'standard', and the maximum injection time mode was set to 'auto'. The MS/MS spectra were acquired with a resolution of 15k at *m/z* 200, a maximum injection time of 30 ms and a normalized collision energy of 30%. The AGC target was also set to standard.

For each animal fossil, the eluted peptides from enamel extractions were analysed for one run under DDA mode on the Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific), either at Capital Medical University or Fudan University. For the Orbitrap Fusion Lumos at Capital Medical University, the liquid chromatography gradient and mass spectrometry parameters were the same as those of the hominin fossils. The Orbitrap Fusion Lumos at Fudan University was also interfaced with an Easy nLC 1200 HPLC system (Thermo Scientific). The peptides were separated on a 75 µm internal diameter × 20 cm analytical column packed with 1.9 µm reversed-phase silica. Mobile phase A consisted of 0.1% formic acid, and mobile phase B consisted of 80% acetonitrile and 0.1% formic acid. An 80 min gradient was used with the following profile: 2–5% B, 3 min, at a flow rate of 200 nl min^{-1}; 5–35% B, 40 min, 200 nl min^{-1}; 35–44% B, 5 min, 200 nl min^{-1}; 44–100% B, 2 min, 200 nl min^{-1}; 100% B, 10 min, 200 nl min^{-1}; 100–5% B, 2 min, 200 nl min^{-1}; 5% B, 2 min, 300 nl min^{-1}; 5–100% B, 6 min, 300 nl min^{-1}; 100% B, 10 min, 300 nl min^{-1}. Full mass spectrometry scans were acquired for the first 65 min, after which the column was washed and re-equilibrated for 15 min without data acquisition. The full mass spectrometry data acquisition was conducted across the *m/z* range 350–1,600, with a resolution of 60k at *m/z* 200, a 100% AGC target, and a maximum injection time of 50 ms. The MS/MS spectra were acquired with a resolution of 15k at *m/z* 200, a 100% AGC target, a 30% normalized collision energy, and a maximum injection time of 30 ms. Blank extractions were processed concurrently to monitor the exogenous contaminants during the procedure.

Data search strategy

MaxQuant (v2.6.0.0)[32], PEAKS Online (v12)[33] and pFind (v3.2.1)[34] were used to search the raw data. The *H. erectus* raw files were searched with the corresponding laboratory blanks and modern *H. sapiens* raw

files[43] against the 'Hominidae enamel database', supplemented with the contaminant database. Unspecific digestion was selected in each software. The animal raw files were searched against the 'mammal enamel database'.

Database composition. The 'Hominidae enamel database' comprised 13 selected enamel proteins from Hominidae. Besides the commonly used 12 proteins (AHSG, ALB, AMBN, AMELX, AMELY, AMTN, COL17A1, ENAM, KLK4, MMP20, ODAM and TUFT1), SERPINC1 was also added because we identified this protein in Harbin with abundant peptides and elevated deamidation rates (nearly 100%), and this protein was also reported in modern enamel[60,61]. The sequences were retrieved from UniProt, downloaded from the 'Hominid Palaeoproteomic Reference Dataset' (https://zenodo.org/records/7728060), translated from the genomes of public projects[30], and specific sequences from published palaeoproteomes[11]. The 'mammal enamel database' was composed of mammal sequences of the same enamel proteins above, retrieved from UniProt using the gene names and 'Mammalia (mammals) [40674]'. The contaminant database was composed of a previously published contaminant database[62], the cRAP database (https://www.thegpm.org/crap/), and the contaminant database from MaxQuant (v2.6.0.0). The 'Hominidae enamel database' was accessible through the ProteomeXchange Consortium (Data availability).

PEAKS search. The precursor ion (MS1) mass tolerance was set to 10 ppm, with a fragment ion (MS2) mass tolerance of 0.02 Da for all PEAKS searches, with unspecific digestion. Variable modifications included deamidation (NQ), oxidation (M), hydroxylation (P), phosphorylation (STY), N-terminal pyro-Glu from E, and N-terminal pyro-Glu from Q, with no fixed modifications and up to three modifications allowed per peptide. The peptide length was set to 6–45. PSMs were filtered using a false discovery rate (FDR) of 1%, and proteins were filtered with criteria of $-10\log P \geq 20$ and average local confidence (ALC) ≥ 50% (de novo only).

After our initial search with PEAKS, additional variable modifications were included in the second-round search for Hexian *H. erectus* samples: chlorination and dichlorination of tyrosine residues, dehydration, dioxidation (W), carbonyl E, dioxidation (M), oxidation (HW), ornithine derived from arginine, tryptophan oxidation to kynurenine, tryptophan oxidation to oxolactone, and proline oxidation to pyroglutamic acid. Up to five modifications per peptide were permitted. The peptide length range was set to 6–30.

After removing peptides from the contaminant database and those detected in the extraction blank samples or matching multiple genes, the deamidation rates of glutamine (Q) and asparagine (N) were calculated for each sample based on PSM counts. We used auxiliary tools for PSM prediction with PEAKS.

pFind search. We included Deamidation [N], Deamidation [Q], Oxidation [M], Oxidation [P], Oxidation [W], Gln->pyro-Glu[AnyN-termQ], Glu->pyro-Glu[AnyN-termE], Pro->pyro-Glu[P], Phospho[S], Phospho[T], Phospho[Y], Dehydrated[S], Dehydrated[T], Dehydrated[Y], Arg->Orn[R], Dioxidation[M], Dioxidation[W], Thiazolidine[W], Trp->Kynurenin[W], Trp->Oxolactone[W], Ammonia-loss[N], His->Asp[H], Pro->HAVA[P], and Amidated[AnyC-term] as variable modifications, with no fixed modifications included. Spectra FDR was set at 1%, and protein FDR was 10%. The mass range of each peptide was set from 350 to 4,000 Da. Open search was enabled. Within pFind's modification configuration, the default mass 'X' is preset to that of isoleucine/leucine. To avoid incorrect identification of the X residue, we set its mass to 6,228.71 Da (Sm_{41}), which substantially exceeds the mass of any natural amino acid. Therefore, any in silico peptide with an X produces abnormal theoretical precursor and fragment ion masses, effectively preventing its matching to experimental spectra during database searches and thus reducing false positives from these ambiguous sequence regions. The other parameters were the same as the settings in PEAKS.

After our first-round search with pFind, some additional variable modifications were selected for inclusion in the second-round search for Hexian *H. erectus* samples: Chlorination[Y], dichlorination[Y], Carbonyl[E], and Dioxidation[P](Pro->Glu[P]).

MaxQuant search. No fixed modifications were specified. Variable modifications included Deamidation (NQ), Phosphorylation (STY), Gln to N-terminal pyro-Glu, Glu to N-terminal pyro-Glu, Dioxidation (MW), Oxidation (M), Oxidation (P), and Oxidation (W). PSM FDR was set at 1% for all 7 samples. The search also enabled the identification of dependent and secondary peptides. The remaining parameters were set to their default values. After the search, the deamidation rates of N and Q were calculated[63] for each sample.

In addition, we determined the extra variable modification for Hexian *H. erectus* samples with PEAKS and pFind results. Cl(Y) and diCl (Y) were added for HX-S1 (both post-translational modifications were self-made in Configuration-Modifications), and Cl (Y) and Dehydrated (STY) were included for HX-S2.

Considering the high resolution of the instrumentation used, we also reduced the precursor ion (MS1) mass tolerance to 5 ppm and repeated all analyses in PEAKS and pFind. The results were similar, and the two main SAPs (AMBN 253 and AMBN 273) within the *Homo* genus were consistently identified (Supplementary Table 8). As the final results do not change significantly, we focus mainly on the results from the commonly used 10 ppm search to make our results more comparable to previous studies.

Construction of the consensus protein sequences and phylogenetic analysis

Consensus sequences of endogenous proteins were reconstructed for phylogenetic analysis (Supplementary Data 1). Peptides shorter than eight amino acids or with abnormal or artificial post-translational modifications were excluded for the consensus sequence reconstruction of each endogenous protein. Only alleles with a PSM count ≥2, a PSM ratio ≥10%, and an intensity ratio ≥10% were considered reliable and retained. For heterozygous sites, at least two peptides were required per allele. The heterozygous variant site, observed only in the Harbin specimen, was identified with 79 peptides for the V allele and 38 for the M allele at position 273 in AMBN (Table 1). We include both alleles of the Harbin consensus in Supplementary Data 2. We performed additional AMELY sequence correction as described in Supplementary Note 4.

The protein data in the phylogenetic tree include Denisova 3 (ref. 52), two Neanderthals (Altai, Vindija33.19)[45,53], two modern humans from the HGDP project (San, HGDP0987; Bougainville, HGDP01027)[30], a chimpanzee (*Pan*)[29], and a *Gorilla*[29]. We used PartitionFinder2 (v2.1.1)[64] to identify the best partitioning schemes and amino acid substitution models for our dataset. Then, we built a consensus Bayesian phylogenetic tree using the software MrBayes (v3.2.6)[65], running 8 Markov Chain Monte Carlo (MCMC) chains for 1 million iterations in 2 independent runs. Sampling was done every 500 generations, and the first 200,000 iterations were discarded as burn-in. The tree was plotted with FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree).

DNA analysis

Sliding window analysis. Variant sites were identified where either Denisova 3 or two Neanderthals (Vindija33.19 and Altai) differed from two modern African diploid genomes (S_Khomani_San-1.DG, S_Mandenka-2.DG) from the phased Simons Genome Diversity Project panel (SGDP) (https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data2021)[66,67]. For each of 129 overlapping 20 kb windows (with 2 kb steps), pairwise matching rates were calculated for variants between each African haploid and each archaic artificial haploid

of either Denisova 3 or the Altai Neanderthal (where each variant is randomly assigned to a haploid to preserve all variants), covering in total 276 kb of DNA sequence surrounding the rs564905233 SNP. For each window, statistical significance of the differences between the African–Denisovan and African–Neanderthal matching rates were assessed using a Wilcoxon rank-sum test, and a permutation test (n = 1,000 permutations). The Wilcoxon rank-sum test was used to compare distributions of pairwise matching rates, with W representing the sum of ranks in the African–Denisovan group. The permutation test was used to evaluate differences in mean matching rates. No adjustments were made for multiple comparisons, since each window was analysed independently to identify localized regions of divergence. All statistical tests were performed as two-sided tests. Full results, including W, exact P values, and the number of comparisons per window, are provided in Supplementary Data 3.

Ethics statement

Permission to test for ancient proteins in the human specimens from this study was granted by the collection room of the IVPP, the Hexian Culture, Tourism, and Sports Bureau, and the Luanchuan County Culture, Radio, Television, and Tourism Bureau. The work was conducted in collaboration with local researchers, who are co-authors because of their contributions to assembling archaeological materials and/or discussions that informed the study.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the proteomic mass spectrometry data and searching database have been deposited in the ProteomeXchange Consortium via the PRIDE[68] partner repository with the dataset identifier PXD068897. Protein consensus sequences for six specimens and the Harbin individual are available in Supplementary Data 2. Reference palaeoproteomic data is available through the PRIDE partner repository (https://www.ebi.ac.uk/pride/archive/) with the dataset identifier PXD054412 (the Penghu 1 specimen), PXD040221 (4 *Paranthropus robustus* specimens: SK14132, SK830, SK835 and SK850), PXD014342 (Dmanisi *H. erectus* specimen D4163, and Atapuerca *H. antecessor* specimen ATD6-92), PXD018721 (specimen Laos_TNH2, and modern human specimens 1692 and 1693), PXD009781 (17 modern human specimens) and PXD012587 (3 modern human specimens). The 'Hominid Palaeoproteomic Reference Dataset' is available through the Zenodo repository (https://doi.org/10.5281/zenodo.7728060 (ref. 56)). All other reference enamel protein sequences are available in the UniProt Knowledgebase (https://www.uniprot.org/). Human reference genome hg19 is available through the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) under accession number PRJNA31257, and the corresponding aligned chimpanzee genome is available at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro6. The previously reported ancient DNA datasets used in this study are available through the Allen Ancient DNA Resource v62.0 (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FFIDCW). All modern human genomes from the HGDP project are available at ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.2019051. The worldwide base maps of land, ocean, and rivers are available in Natural Earth (https://www.naturalearthdata.com/).

57. Rao, H. et al. Palaeoproteomic analysis of Pleistocene cave hyenas from east Asia. *Sci. Rep.* **10**, 16674 (2020).
58. Buckley, M., Collins, M., Thomas-Oates, J. & Wilson, J. C. Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **23**, 3843–3854 (2009).
59. Strohalm, M., Kavan, D., Novák, P., Volný, M. & Havlíček, V. mMass 3: a cross-platform software environment for precise analysis of mass spectrometric data. *Anal. Chem.* **82**, 4648–4651 (2010).
60. Jágr, M. et al. Proteomic analysis of dentin–enamel junction and adjacent protein-containing enamel matrix layer of healthy human molar teeth. *Eur. J. Oral Sci.* **127**, 112–121 (2019).
61. Gil-Bona, A. & Bidlack, F. B. Tooth enamel and its dynamic protein matrix. *Int. J. Mol. Sci.* **21**, 4458 (2020).
62. Hendy, J. et al. A guide to ancient protein studies. *Nat. Ecol. Evol.* **2**, 791–799 (2018).
63. Mackie, M. et al. Palaeoproteomic profiling of conservation layers on a 14th century Italian wall painting. *Angew. Chem. Int. Ed. Engl.* **57**, 7369–7374 (2018).
64. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772–773 (2017).
65. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
66. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
67. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Publisher Correction: Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 412 (2021).
68. Y, P.-R. et al. The PRIDE database at 20 years: 2025 update. *Nucleic Acids Res.* **53**, D543–D553 (2025).

Acknowledgements We thank F. Bai, K. Li, L. Duan and Z. Zou for their initial visualization of parts of Figs. 1 and 3b; X. Yang for initial testing; S. Pääbo for comments; J. Kelso and S. Peyrégne for permission to use the data of one position of the Denisova 25 genome; the BSI team for making improvements to PEAKS Online; and Natural Earth for providing the map data used in this study. This work was supported by the National Natural Science Foundation of China (L2424324), the Chinese Academy of Sciences (CAS) (YSBR-019), the Archaeological Talent Promotion Program of China (2024-278) and the New Cornerstone Science Foundation.

Author contributions Q.F. conceived and designed the research project. S.X., Q.J., Z.D., X.G., Y.D. and J.X. prepared the archaeological samples and materials. Q.F., H.R. and K.Z. performed or supervised wet laboratory work. Z.D., X.G., Y.D. and J.X. provided archaeological interpretations. Q.F., Z.W., and X.F. processed the data. Q.F., Z.W. and E.A.B. analysed data. Q.F. wrote the manuscript. Q.F. and E.A.B. revised the manuscript. Q.F., Z.W., E.A.B., H.R. and K.Z. wrote the methods. Q.F., S.X. and Z.W. wrote and prepared the supplementary information. All authors discussed, critically revised and approved the final version of the manuscript.

Competing interests The authors declare no competing interests.

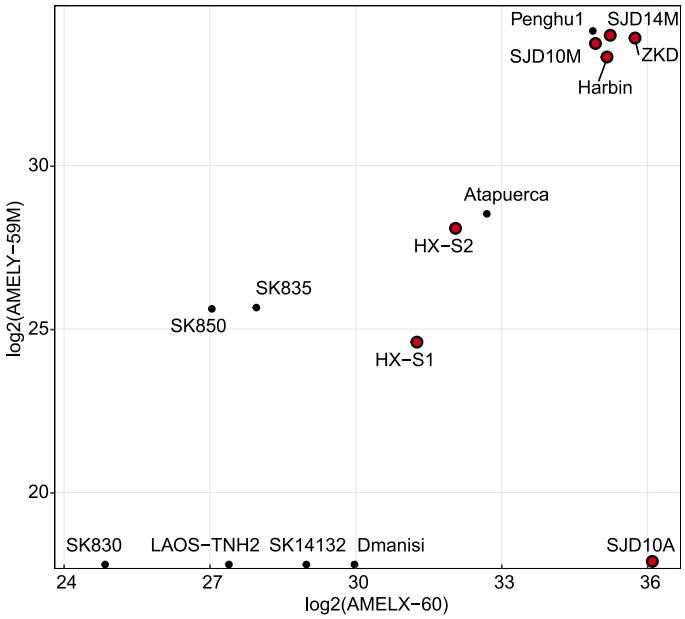
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10478-8.

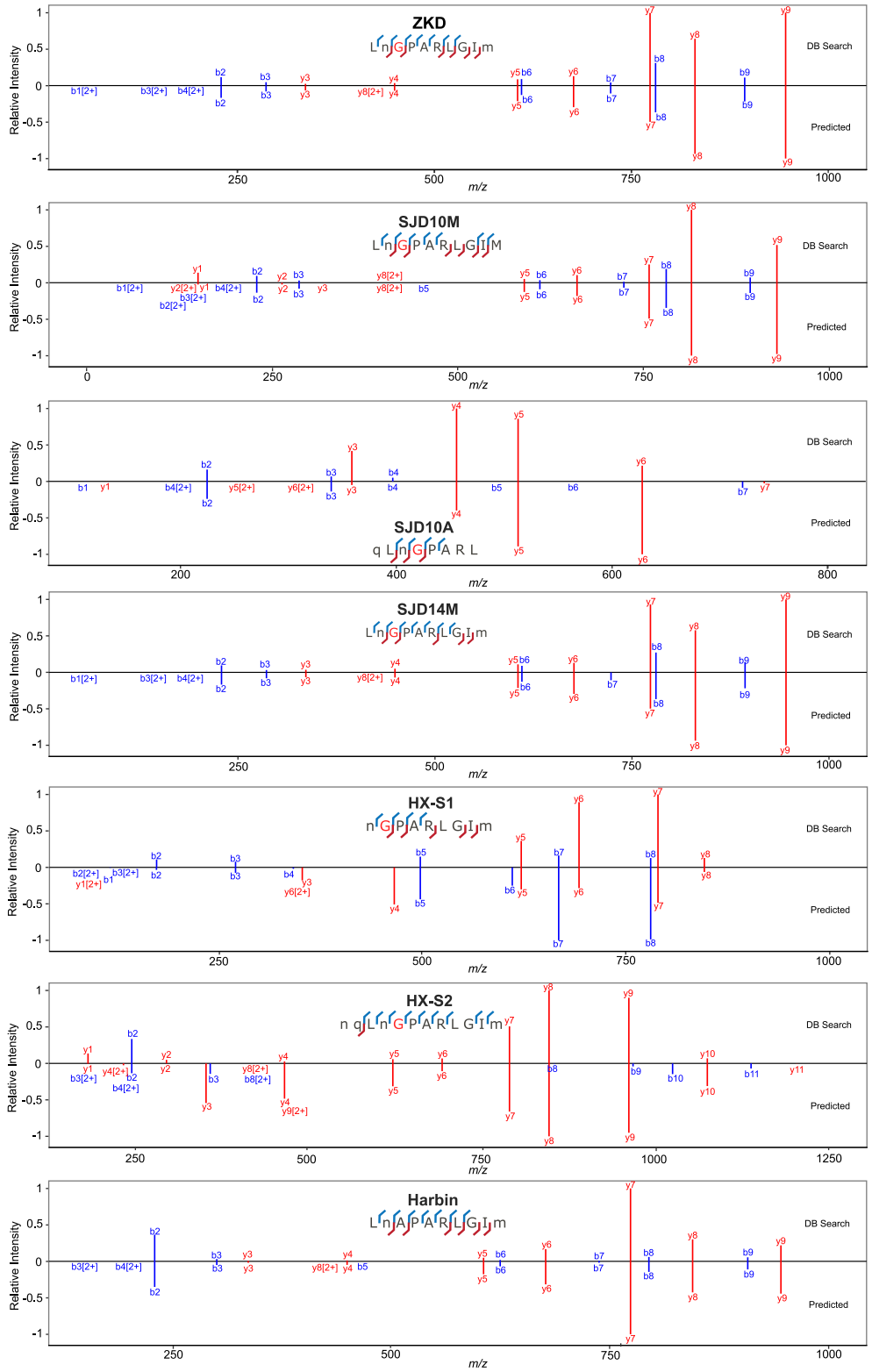
Correspondence and requests for materials should be addressed to Qiaomei Fu.

Peer review information *Nature* thanks Enrico Cappellini who co-reviewed with Palesa Madupe; Daniel Green and Ryan Paterson for their contribution to the peer review of this work. Peer reviewer reports are available.

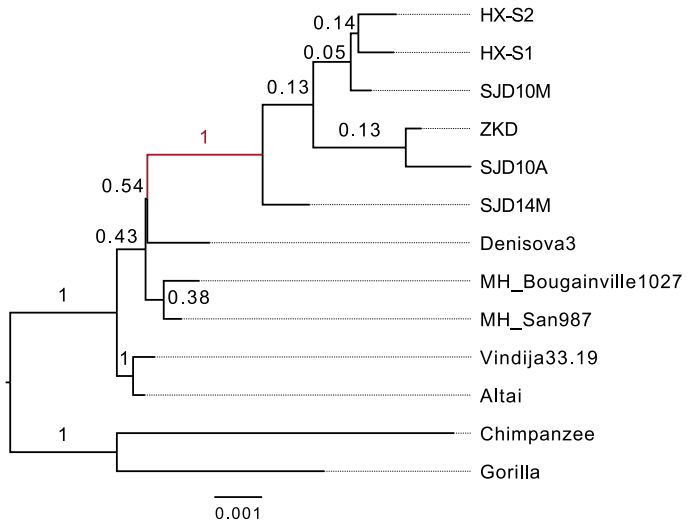
Reprints and permissions information is available at http://www.nature.com/reprints.



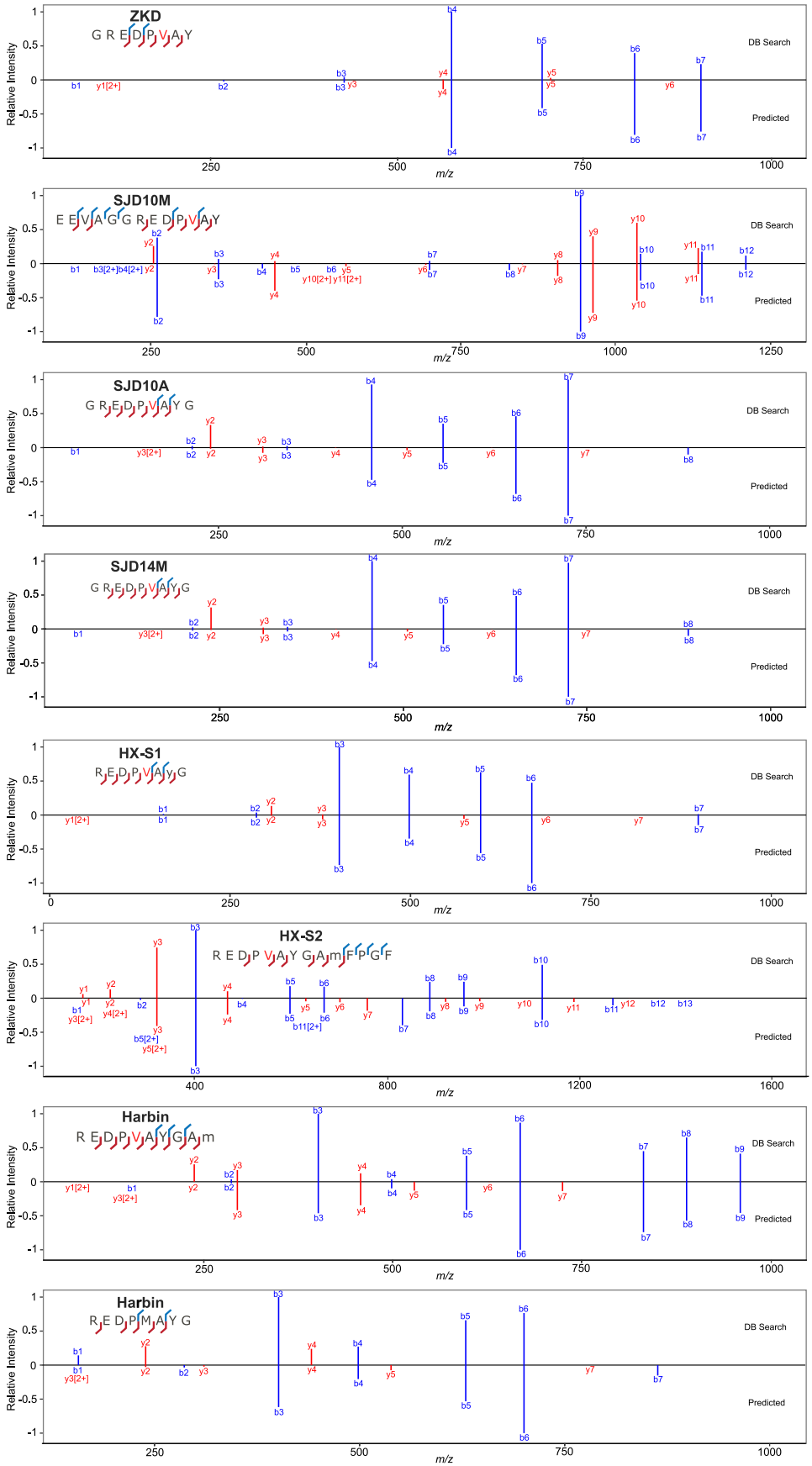
Extended Data Fig. 1 | The intensity of AMELY-59M site as a function of the intensity of AMELX-60 site. The Y-axis and X-axis represent the $\log_2$-transformed sum of intensities of all peptides covering the AMELY-59M and AMELX-60 sites, respectively. Black dots denote reference samples of known sex; red dots denote new samples.



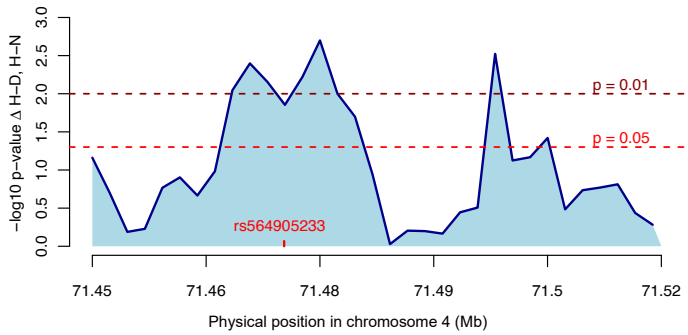
Extended Data Fig. 2 | The mirror plot for the fragment ions of observed and predicted spectra for SAP AMBN−253G in the six *H. erectus* samples and AMBN-253A in the Harbin specimen. n = deamidated N, m = oxidated M (in SJD10M, SJD10A, SJD14M and ZKD) or dioxidated M (in HX-S1 and HX-S2).



Extended Data Fig. 3 | Phylogenetic tree of six Middle Pleistocene *H. erectus* specimens, one Denisovan, two Neanderthals (Vindija33.19 and Altai), an African (MH_San987), an Oceanian (MH_Bougainville1027), a chimpanzee, and a gorilla, using Bayesian analysis. Numbers on the tree indicate posterior probabilities for each clade, and those below 0.2 were not shown. The scale bar represents substitution per site.



Extended Data Fig. 4 | The mirror plot for the fragment ions of observed and predicted spectra for SAP AMBN-273V in the six *H. erectus* samples and AMBN-273V/M in the Harbin specimen. y = dichlorinated Y (in HX-S1), or chlorinated Y (in HX-S2).



Extended Data Fig. 5 | Permutation test p-values for DNA sequence divergence between modern Africans compared with a Denisovan and a Neanderthal. Significance calculated in sliding 20 kb windows with 2 kb steps along a 70 kb segment of chromosome 4 containing rs564905233 (SNP position marked in red). Higher negative log p-values indicate more significant sequence divergence between Africans and Denisova 3 than between Africans and the Altai Neanderthal. The red dotted line shows that SNP rs564905233 is found within a contiguous 34 kb window of sequence divergence $p < 0.05$. The X-axis shows the chromosome 4 coordinates aligning to the center of each window (full results in Supplementary Data 3).

Extended Data Table 1 | Enamel data information for all six H. erectus samples and the Harbin tooth

Sample ID	Before filter			After filter			Consensus position	Sex
	Proteins	Peptides	PSM	Proteins	Peptides	PSM		
ZKD	26	2201	22107	8	2134	21865	497	M
SJD10M	10	2809	15467	9	2808	15466	599	M
SJD10A	8	2015	12520	8	2013	12520	440	F
SJD14M	9	2721	15547	9	2720	15541	578	M
HX-S1	9	652	4097	8	650	4094	269	M
HX-S2	7	821	5182	6	820	5181	310	M
Harbin	19	3476	17852	11	3457	17814	903	M

Data is provided before and after application of the filtering criteria described in the text. Consensus positions are listed for filtered data.

Extended Data Table 2 | Protein composition and coverage details for all six H. erectus samples, the Harbin tooth, and blank controls (_BK)

Sample ID	Peptide number of proteins													Endogenous (%)	Exogenous(%)
	AHSG	ALB	AMBN	AMEL	AMTN	COL17A1	ENAM	KLK4	MMP20	ODAM	SERPINC1	Others (Possible Contaminant)	Sum		
ZKD	2	6	329	1184	0	6	599	0	8	0	0	67	2201	97	3
SJD10M	14	4	528	1542	37	6	675	0	0	2	0	1	2809	100	0
SJD10A	3	5	201	1251	2	7	538	0	0	0	6	2	2015	99.9	0.1
SJD14M	15	4	468	1571	14	11	635	0	0	2	0	1	2721	100	0
HX-S1	0	4	57	483	0	1	103	0	1	1	0	2	652	99.7	0.3
HX-S2	0	3	57	659	0	0	100	0	0	1	0	1	821	99.9	0.1
Harbin	6	186	760	1569	37	61	795	2	1	0	40	19	3476	99.5	0.5
ZKD_BK	0	0	0	0	0	0	0	0	0	0	0	146	146	0	100
SJD_BK	0	0	0	0	0	0	0	0	0	0	0	18	18	0	100
HX_BK	0	0	0	0	0	0	0	0	0	0	0	21	21	0	100

Total proteins recovered from all samples, including possible contaminants, along with the percentage of endogenous proteins. The peptide count for AMEL includes peptides from both AMELX and AMELY. The main proteins in "Others" are typical lab contaminants such as keratins and trypsin.

nature portfolio

Corresponding author(s): Qiaomei Fu

Last updated by author(s): 2026/3/26

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	All software packages used for analysis are cited in the Methods section and in the Supplementary Information and all used packages are publicly available: mMass: http://www.mmass.org version: 5.5.0 PEAKS Online: https://www.bioinfor.com/peaks-online/ version: 12 MaxQuant https://maxquant.org/ version: 2.6.0 pFind https://pfind.ict.ac.cn/ version: 3.2.1 deamidation: https://github.com/dblyon/deamidation version: 1 Partition Finder: http://www.robertlanfear.com/partitionfinder/ version: 2.1.1 Mrbayes: https://nbisweden.github.io/MrBayes version: 3.2.6 x64 FigTree: https://github.com/rambaut/figtree version: 1.4.4 LiftOverVcf: http://broadinstitute.github.io/picard/ version: 2.22.1 protSexInferer: https://github.com/QFuLab/protSexInferer

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All the proteomic mass spectrometry data and searching database have been deposited in the ProteomeXchange Consortium via the PRIDE 65 partner repository with the dataset identifier PXD068897. Protein consensus sequences for six specimens and the Harbin individual are available in Supplementary Data 2. Reference paleoproteomic data is available through the PRIDE partner repository (https://www.ebi.ac.uk/pride/archive/) with the dataset identifier PXD054412 (the Penghu1 specimen), PXD040221 (4 Paranthropus robustus specimens: SK14132, SK830, SK835, and SK850), PXD014342 (Dmanisi H. erectus specimen D4163, and Atapuerca H. antecessor specimen ATD-6-92), PXD018721 (specimen Laos_TNH2, and modern human specimens 1692 and 1693), PXD009781 (17 modern human specimens) and PXD012587 (3 modern human specimens). The 'Hominid Palaeoproteomic Reference Dataset' is available through the Zenodo repository (https://zenodo.org/) with identifier 7728060. All other reference enamel protein sequences are available in the UniProt Knowledgebase (https://www.uniprot.org/). Human reference genome hg19 is available through the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) under accession number PRJNA31257, and the corresponding aligned chimpanzee genome is available at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro6. The previously reported ancient DNA datasets used in this study are available through the Allen Ancient DNA Resource v.62.0 (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FFIDCW). And all modern human genomes from the HGDP project are available at ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.2019051. The worldwide base maps of land, ocean, and rivers are available in Natural Earth (https://www.naturalearthdata.com/).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	*Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data exclusions	
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*
Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Proteomic analysis of the Middle Pleistocene bone specimens
Research sample	Teeth
Sampling strategy	protein extraction
Data collection	All the pretreatment of the sample were conducted by the technicians at the Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology (IVPP), Chinese Academy of Sciences (CAS). And the LC MS/MS analysis was conducted by technicians in State Key Laboratory of Genetic Engineering (Fudan University) and Central Laboratory (Capital Medical University).
Timing and spatial scale	~0.4 Ma, China
Data exclusions	No data was excluded during the study.
Reproducibility	Different fractions were collected and loaded for LC MSMS analysis in two independent labs using different equipments. These results can be cross-validated.
Randomization	Randomization was not relevant to this study.
Blinding	Blinding was not relevant to this study.

Did the study involve field work? ☐ Yes ☒ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*

Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☐	☒ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	Permission was granted to Qiaomei Fu of the IVPP,CAS in 2021, 2024 for the sampling of the specimens for molecular analyses.
Specimen deposition	The collection room of the IVPP, the Hexian Culture, Tourism, and Sports Bureau, and the Luanchuan County Culture, Radio, Television, and Tourism Bureau.
Dating methods	The specimens are dated by U-series in previous study

☒ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	Permission to test for ancient proteins in the human specimens from this study was granted by the collection room of the IVPP, the Hexian Culture, Tourism, and Sports Bureau, and the Luanchuan County Culture, Radio, Television, and Tourism Bureau. The work was conducted in collaboration with local researchers, who are co-authors because of their contributions to assembling archaeological materials and/or discussions that informed the study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

(See Eklund et al. 2016)

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.