LETTER

https://doi.org/10.1038/s41586-019-1555-y

Early Pleistocene enamel proteome from Dmanisi resolves *Stephanorhinus* phylogeny

Enrico Cappellini[1,2]*, Frido Welker[2,3], Luca Pandolfi[4], Jazmín Ramos-Madrigal[2], Diana Samodova[5], Patrick L. Rüther[5], Anna K. Fotakis[2], David Lyon[5], J. Víctor Moreno-Mayar[1], Maia Bukhsianidze[6], Rosa Rakownikow Jersie-Christensen[5], Meaghan Mackie[2,5], Aurélien Ginolhac[7], Reid Ferring[8], Martha Tappen[9], Eleftheria Palkopoulou[10], Marc R. Dickinson[11], Thomas W. Stafford Jr[12], Yvonne L. Chan[13], Anders Götherström[14], Senthilvel K. S. S. Nathan[15], Peter D. Heintzman[16,17], Joshua D. Kapp[16], Irina Kirillova[18], Yoshan Moodley[19], Jordi Agusti[20,21], Ralf-Dietrich Kahlke[22], Gocha Kiladze[23], Bienvenido Martínez-Navarro[20,21,24], Shanlin Liu[2,25], Marcela Sandoval Velasco[2], Mikkel-Holger S. Sinding[2,26], Christian D. Kelstrup[5], Morten E. Allentoft[1], Ludovic Orlando[1,27], Kirsty Penkman[11], Beth Shapiro[16,28], Lorenzo Rook[4], Love Dalén[13], M. Thomas P. Gilbert[2,29], Jesper V. Olsen[5]*, David Lordkipanidze[6] & Eske Willerslev[1,30,31,32]*

The sequencing of ancient DNA has enabled the reconstruction of speciation, migration and admixture events for extinct taxa[1]. However, the irreversible post-mortem degradation[2] of ancient DNA has so far limited its recovery—outside permafrost areas—to specimens that are not older than approximately 0.5 million years (Myr)[3]. By contrast, tandem mass spectrometry has enabled the sequencing of approximately 1.5-Myr-old collagen type I[4], and suggested the presence of protein residues in fossils of the Cretaceous period[5]—although with limited phylogenetic use[6]. In the absence of molecular evidence, the speciation of several extinct species of the Early and Middle Pleistocene epoch remains contentious. Here we address the phylogenetic relationships of the Eurasian Rhinocerotidae of the Pleistocene epoch[7–9], using the proteome of dental enamel from a *Stephanorhinus* tooth that is approximately 1.77-Myr old, recovered from the archaeological site of Dmanisi (South Caucasus, Georgia)[10]. Molecular phylogenetic analyses place this *Stephanorhinus* as a sister group to the clade formed by the woolly rhinoceros (*Coelodonta antiquitatis*) and Merck's rhinoceros (*Stephanorhinus kirchbergensis*). We show that *Coelodonta* evolved from an early *Stephanorhinus* lineage, and that this latter genus includes at least two distinct evolutionary lines. The genus *Stephanorhinus* is therefore currently paraphyletic, and its systematic revision is needed. We demonstrate that sequencing the proteome of Early Pleistocene dental enamel overcomes the limitations of phylogenetic inference based on ancient collagen or DNA. Our approach also provides additional information about the sex and taxonomic assignment of other specimens from Dmanisi. Our findings reveal that proteomic investigation of ancient dental enamel—which is the hardest tissue in vertebrates[11], and is highly abundant in the fossil record—can push the reconstruction of molecular evolution further back into the Early Pleistocene epoch, beyond the currently known limits of ancient DNA preservation.

The phylogenetic placement of extinct species relies increasingly on the sequencing of ancient DNA. Efforts to improve the molecular tools that underlie the recovery of ancient DNA have enabled the reconstruction of approximately 0.4-Myr-old and approximately 0.7-Myr-old DNA sequences from temperate deposits[3] and subpolar regions[12], respectively. However, no ancient DNA data have so far been generated from species that became extinct beyond this time range. By contrast, ancient proteins represent a more-durable source of genetic information, and have been reported[13] to survive (in eggshell) for up to 3.8 Myr. Ancient protein sequences can carry taxonomic and phylogenetic information that is useful for tracing the evolutionary relationships between extant and extinct species[14,15]. However, the recovery of ancient mammal proteins from sites that are too old or too warm to be compatible with the preservation of ancient DNA has so far mostly been limited to collagen type I (COL1). This protein is not an ideal phylogenetic marker, as it is highly conserved[16]. For example, regardless of endogeneity[17], the phylogenetic placement of Dinosauria in relation to extant Aves on the basis of collagen appears to be unstable[6]. This suggests that the exclusive use of COL1 constrains deep-time molecular phylogenetics. Here we sought to overcome these limitations by testing whether dental enamel can be used as an abundant source of larger, and more phylogenetically informative, sets of ancient proteins that are preserved longer than COL1.

The archaeological site of Dmanisi (South Caucasus, Georgia) (Fig. 1a) has been dated to approximately 1.77 Myr ago by a combination of $^{40}Ar/^{39}Ar$ dating, palaeomagnetism and biozonation[18,19]; this age represents a context that is currently considered to be outside the scope of the recovery of ancient DNA. This site has been excavated since 1983, which has resulted in the discovery—along with stone tools and contemporaneous fauna (Supplementary Table 1)—of almost 100 hominin fossils, including 5 skulls, that represent the 'georgicus' palaeodeme within *Homo erectus*[10]. These are the earliest fossils of the genus *Homo* outside of Africa.

The geology of the Dmanisi deposits favours the preservation of faunal material (see Supplementary Information), as the primary aeolian deposits provide rapid burial in fine-grained calcareous sediments. We studied 11 bone, 1 dentine and 14 enamel samples (these enamel samples were occasionally associated with traces of dentine (enamel + dentine))

[1]Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark. [2]Evolutionary Genomics Section, Globe Institute, University of Copenhagen, Copenhagen, Denmark. [3]Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [4]Dipartimento di Scienze della Terra, Università degli Studi di Firenze, Florence, Italy. [5]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark. [6]Georgian National Museum, Tbilisi, Georgia. [7]Life Sciences Research Unit, University of Luxembourg, Belvaux, Luxembourg. [8]Department of Geography and Environment, University of North Texas, Denton, TX, USA. [9]Department of Anthropology, University of Minnesota, Minneapolis, MN, USA. [10]Department of Genetics, Harvard Medical School, Cambridge, MA, USA. [11]Department of Chemistry, University of York, York, UK. [12]Stafford Research, Lafayette, CO, USA. [13]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden. [14]Department of Archaeology and Classical Studies, Stockholm University, Stockholm, Sweden. [15]Sabah Wildlife Department, Kota Kinabalu, Malaysia. [16]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA, USA. [17]Tromsø University Museum, The Arctic University of Norway (UiT), Tromsø, Norway. [18]Ice Age Museum, National Alliance of Shidlovskiy 'Ice Age', Moscow, Russia. [19]Department of Zoology, University of Venda, Thohoyandou, South Africa. [20]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [21]Institut Català de Paleoecologia Humana i Evolució Social, Universitat Rovira i Virgili, Tarragona, Spain. [22]Senckenberg Research Station of Quaternary Palaeontology, Weimar, Germany. [23]Geology Department, Tbilisi State University, Tbilisi, Georgia. [24]Departament d'Història i Geografia, Universitat Rovira i Virgili, Tarragona, Spain. [25]BGI Shenzhen, Shenzen, China. [26]Greenland Institute of Natural Resources, Nuuk, Greenland. [27]Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, Université Paul Sabatier, Toulouse, France. [28]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA, USA. [29]University Museum, Norwegian University of Science and Technology, Trondheim, Norway. [30]Department of Zoology, University of Cambridge, Cambridge, UK. [31]Wellcome Trust Sanger Institute, Hinxton, UK. [32]Danish Institute for Advanced Study, University of Southern Denmark, Odense, Denmark. *e-mail: ecappellini@bio.ku.dk; jesper.olsen@cpr.ku.dk; ewillerslev@bio.ku.dk

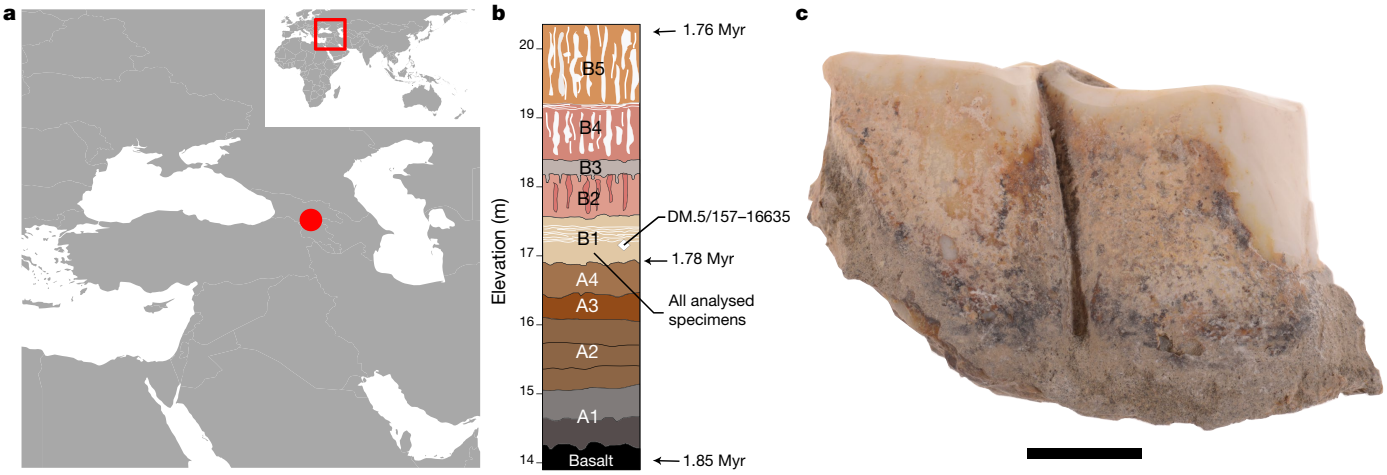


Fig. 1 | Location of Dmanisi, stratigraphy, and specimen Dm.5/157–16635. **a**, Location of Dmanisi in the South Caucasus. The base map was generated using public domain data from www.naturalearthdata.com. **b**, Generalized stratigraphic profile, indicating origin and age of the analysed specimens. Elevation is referred to the local datum. **c**, Isolated left lower molar (m1 or m2) of *Stephanorhinus* ex gr. *etruscus/hundsheimensis*, from Dmanisi (labial view). Scale bar, 1 cm.

from 23 specimens of large mammals from multiple excavation units within stratum B1 (Fig. 1b, Extended Data Fig. 1, Extended Data Table 1, Supplementary Table 3). This is an ashfall deposit that contains faunal remains in a range of geomorphic contexts that are firmly dated to between 1.85 and 1.76 Myr ago[19]. High-resolution tandem mass spectrometry was used to confidently sequence ancient proteins from the set of faunal remains, after proteolytic digestion-based (protocols A and B) or digestion-free (protocol C) preparation of samples (for details of protocols, see Methods, Supplementary Information). Analysis of ancient DNA was attempted, unsuccessfully, on a subset of five bone and dentine specimens (Methods).

We recovered endogenous proteins from 15 out of the 23 specimens that we studied. Digestion-based peptide extraction from bone, dentine and enamel + dentine specimens led to the sporadic recovery (6 out of 19) of a limited number of collagen fragments. By contrast, digestion-free peptide extraction of enamel + dentine and bone specimens resulted in high rates of recovery of the enamel proteome (13 out of 14 specimens) (Extended Data Table 1).

The small proteome[20,21] of mature dental enamel consists of structural proteins (amelogenin (multiple species express the X isoform, AMELX; and males also express the Y isoform, AMELY), enamelin (ENAM), amelotin (AMTN) and ameloblastin (AMBN)) as well as enamel-specific proteases that are secreted during amelogenesis (matrix metalloproteinase-20 (MMP20) and kallikrein 4 (KLK4)). The presence of non-specific proteins—such as serum albumin (ALB)—has also previously been reported in mature dental enamel[20] (Extended Data Table 2). The depth of coverage for these proteins varied considerably across their sequence, with some regions covered by over 1,000 peptide–spectrum matches (Extended Data Fig. 2). The high depth of coverage also enabled us to identify multiple isoforms of AMELX (Extended Data Fig. 3).

Multiple lines of evidence support the authenticity and the endogenous origin of the sequences that we recovered. Dental enamel proteins are extremely tissue-specific, and are confined to this mineral matrix[20]. The amino acid composition of the intra-crystalline protein fraction (measured by amino acid racemization analysis) indicates that the dental enamel behaves as a closed system, and is unaffected by exchanges of amino acids and protein residues with the burial environment (Extended Data Fig. 4). The measured rate of asparagine and glutamine deamidation, which is a spontaneous form of hydrolytic damage that is consistently observed in ancient samples[22], is particularly advanced. Deamidation in enamel from Dmanisi is higher than in the control sample of enamel, which provides support for the antiquity of the peptides that we recovered (Fig. 2a, Supplementary Information). Other forms of non-enzymatic modifications are abundantly present as well. Tyrosine (Y) experienced mono- and di-oxidation, and tryptophan (W) was extensively converted into multiple oxidation products (Fig. 2b, Supplementary Information). The oxidative degradation of histidine (H) and conversion of arginine (R), leading to the accumulation of ornithine, were also observed (Supplementary Information). These modifications are absent or much less frequent in the control sample. Similarly, unlike in the control sample, the distribution of peptide lengths in the Dmanisi dataset is dominated by shorter fragments that are generated by advanced, diagenetically induced terminal hydrolysis[23] (Fig. 2c, d). Together, these independent lines of evidence clearly define the substantial biomolecular damage that has affected the enamel proteomes that we retrieved and independently support the authenticity of the amino acid sequences that we reconstructed. To demonstrate beyond reasonable doubt the correct peptide sequence assignments of our tandem mass spectra, we performed manual validation of peptide–spectrum matches, conducted fragment-ion intensity predictions and generated synthetic peptides for a range of phylogenetically informative and phosphorylated peptides (Methods, Supplementary Data).

We confidently detected site-specific phosphorylation (Fig. 3, Extended Data Figs. 2, 5), a physiological post-translational modification that is highly stable and tightly regulated in vivo and that has previously been detected in dental enamel proteins[24,25]. Most of the phosphorylated sites that we identified belong to the S-X-E or S-X-phosphorylated S motifs, which are recognized by the secreted kinases of the FAM20C family; these kinases are involved in the phosphorylation of extracellular proteins and regulation of biomineralization[26]. Spectra that supported the identification of serine phosphorylation were validated manually and by comparison with tandem mass spectra recorded from synthetic peptides (Supplementary Information), which confirmed the automated identifications from MaxQuant software. Phosphorylated serine and threonine residues may be subjected to spontaneous dephosphorylation. However, by complexing with the Ca^{2+} ions in the enamel hydroxyapatite matrix, the peptide-bound phosphate groups can remain stable over millennia, as recently observed for ancient bone[27]. Previous studies have demonstrated that when complexed with the mineral matrix, approximately 3.8-Myr-old protein remains can be retrieved from sub-tropical environments[13]. The limited availability of free water in the enamel matrix further reduces spontaneous dephosphorylation via β-elimination. These observations demonstrate that the heavily modified proteome of the dental enamel retrieved from the approximately 1.77-Myr-old faunal material from Dmanisi is endogenous and almost complete.

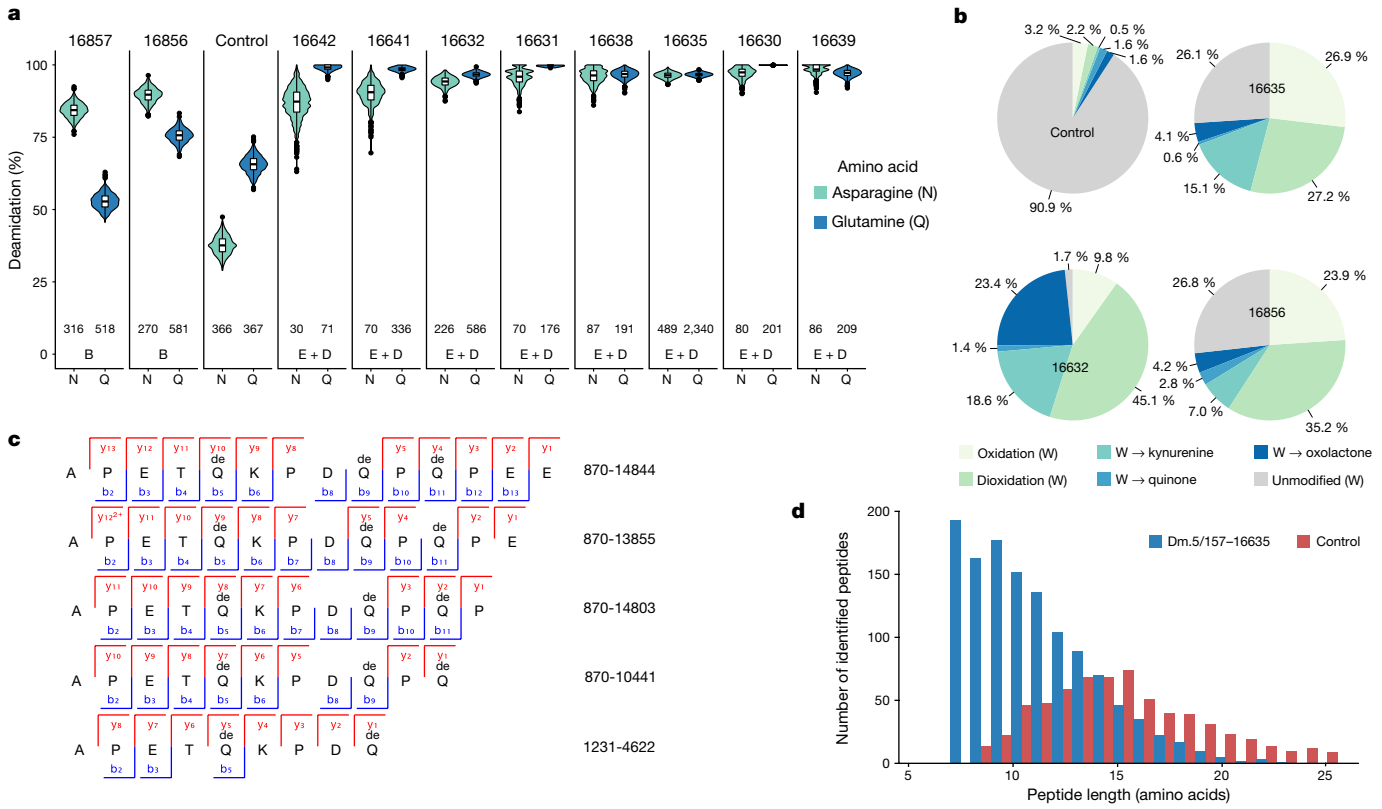


Fig. 2 | Degradation of the enamel proteome. a, Deamidation of asparagine (N) and glutamine (Q). Violin plots are based on 1,000 bootstrap replicates. The box plots define the range of the data, with whiskers extending to 1.5× the interquartile range, boxes showing the 25th to 75th percentiles and dots indicate the median. The tissue source is indicated: B, bone; D, dentine; and E, enamel. The number of peptides used for the calculation are shown at the bottom of the plots. Numbers above the plots refer to the CGG reference numbers for the specimens. **b**, Extent of tryptophan (W) oxidation leading to several diagenetic products, measured as relative spectral counts. **c**, Alignment of deamidated (de) peptides (positions 124–137, ENAM) retrieved by digestion-free acid demineralization from specimen Dm.5/157–16635. Numbers to the right refer to stage-tip (see Supplementary Table 3 in the Supplementary Information) and MS/MS scan (see 'Data availability' in Methods) numbers. **d**, Bar plot of distribution of peptide lengths in undigested proteomes from the dental enamel of specimen Dm.5/157–16635 and of a control (mediaeval sheep or goat).

We used the proteome-sequence information that we recovered to improve taxonomic assignment and achieve sex attribution for some of the faunal remains from Dmanisi. Phylogenetic analysis of the five largest enamel + dentine proteomes, and of a moderately large bone proteome, enabled us to confirm or improve the morphological identification of their specimens of origin (Extended Data Fig. 6, Supplementary Figs. 10–15). Our confident identification of peptides specific for isoform Y of amelogenin (AMELY), which is encoded on the non-recombinant portion of the Y chromosome, indicates that four tooth specimens—Dm.6/151.4.A4.12–16630 (*Pseudodama*) (Dm. code refers to the accession number in the Georgian National Museum (GNM) and the appended five-digit number refers to the reference code of the sample from the Centre for GeoGenetics (CGG)), Dm.69/64.3.B1.53–16631 (Cervidae), Dm.8/154.4.A4.22–16639 (Bovidae) and Dm.M6/7.II.296–16856 (Cervidae)—belonged to male individuals[21] (Extended Data Fig. 7a–d).

An enamel + dentine fragment from specimen Dm.5/157–16635, a lower molar assigned to a *Stephanorhinus* of the group that includes *Stephanorhinus etruscus* and *Stephanorhinus hundsheimensis* (*Stephanorhinus* ex gr. *etruscus/hundsheimensis*) (Fig. 1c, Supplementary Information), returned the highest proteome-sequence coverage, which encompassed a total of 875 amino acids across 987 peptides (6 proteins) (Extended Data Fig. 2, Supplementary Information). Following the alignment of the enamel protein sequences retrieved from Dm.5/157–16635 against their homologues from all extant rhinoceros species plus the extinct woolly rhinoceros (*C. antiquitatis*) and Merck's rhinoceros (*S. kirchbergensis*), phylogenetic reconstructions place the Dmanisi specimen closer to these two extinct rhinoceroses than to the extant Sumatran rhinoceros (*Dicerorhinus sumatrensis*), as an early divergent sister lineage (Fig. 4, Extended Data Fig. 8).

Our phylogenetic reconstruction confidently recovered the expected differentiation of the *Rhinoceros* genus from other genera that we considered, and is consistent with previous cladistic[28] and genetic analyses[29] (Supplementary Information). This topology defines two-horned

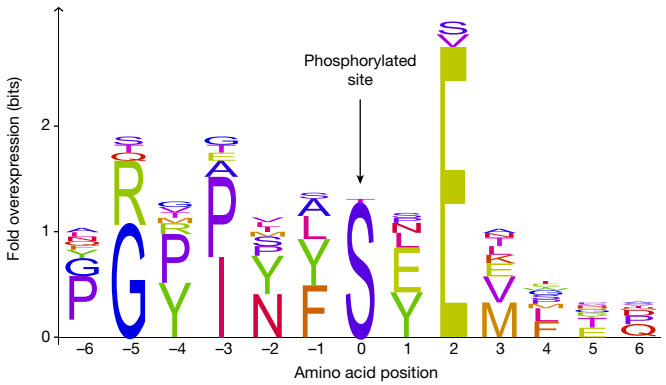


Fig. 3 | Sequence motif analysis of phosphorylation sites in the proteome of ancient enamel. The over-representation of specific amino acids within six positions of the N and C termini of the phosphorylated amino acid (position 0) is indicated. Extended Data Figure 5 provides tandem mass spectra examples of both S-X-E and S-X-phosphorylated S motifs.

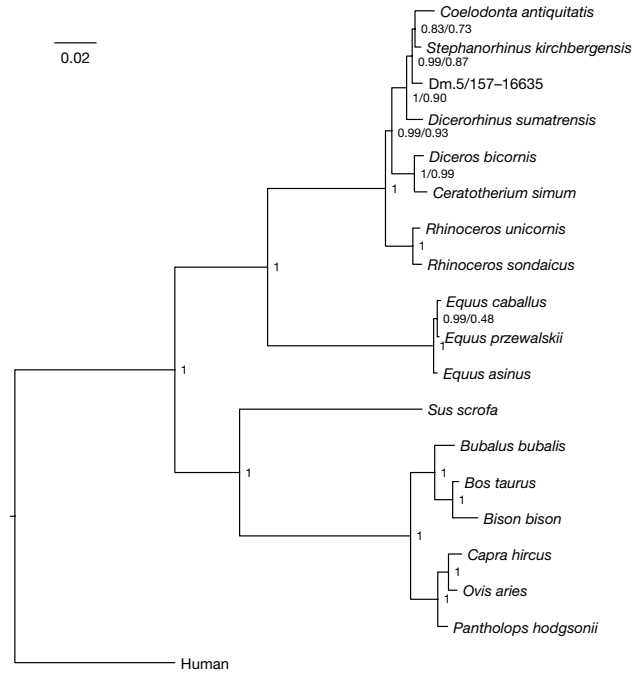


Fig. 4 | Phylogenetic relationships between the comparative dataset of enamel proteomes and specimen Dm.5/157–16635. Consensus tree from Bayesian inference on the concatenated alignment of six enamel proteins, using *Homo sapiens* as an outgroup. For each bipartition, we show the posterior probability obtained from the Bayesian inference. For bipartitions for which the Bayesian and the maximum-likelihood inference support differs, we show the support obtained using the latter on the right. Scale indicates the estimated branch lengths.

rhinoceroses as monophyletic and the one-horned condition as plesiomorphic, as previously proposed (Supplementary Information). We caution, however, that the higher-level relationships that we observe between the rhinoceros monophyletic clades might be affected by demographic events such as incomplete lineage sorting[30] and/or gene flow between groups[31], owing to the limited number of markers that we considered. A confident and stable reconstruction of the structure of the Rhinocerotidae family needs the strong support that only high-resolution whole-genome sequencing can provide. Regardless, the highly supported placement of the Dmanisi rhinoceros in the clade formed by *Stephanorhinus*, the woolly rhinoceros and the Sumatran rhinoceros will remain unaffected, should the deeper phylogenetic relationships between the *Rhinoceros* genus and other family members be revised (Extended Data Fig. 8).

The phylogenetic relationships of the genus *Stephanorhinus* within the family Rhinocerotidae—as well as those of the several species recognized within this genus—are contentious. *Stephanorhinus* was initially included in the extant southeast-Asian genus *Dicerorhinus*, which is represented by *D. sumatrensis*[32]. This hypothesis has been rejected, and *Stephanorhinus* has been identified on the basis of morphological data as a sister taxon of the woolly rhinoceros[33]. Furthermore, analysis of ancient DNA supports a sister relationship between the woolly rhinoceros and *D. sumatrensis*[7,34,35].

As the *Stephanorhinus* ex gr. *etruscus*/*hundsheimensis* sequences from Dmanisi branch off basal to the common ancestor of the woolly rhinoceros and Merck's rhinoceros, these two species most probably derived from an early *Stephanorhinus* lineage that expanded eastward from western Eurasia. Throughout the Pliocene and Pleistocene epochs, *Coelodonta* adapted to continental and, later, to cold-climate habitats in central Asia. The earliest representative of this genus, *Coelodonta thibetana*, displayed some clear *Stephanorhinus*-like anatomical features[33]. The genus *Stephanorhinus* was present in eastern Europe and Anatolia[35] at least since the late Miocene epoch, and the Dmanisi specimen most probably represents an Early Pleistocene descendent of the western-Eurasian branch of this genus.

Our phylogenetic reconstructions show that, as currently defined, the genus *Stephanorhinus* is paraphyletic, which is consistent with previous morphological and palaeo-biogeographical evidence (Supplementary Information). Accordingly, a systematic revision of the genera *Stephanorhinus* and *Coelodonta*, as well as their closest relatives, is needed.

In this study, we show that the mass spectrometric sequencing of the enamel proteome can overcome the time limitations of the preservation of ancient DNA, as well as the reduced phylogenetic content of COL1 sequences. Given the abundance of teeth in the palaeontological record, the approach presented here holds the potential to address a wide range of questions that pertain to the Early and Middle Pleistocene evolutionary history of a large number of mammals (including hominins), at least in temperate climates.

Online content

Any methods, additional references, Nature Research 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/s41586-019-1555-y.

Received: 16 October 2018; Accepted: 12 August 2019;
Published online: 11 September 2019

1. Cappellini, E. et al. Ancient biomolecules and evolutionary inference. *Annu. Rev. Biochem.* **87**, 1029–1060 (2018).
2. Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harb. Perspect. Biol.* **5**, a012567 (2013).
3. Meyer, M. et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531**, 504–507 (2016).
4. Wadsworth, C. & Buckley, M. Proteome degradation in fossils: investigating the longevity of protein survival in ancient bone. *Rapid Commun. Mass Spectrom.* **28**, 605–615 (2014).
5. Schweitzer, M. H. et al. Analyses of soft tissue from *Tyrannosaurus rex* suggest the presence of protein. *Science* **316**, 277–280 (2007).
6. Schroeter, E. R. et al. Expansion for the *Brachylophosaurus canadensis* collagen I sequence and additional evidence of the preservation of Cretaceous protein. *J. Proteome Res.* **16**, 920–932 (2017).
7. Willerslev, E. et al. Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. *BMC Evol. Biol.* **9**, 95 (2009).
8. Welker, F. et al. Middle Pleistocene protein sequences from the rhinoceros genus *Stephanorhinus* and the phylogeny of extant and extinct Middle/Late Pleistocene Rhinocerotidae. *PeerJ* **5**, e3033 (2017).
9. Kirillova, I. et al. Discovery of the skull of *Stephanorhinus kirchbergensis* (Jäger, 1839) above the Arctic Circle. *Quat. Res.* **88**, 537–550 (2017).
10. Lordkipanidze, D. et al. A complete skull from Dmanisi, Georgia, and the evolutionary biology of early *Homo*. *Science* **342**, 326–331 (2013).
11. Eastoe, J. E. Organic matrix of tooth enamel. *Nature* **187**, 411–412 (1960).
12. Orlando, L. et al. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78 (2013).
13. Demarchi, B. et al. Protein sequences bound to mineral surfaces persist into deep time. *eLife* **5**, e17092 (2016).
14. Welker, F. et al. Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature* **522**, 81–84 (2015).
15. Chen, F. et al. A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature* **569**, 409–412 (2019).
16. Nei, M. *Molecular Evolutionary Genetics* Vol. 75, 39–63 (Columbia Univ. Press, 1987).
17. Buckley, M., Warwood, S., van Dongen, B., Kitchener, A. C. & Manning, P. L. A fossil protein chimera; difficulties in discriminating dinosaur peptide sequences from modern cross-contamination. *Proc. R. Soc. Lond. B* **284**, 20170544 (2017).
18. Gabunia, L. et al. Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: taxonomy, geological setting, and age. *Science* **288**, 1019–1025 (2000).
19. 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).
20. Castiblanco, G. A. et al. Identification of proteins from human permanent erupted enamel. *Eur. J. Oral Sci.* **123**, 390–395 (2015).
21. Stewart, N. A. et al. The identification of peptides by nanoLC-MS/MS from human surface tooth enamel following a simple acid etch extraction. *RSC Advances* **6**, 61673–61679 (2016).
22. van Doorn, N. L., Wilson, J., Hollund, H., Soressi, M. & Collins, M. J. Site-specific deamidation of glutamine: a new marker of bone collagen deterioration. *Rapid Commun. Mass Spectrom.* **26**, 2319–2327 (2012).

23. Catak, S., Monard, G., Aviyente, V. & Ruiz-López, M. F. Computational study on nonenzymatic peptide bond cleavage at asparagine and aspartic acid. *J. Phys. Chem. A* **112**, 8752–8761 (2008).
24. Hunter, T. Why nature chose phosphate to modify proteins. *Phil. Trans. R. Soc. Lond. B* **367**, 2513–2516 (2012).
25. Hu, J. C. C., Yamakoshi, Y., Yamakoshi, F., Krebsbach, P. H. & Simmer, J. P. Proteomics and genetics of dental enamel. *Cells Tissues Organs* **181**, 219–231 (2005).
26. Tagliabracci, V. S. et al. Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. *Science* **336**, 1150–1153 (2012).
27. Cleland, T. P. Solid digestion of demineralized bone as a method to access potentially insoluble proteins and post-translational modifications. *J. Proteome Res.* **17**, 536–542 (2018).
28. Antoine, P.-O. et al. A revision of *Aceratherium blanfordi* Lydekker, 1884 (Mammalia: Rhinocerotidae) from the Early Miocene of Pakistan: postcranials as a key. *Zool. J. Linn. Soc.* **160**, 139–194 (2010).
29. Steiner, C. C. & Ryder, O. A. Molecular phylogeny and evolution of the Perissodactyla. *Zool. J. Linn. Soc.* **163**, 1289–1303 (2011).
30. Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H. & Mailund, T. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* **21**, 349–356 (2011).
31. Rieseberg, L. H. Evolution: replacing genes and traits through hybridization. *Curr. Biol.* **19**, R119–R122 (2009).
32. Guérin, C. *Les Rhinocéros (Mammalia, Perissodactyla) du Miocène Terminal au Pleistocène Supérieur en Europe occidentale, Comparaison avec les Espèces Actuelles (Documents du Laboratoire de Geologie de la Faculte des Sciences de Lyon*, volume 79) (Univ. Claude-Bernard, 1980).
33. Deng, T. et al. Out of Tibet: Pliocene woolly rhino suggests high-plateau origin of Ice Age megaherbivores. *Science* **333**, 1285–1288 (2011).
34. Orlando, L. et al. Ancient DNA analysis reveals woolly rhino evolutionary relationships. *Mol. Phylogenet. Evol.* **28**, 485–499 (2003).
35. Yuan, J. et al. Ancient DNA sequences from *Coelodonta antiquitatis* in China reveal its divergence and phylogeny. *Sci. China Earth Sci.* **57**, 388–396 (2014).

METHODS

Dmanisi and sample selection. Dmanisi is located about 65 km southwest of Tbilisi, in the Kvemo Kartli region of Georgia, at an elevation of 910 m above sea level (41° 20′ N, 44° 20′ E)[10,18]. The 23 fossil specimens that we analysed were retrieved from stratum B1, in excavation blocks M17, M6, block 2 and area R11 (Extended Data Fig. 1, Extended Data Table 1). Stratum B deposits date to between 1.78 and 1.76 Myr ago[19]. All of the specimens that we analysed were collected between 1984 and 2014, and their taxonomic identification was based on traditional comparative anatomy.

After the sample preparation and data acquisition for the Dmanisi specimens was concluded, we applied the whole experimental procedure to a mediaeval-period enamel + dentine specimen from a sheep or goat (ovicaprine); this was used as control. For this sample, we used extraction protocol C (see 'Extraction protocol C (digestion-free acid demineralization)') and generated tandem mass spectrometry data using a Q Exactive HF mass spectrometer (Thermo Fisher Scientific). The data were searched against the goat proteome, downloaded from the NCBI Reference Sequence Database (RefSeq) archive on 31 May 2017 (Supplementary Information). The ovicaprine specimen was found at the 'Hotel Skandinavia' site (Århus) and stored at the Natural History Museum of Denmark (Copenhagen).

Biomolecular preservation. We assessed the potential for preservation of ancient proteins before proteomic analysis by measuring the extent of amino acid racemization in a subset of samples (6 out of 23)[36]. Enamel chips with all dentine removed were powdered, and two subsamples per specimen were subjected to analysis of their free and total hydrolysable amino acid fractions. Samples were analysed in duplicate by reverse-phase high-performance liquid chromatography, with standards and blanks run alongside each of them (Supplementary Information). The values of the D over L ratios of aspartic acid plus asparagine, glutamic acid plus glutamine, phenylalanine and alanine were assessed (Extended Data Fig. 4) to provide an overall estimate of intra-crystalline protein decomposition.

Proteomics. All of the sample preparation procedures for mass spectrometric analysis of ancient proteins were conducted in laboratories dedicated to the analysis of ancient DNA and ancient proteins, in clean rooms fitted with filtered ventilation and positive pressure (consistent with recent recommendations for ancient protein analysis[37]). A mock 'extraction blank', containing no starting material, was prepared, processed and analysed together with each batch of ancient samples.

Sample preparation. The external surface of bone samples was gently removed, and the remaining material was subsequently powdered. Enamel fragments, occasionally mixed with small amounts of dentine, were removed from teeth with a cutting disc and subsequently crushed into a rough powder. Unless otherwise specified, ancient protein residues were extracted from approximately 180–220 mg of mineralized material using three different extraction protocols: protocols A, B and C (see below and Supplementary Information for more detailed descriptions of protocols).

Extraction protocol A (filter-aided sample preparation). Tryptic peptides were generated using a filter-aided sample preparation approach[38], as previously performed on ancient samples[39].

Extraction protocol B (GuHCl solution and digestion). Bone or enamel + dentine powder was demineralized in 1 ml 0.5 M EDTA pH 8.0. After removal of the supernatant, all demineralized pellets were resuspended in a 300 μl solution containing 2 M guanidine hydrochloride (GuHCl, Thermo Scientific), 100 mM Tris pH 8.0, 20 mM 2-chloroacetamide (CAA), 10 mM Tris (2-carboxyethyl)phosphine (TCEP) in ultrapure water[40,41]. A total of 0.2 μg of mass-spectrometry-grade rLysC (Promega P/N V1671) enzyme was added before the samples were incubated for 3–4 h at 37 °C with agitation. Samples and negative controls were subsequently diluted to 0.6 M GuHCl, and 0.8 μg of mass-spectrometry-grade trypsin (Promega P/N V5111) was added. Next, samples and negative controls were incubated overnight under mechanical agitation at 37 °C. On the following day, samples were acidified, and the tryptic peptides were purified on C18 stage-tips, as previously described[42].

Extraction protocol C (digestion-free acid demineralization). Dental enamel powder, with possible trace amounts of dentine, was demineralized in 1.2 M HCl at room temperature, after which the solubilized protein residues were directly cleaned and concentrated on stage-tips (Supplementary Information, section 5.1). The sample prepared on stage-tip no. 1217 was processed with 10% trifluoroacetic acid (TFA) instead of 1.2 M HCl. All other parameters and procedures were identical to those used for all other samples extracted with protocol C.

Tandem mass spectrometry. Different sets of samples (Supplementary Information, sections 5.1, 5.2) were analysed by nanoflow liquid chromatography coupled to tandem mass spectrometry (MS/MS) on an EASY-nLC 1000 or 1200 system connected to a Q-Exactive, a Q-Exactive Plus or a Q-Exactive HF (Thermo Scientific) mass spectrometer. Before and after each MS/MS run measuring ancient or extraction blank samples, two successive MS/MS runs were included in the sample queue to prevent carryover contamination between the samples. These consisted of an MS/MS run with an injection exclusively of the buffer used to resuspend the samples (0.1% TFA and 5% acetonitrile) ('MS/MS blank'), followed by a second MS/MS run with no injection ('MS/MS wash').

Data analysis. Raw data files generated during MS/MS spectral acquisition were searched using MaxQuant[43] version 1.5.3.30 and PEAKS[44] version 7.5. A two-stage peptide–spectrum matching approach was adopted (Supplementary Information, section 5.3). Raw files were initially searched against a target and reversed database of collagen and enamel proteins retrieved from the UniProt and RefSeq archives[45,46], taxonomically restricted to mammalian species. A database of partial COL1A1 and COL1A2 sequences from cervid species[47] was also included. The results from the preliminary analysis were used for a first provisional reconstruction of protein sequences (MaxQuant search 1, MQ1).

For specimens with a dataset that resulted in a narrower—although not fully resolved—initial taxonomic placement, a second MaxQuant search (MQ2) was performed using a new protein database taxonomically restricted to the 'order' taxonomic rank, as determined after MQ1. For the MQ2 matching of the MS/MS spectra from specimen Dm.5/157–16635, partial sequences of serum albumin and enamel proteins from Sumatran rhinoceros (*D. sumatrensis*), Javan rhinoceros (*Rhinoceros sondaicus*), Indian rhinoceros (*Rhinoceros unicornis*), woolly rhinoceros (*C. antiquitatis*), Merck's rhinoceros (*S. kirchbergensis*) and black rhinoceros (*Diceros bicornis*) were also added to the protein database. All of the protein sequences from these species were reconstructed from draft genomes for each species (Supplementary Information, L. Dalen and M. T. P. Gilbert, unpublished data).

For each MaxQuant and PEAKS search, enzymatic digestion was set to 'unspecific' and the following variable modifications were included: oxidation (for M), deamidation (for N and Q), N-terminal pyro-Glu (for Q), N-terminal pyro-Glu (for E), hydroxylation (for P) and phosphorylation (for S). The error tolerance was set to 5 ppm for the precursor and to 20 ppm (or 0.05 Da) for the fragment ions in MaxQuant and PEAKS, respectively. For searches of data generated from sample fractions partially or exclusively digested with trypsin, another MaxQuant and PEAKS search was conducted using the 'enzyme' parameter set to 'Trypsin/P'. Carbamidomethylation (for C) was set: (i) as a fixed modification, for searches of data generated from sets of sample fractions exclusively digested with trypsin or (ii) as a variable modification, for searches of data generated from sets of sample fractions partially digested with trypsin. For searches of data generated exclusively from undigested sample fractions, carbamidomethylation was not included as a modification.

The datasets that were re-analysed with MQ2 search were also processed with the PEAKS software using the entire workflow (PEAKS de novo to PEAKS SPIDER) to detect hitherto unreported single-amino-acid polymorphisms. Any amino acid substitution detected by the SPIDER homology search algorithm was validated by repeating the MaxQuant search (MQ3). In MQ3, the protein database used for MQ2 was modified to include the amino acid substitutions detected by the SPIDER algorithm.

Reconstruction of ancient protein sequences. The peptide sequences confidently identified by the MQ1, MQ2 and MQ3 were aligned using the software Geneious[48] (v.5.4.4, substitution matrix BLOSUM62). The peptide sequences confidently identified by the PEAKS searches were aligned using an in-house-generated R-script. A consensus sequence for each protein from each specimen was generated in FASTA format, without filtering on depth of coverage. Amino acid positions that were not confidently reconstructed were replaced by an X. Newly identified single-amino-acid polymorphisms discovered through PEAKS were only accepted if these were further validated by repeating the MaxQuant search (MQ3). All isoleucines were converted into leucines, as standard MS/MS cannot differentiate between these two isobaric amino acids. For possible deamidated sites, we checked whether there were positions in our reference sequence database at which both Q and E or both N and D occurred in the same position, and for which we also had ancient sequences matching. For specimen Dm.5/157–16635, only one such position existed, and this was replaced by an X in our consensus sequence. Based on parsimony, for other Q, E, N and D positions we called the amino acid present in the reference proteome, regardless of the phylogenetic relevance. The output of MQ2 and MQ3 was used to extend the coverage of the ancient protein sequences initially identified in MQ1. For specimen DM.5/157–16335, all of the experimentally identified peptides—as well as the respective best-matching MS/MS spectra covering the sites informative for Rhinocerotidae phylogenetic inference—are provided as Supplementary Data. All of the reported MS/MS spectra are annotated using the advanced annotation mode of MaxQuant. Selected spectra matching peptides that cover phylogenetically informative amino acid positions were manually inspected, validated and annotated by an experienced mass spectrometrist, in all cases in full agreement with bioinformatic sequence assignment (Supplementary Data). We used MS^{2}PIP fragment-ion spectral-intensity prediction[49] (version v.20190312, model version 20190107 HCD) to demonstrate that the experimentally observed fragment-ion intensities are highly correlated with the theoretical ones (Supplementary Fig. 3). Finally, we generated synthetic peptides for 19 selected peptides covering Rhinocerotidae single-amino-acid polymorphisms in DM.5/157–16635.

Post-translational modifications. *Deamidation.* After removal of likely contaminants, the extent of glutamine and asparagine deamidation was estimated for

individual specimens using the MaxQuant output files, as previously published[41] (Supplementary Information).

Other spontaneous chemical modifications. Spontaneous post-translational modifications (PTMs) associated with chemical protein damage were searched using the PEAKS PTM tool and the dependent-peptides search mode[50] in MaxQuant. In the PEAKS PTM search, all modifications in the Unimod database were considered. The mass error was set to 5.0 ppm and 0.5 Da for precursor and fragment, respectively. For PEAKS, the de novo average-local-confidence (ALC) score was set to a threshold of 15% and the peptide hit threshold was set to 30. The results were filtered with a false discovery rate of 5%, de novo ALC score of 50% and a protein hit threshold of ≥ 20. The MaxQuant dependent-peptide search was carried out with the same search settings as described in 'Data analysis' and with a dependent-peptide false discovery rate of 1% and a mass bin size of 0.0065 Da.

Phosphorylation. Class I phosphorylation sites were selected with localization probabilities of ≥ 0.98 in the Phospho(ST)Sites MaxQuant output file. Sequence windows of ± 6 amino acids from all identified sites were compared against a background file containing all unphosphorylated peptides, using a linear kinase sequence motif enrichment analysis in IceLogo (version 1.3.8)[51].

Phylogenetic analysis. *Reference datasets.* We assembled a reference dataset that consisted of publicly available protein sequences from representative ungulate species belonging to the following families: Equidae, Rhinocerotidae, Suidae and Bovidae (Supplementary Information, sections 7, 8). As Cervidae and carnivores are absent from protein sequence databases to varying extents, we did not attempt phylogenetic placement of samples from these taxa. Instead, we conducted our phylogenetic analysis on the five best-performing enamel proteomes (Dm.5/154.2.A4.38–16632, Dm.5/157–16635, Dm.5/154.1.B1.1–16638, Dm.8/154.4.A4.22–16639 and Dm.8/152.3.B1.2–16641) and the largest bone proteome (Dm.bXI.North.B1a.collection–16658) that we recovered (Extended Data Table 2).

We extended this dataset with the protein sequences from extinct and extant rhinoceros species, including woolly rhinoceros (*C. antiquitatis*), Merck's rhinoceros (*S. kirchbergensis*), Sumatran rhinoceros (*D. sumatrensis*), Javan rhinoceros (*R. sondaicus*), Indian rhinoceros (*R. unicornis*) and black rhinoceros (*D. bicornis*). Their corresponding protein sequences were obtained following translation of high-throughput DNA sequencing data, after filtering reads with mapping quality lower than 30 and nucleotides with base quality lower than 20, and calling the majority rule consensus sequence using ANGSD[52]. For the woolly rhinoceros and Merck's rhinoceros, we excluded the first and last five nucleotides of each DNA fragment to minimize the effect of post-mortem damage to the ancient DNA[53]. Each consensus sequence was formatted as a separate blast nucleotide database. We then performed a tblastn[54] alignment using the corresponding white rhinoceros sequence as a query, favouring ungapped alignments to recover translated and spliced protein sequences. The resulting alignments were processed using ProSplign algorithm from the NCBI Eukaryotic Genome Annotation Pipeline[55] to recover the spliced alignments and translated protein sequences.

Construction of phylogenetic trees. For each specimen, multiple sequence alignments for each protein were built using MAFFT[56] and concatenated onto a single alignment per specimen. These were inspected visually to correct obvious alignment mistakes, and all of the isoleucine residues were substituted with leucine ones to account for indistinguishable isobaric amino acids at the positions at which the ancient protein carried one of these amino acids. On the basis of these alignments, we inferred the phylogenetic relationship between the ancient samples and the species included in the reference dataset using three approaches: distance-based neighbour joining, maximum-likelihood and Bayesian phylogenetic inference (Supplementary Information).

Neighbour-joining trees were built using the phangorn[57] R package, restricting to sites covered in the ancient samples. Genetic distances were estimated using the JTT model, considering pairwise deletions. We estimated bipartition support through a non-parametric bootstrap procedure using 500 pseudoreplicates. We used PHyML 3.1[58] for maximum-likelihood inference on the basis of the whole concatenated alignment. For likelihood computation, we used the JTT substitution model with two additional parameters for modelling rate heterogeneity and the proportion of invariant sites. Bipartition support was estimated using a non-parametric bootstrap procedure with 500 replicates. Bayesian phylogenetic inference was carried out using MrBayes 3.2.6[59] on each concatenated alignment, partitioned per gene. Although we chose the JTT substitution model in the two approaches above, we allowed the Markov chain to sample parameters for the substitution rates from a set of predetermined matrices, as well as the shape parameter of a gamma distribution for modelling across-site rate variation and the proportion of invariable sites. The Markov chain Monte Carlo algorithm was run with 4 chains for 5,000,000 cycles. Sampling was conducted every 500 cycles and the first 25% was discarded as burn-in. Convergence was assessed using Tracer v.1.6.0, which estimated an effective sample size greater than 5,500 for each individual, which indicates that there was reasonable convergence for all runs.

Analysis of ancient DNA. The samples were processed using strict ancient DNA guidelines in a clean laboratory facility at the Natural History Museum of Denmark (University of Copenhagen). DNA extraction was attempted on five of the ancient animal samples (Supplementary Information, sections 9, 13). Powdered samples (120–140 mg) were extracted using a silica-in-solution method[12,60]. To prepare the samples for next-generation sequencing, 20 μl of DNA extract was built into a blunt-end library using the NEBNext DNA Sample Prep Master Mix Set 2 (E6070) with Illumina-specific adapters. The libraries were PCR-amplified with inPE1.0 forward primers and custom-designed reverse primers with a six-nucleotide index[61]. Two extracts (MA399 and MA2481, from specimens D4–16859 and Dm.5/157–16635, respectively) yielded detectable DNA concentrations (Supplementary Table 9). The libraries generated from specimen 16859 and 16635 were processed on different flow cells. They were pooled with others for sequencing on an Illumina 2000 platform (MA399_L1 and MA399_L2) using 100-bp single-read chemistry, and on an Illumina 2500 platform (MA2481_L1) using 81-bp single-read chemistry.

The data were base-called using the Illumina software CASAVA 1.8.2 and sequences were demultiplexed with a requirement of a full match of the six nucleotide indexes that were used. Raw reads were processed using the PALEOMIX pipeline following published guidelines[62], mapping against the cow nuclear genome (*Bos taurus* 4.6.1, accession GCA_000003205.4), the cow mitochondrial genome (*Bos taurus*), the red deer mitochondrial genome (*Cervus elaphus*, accession AB245427.2) and the human nuclear genome (GRCh37/hg19) using BWA backtrack[63] v.0.5.10 with the seed disabled. All other parameters were set as default. PCR duplicates from mapped reads were removed using the picard tool MarkDuplicate (http://picard.sourceforge.net/).

Morphological measurements of specimen Dm.5/157–16635. We followed a previously published methodology[32]. The maximal length of the tooth was measured with a digital calliper at the lingual side of the tooth and parallel to the occlusal surface. All measurements are given in mm (Supplementary Information, section 3).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All of the mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD011008. Genomic BAM files used for Rhinocerotidae protein sequence translation and protein sequence alignments used for phylogenetic reconstruction are available on Figshare (https://doi.org/10.6084/m9.figshare.7212746).

Code availability

The in-house R script used to align the peptide sequences confidently identified by the PEAKS searches is available to everyone upon request to the corresponding authors.

36. Penkman, K. E. H., Kaufman, D. S., Maddy, D. & Collins, M. J. Closed-system behaviour of the intra-crystalline fraction of amino acids in mollusc shells. *Quat. Geochronol.* **3**, 2–25 (2008).
37. Hendy, J. et al. A guide to ancient protein studies. *Nat. Ecol. Evol.* **2**, 791–799 (2018).
38. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).
39. Cappellini, E. et al. Resolution of the type material of the Asian elephant, *Elephas maximus* Linnaeus, 1758 (Proboscidea, Elephantidae). *Zool. J. Linn. Soc.* **170**, 222–232 (2014).
40. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
41. Mackie, M. et al. Palaeoproteomic profiling of conservation layers on a 14th century Italian wall painting. *Angew. Chem. Int. Edn* **57**, 7369–7374 (2018).
42. Cappellini, E. et al. Proteomic analysis of a Pleistocene mammoth femur reveals more than one hundred ancient bone proteins. *J. Proteome Res.* **11**, 917–926 (2012).
43. 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).
44. Zhang, J. et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **11**, M111.010587 (2012).
45. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
46. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
47. Welker, F. et al. Palaeoproteomic evidence identifies archaic hominins associated with the Châtelperronian at the Grotte du Renne. *Proc. Natl Acad. Sci. USA* **113**, 11162–11167 (2016).

48. Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
49. Gabriels, R., Martens, L. & Degroeve, S. Updated MS^2PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Res.* **47**, W295–W299 (2019).
50. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protocols* **11**, 2301–2319 (2016).
51. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6**, 786–787 (2009).
52. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
53. Briggs, A. W. et al. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).
54. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
55. Sea Urchin Genome Sequencing Consortium. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941–952 (2006).
56. Katoh, K. & Frith, M. C. Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics* **28**, 3144–3146 (2012).
57. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
58. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
59. Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
60. Rohland, N. & Hofreiter, M. Comparison and optimization of ancient DNA extraction. *Biotechniques* **42**, 343–352 (2007).
61. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, pdb.prot5448 (2010).
62. Schubert, M. et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protocols* **9**, 1056–1082 (2014).
63. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
64. Dickinson, M. R., Lister, A. M. & Penkman, K. E. H. A new method for enamel amino acid racemization dating: a closed system approach. *Quat. Geochronol.* **50**, 29–46 (2019).

Acknowledgements E.C. and F.W. are supported by the VILLUM FONDEN (grant number 17649) and by the European Commission through a Marie Skłodowska Curie (MSC) Individual Fellowship (grant number 795569). E.W. is supported by the Lundbeck Foundation, the Danish National Research Foundation, the Novo Nordisk Foundation, the Carlsberg Foundation, KU2016 and the Wellcome Trust. E.C., C.K., J.V.O., P.R. and D.S. are supported by the European Commission through the MSC European Training Network 'TEMPERA' (grant number 722606). M.M. and R.R.J.-C. are supported by the University of Copenhagen KU2016 (UCPH Excellence Programme) grant. M.M. is also supported by the Danish National Research Foundation award PROTEIOS (DNRF128). Work at the Novo Nordisk Foundation Center for Protein Research is funded in part by a donation from the Novo Nordisk Foundation (grant number NNF14CC0001). M.R.D. is supported by a PhD DTA studentship from NERC and the Natural History Museum (NE/K500987/1 & NE/L501761/1). K.P. is supported by the Leverhulme Trust (PLP -2012-116). L.R. and L.P. are supported by the Italian Ministry for Foreign Affairs (MAECI, DGSP-VI). L.P. was also supported by the EU-SYNTHESYS project (AT-TAF-2550, DE-TAF-3049, GB-TAF-2825, HU-TAF-3593 and ES-TAF-2997) funded by the European Commission. L.D. is supported by the Swedish Research Council (grant number 2017-04647) and FORMAS (grant number 2015-676). M.T.P.G. is supported by ERC Consolidator Grant 'Extinction genomics' (grant number 681396). L.O. is supported by the ERC Consolidator Grant 'PEGASUS' (grant agreement number 681605). B.S., J.K. and P.D.H. are supported by the Gordon and Betty Moore foundation. B.M.-N. is supported by the Spanish Ministry of Sciences (grant number CGL2016-80975-P) and the Generalitat de Catalunya, Spain (grant number 2017SGR 859). J.A. is supported by the Spanish Ministry of Sciences (grant number CGL2016-80000-P). R.F. is supported by National Science Foundation (grant number 1025245). The ancient DNA analysis was carried out using the facilities of the University of Luxembourg, the Swedish Museum of Natural History and UC Santa Cruz. We acknowledge support from the Science for Life Laboratory, the National Genomics Infrastructure (Sweden) and UPPMAX for providing assistance with massive parallel sequencing and computational infrastructure. Research at Dmanisi is supported by the John Templeton Foundation (grant number 52935), and the Shota Rustaveli Science Foundation (grant number 18-27262). We thank B. Triozzi and K. Murphy Gregersen for technical support.

Author contributions E.C., D. Lordkipanidze and E.W. designed the study. A.K.F., M.M., R.R.J.-C., M.E.A., M.R.D., K.P. and E.C. performed laboratory experiments. M.B., M.T., R.F., E.P., T.W.S. Jr, Y.L.C., A. Götherström, S.K.S.S.N., P.D.H., J.D.K., I.K., Y.M., J.A., R.-D.K., G.K., B.M.-N., M.-H.S.S., S.L., M.S.V., B.S., L.D., M.T.P.G. and D. Lordkipanidze provided ancient samples or modern reference material. E.C., F.W., L.P., J.R.-M., D. Lyon, J.V.M.-M., D.S., C.D.K., A. Ginolhac, L.O., L.R., J.V.O., P.L.R., M.R.D. and K.P. performed analyses and data interpretation. E.C., F.W., J.R.-M., L.P. and E.W. wrote the manuscript with contributions from all authors.

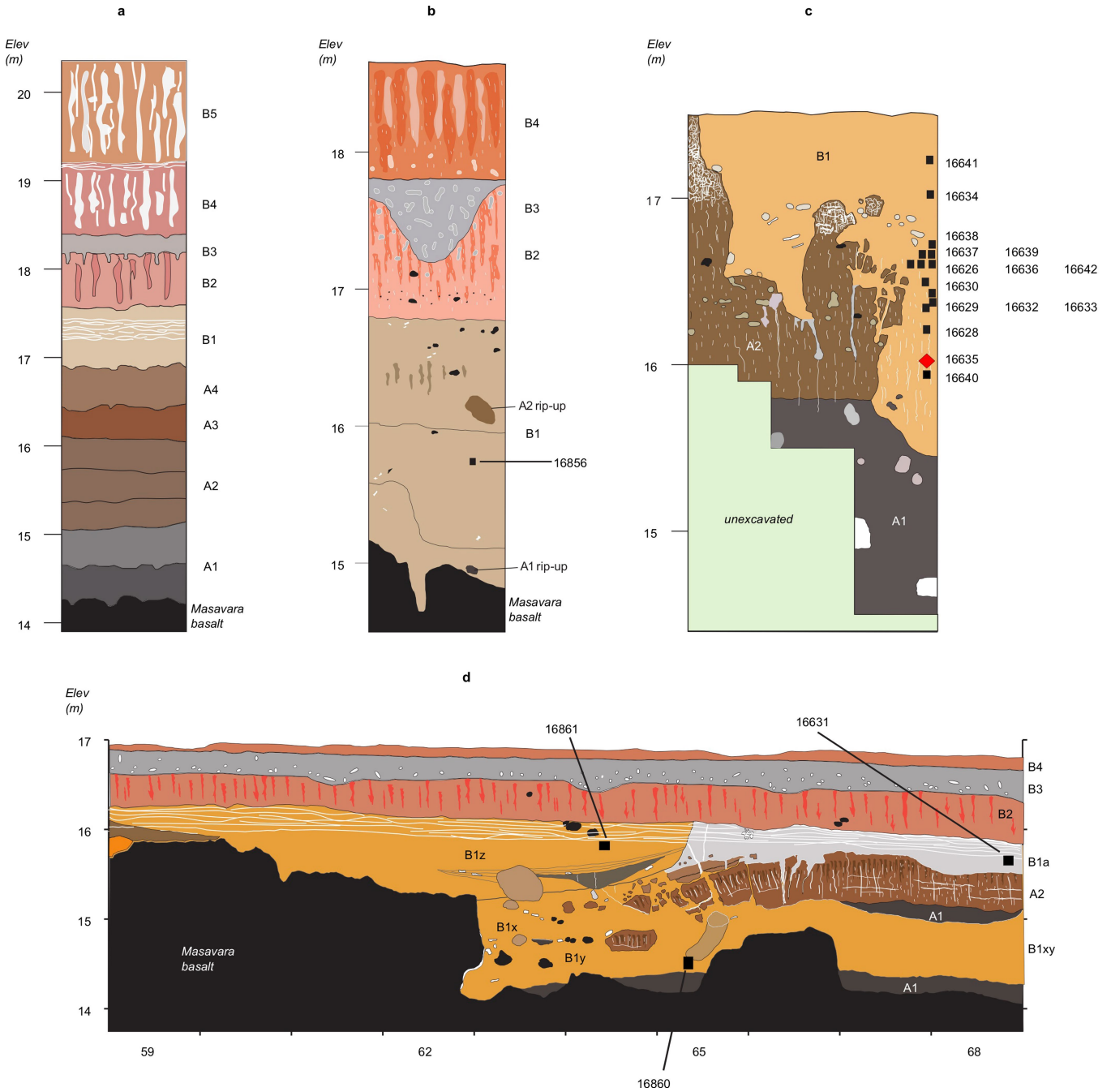
Competing interests The authors declare no competing interests.

Additional information

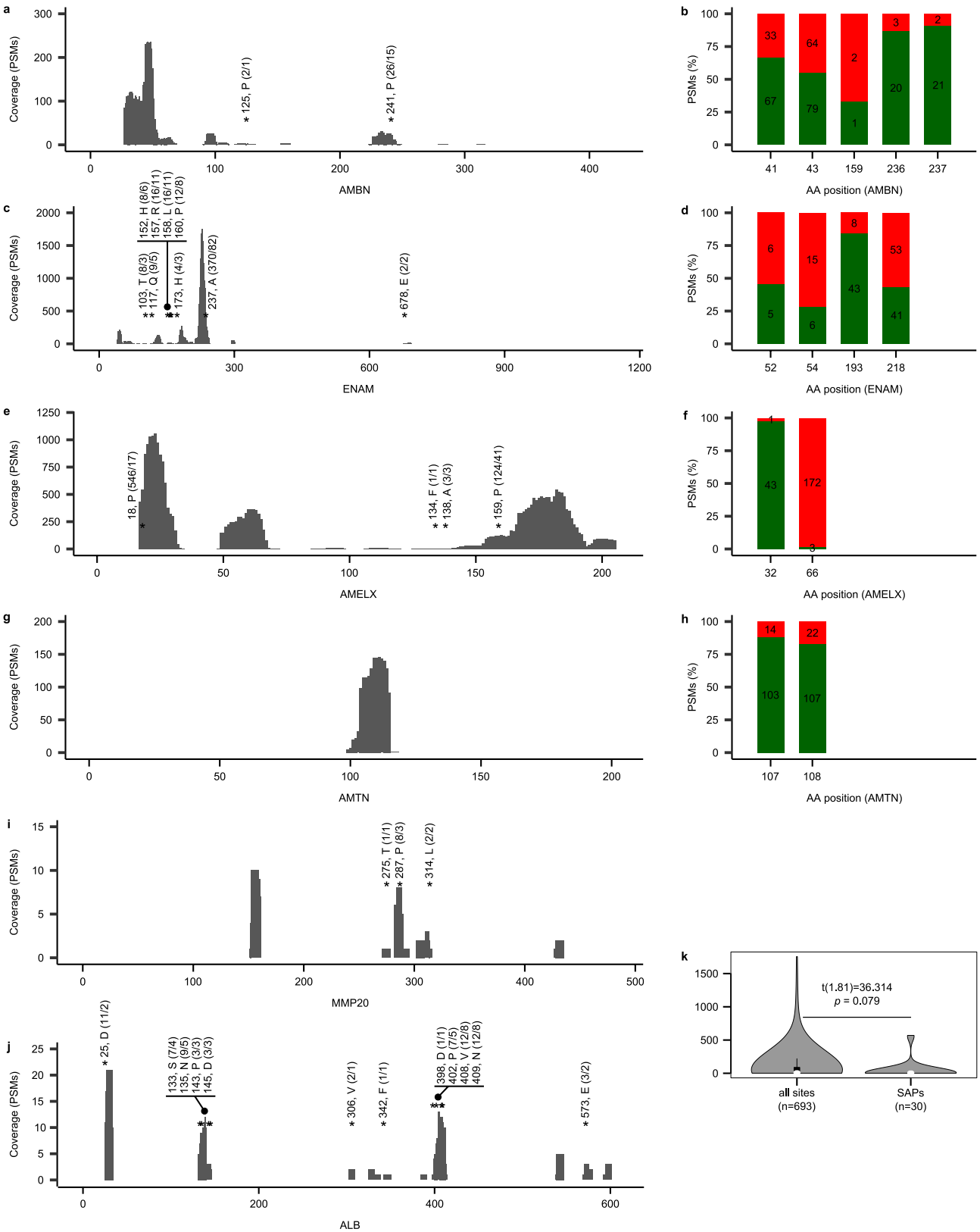
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1555-y.

Correspondence and requests for materials should be addressed to E.C., J.V.O. or E.W.

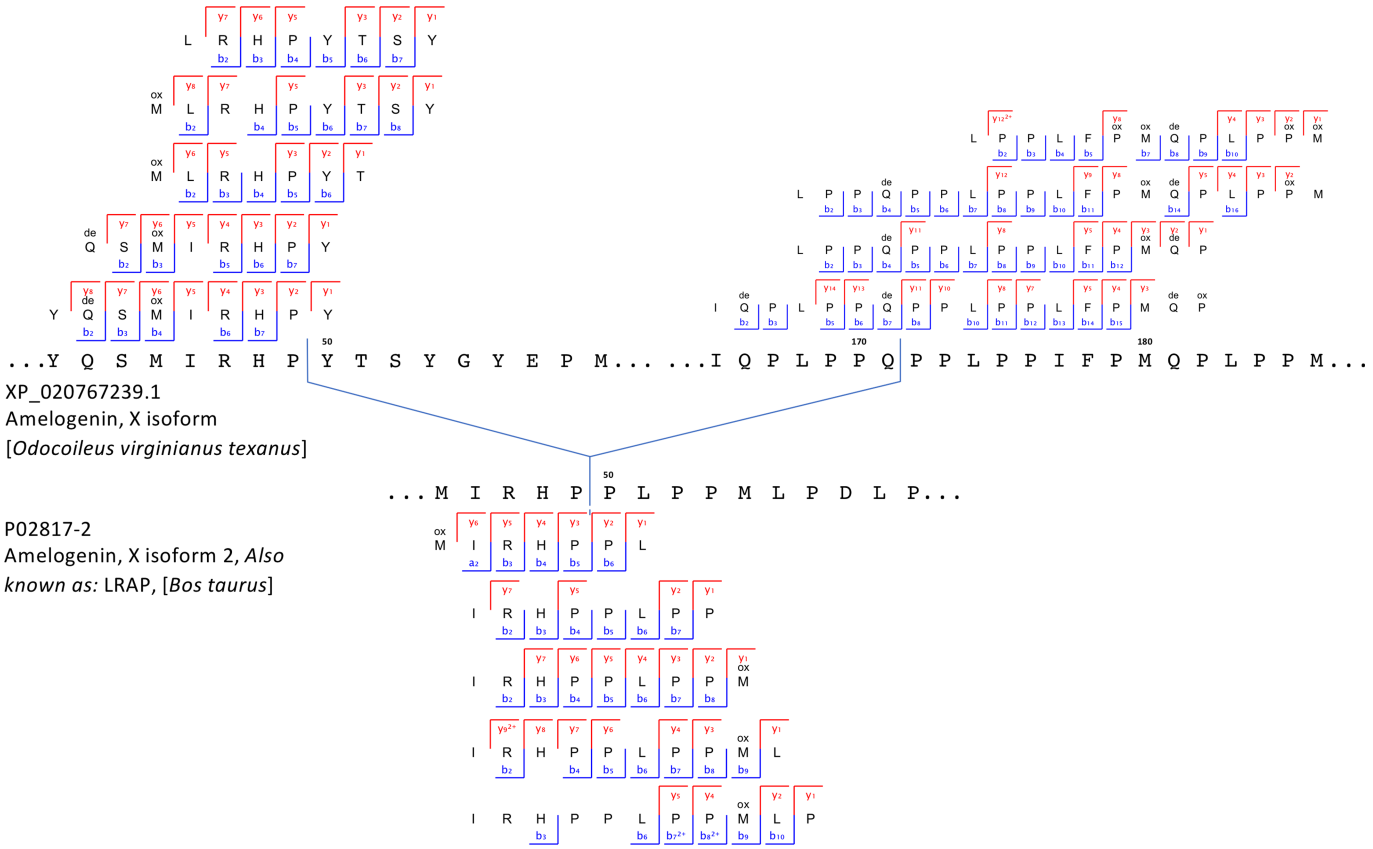
Peer review information *Nature* thanks Benedikt Kessler, Tina Warinner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Generalized stratigraphic profiles for Dmanisi, indicating origins of the specimens. a, Type section of the Dmanisi M5 excavation block. **b**, Stratigraphic profile of excavation area M6. M6 preserves a larger gully associated with the pipe-gully phase of stratigraphic–geomorphic development in stratum B1. The thickness of the stratum B1 gully fill extends to the basalt surface but includes 'rip-ups' of strata A1 and A2, showing that the deposits in stratum B1 post-date those of stratum A. **c**, Stratigraphic section of excavation area M17. Here, Stratum B1 was deposited after the erosion of stratum A deposits. The stratigraphic position of specimen Dm.5/157–16635 is highlighted with a red diamond. The Masavara basalt is about 50 cm below the base of the profile shown. **d**, Northern section of block 2. Following the collapse of a pipe and erosion to the basalt, the deeper part of this area was filled with local gully fill of strata B1x, B1y and B1z. Note the uniform burial of all stratum B1 deposits by strata B2, B3 and B4. The sampled specimens are indicated by the five-digit CGG numbers. Extended Data Table 1 provides both the CGG and GNM specimen numbers.



Extended Data Fig. 2 | Proteome-sequence coverage for specimen Dm.5/157-16635. a, **c**, **e**, **g**, **i**, **j**, Peptide–spectrum match (PSM) sequence coverage of the proteins AMBN (**a**), ENAM (**c**), AMELX (**e**), AMTN (**g**), MMP20 (**i**) and ALB (**j**). Annotations include 'amino acid position, amino acid called in that position (number of PSMs and peptides covering that position)' for the phylogenetically informative single-amino-acid polymorphisms within Rhinocerotidae. **b**, **d**, **f**, **h**, Frequency (per cent) of phosphorylated (green) and unphosphorylated (red) PSMs per amino acid position for AMBN (**b**), ENAM (**d**), AMELX (**f**) and AMTN (**h**). Numbers within the bars provide the PSM counts. **k**, Violin plot of distribution of PSM coverage for all covered sites ($n = 693$), and for sites of phylogenetic relevance (single-amino-acid polymorphisms, $n = 30$). The box plots define the range of the data, with whiskers extending to 1.5× interquartile range, boxes denoting the 25th and 75th percentiles and dots indicating the median. All panels are based only on MaxQuant search results. The Supplementary Data contains examples of MS/MS spectra, and fragment-ion series alignments for each of the marked single-amino-acid polymorphisms.



Extended Data Fig. 3 | Peptide and fragment-ion coverage of AMELX isoform 1 and isoform 2 from specimen Dm.M6/7.II.296–16856. Peptides specific to AMELX isoform 1 and isoform 2 appear in the top and bottom parts of the figure, respectively. No AMELX isoform 2 is currently reported in public databases for the Cervidae group. Accordingly, the AMELX-isoform-2-specific peptides were identified by MaxQuant spectral matching against bovine (*Bos taurus*) AMELX isoform 2 (UniProt accession number P02817-2). AMELX isoform 2 (also known as leucine-rich amelogenin peptide (LRAP)) is a naturally occurring isoform of AMELX from the translation product of an alternatively spliced transcript.