Early Pleistocene enamel proteome sequences from Dmanisi resolve *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 KSS 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[6], Bienvenido Martínez–Navarro[20,21,23], Shanlin Liu[2,24], Marcela Sandoval Velasco[2], Mikkel-Holger S. Sinding[2,25], Christian D. Kelstrup[5], Morten E. Allentoft[1], Ludovic Orlando[1,26], Kirsty Penkman[11], Beth Shapiro[16,27], Lorenzo Rook[4], Love Dalén[13], M. Thomas P. Gilbert[2,28], Jesper V. Olsen[5,*], David Lordkipanidze[6,29], Eske Willerslev[1,30,31,32,*]

[1] Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Denmark.
[2] Evolutionary Genomics Section, Globe Institute, University of Copenhagen, Denmark.
[3] Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, Germany.
[4] Dipartimento di Scienze della Terra, Università degli Studi di Firenze, Italy.
[5] Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark.
[6] Georgian National Museum, Tbilisi, Georgia.
[7] Life Sciences Research Unit, University of Luxembourg, Luxembourg.
[8] Department of Geography and Environment, University of North Texas, USA.
[9] Department of Anthropology, University of Minnesota, USA.
[10] Department of Genetics, Harvard Medical School, USA.
[11] Department of Chemistry, University of York, UK.
[12] Stafford Research LLC, Lafayette, 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, USA.
[17] Tromsø University Museum, UiT - The Arctic University of Norway, Tromsø, Norway.
[18] National Alliance of Shidlovskiy "Ice Age", Moscow, Russia.
[19] Department of Zoology, University of Venda, Republic of South Africa.
[20] Institut Català de Paleoecologia Humana i Evolució Social, Universitat Rovira i Virgili, Spain.
[21] Institució Catalana de Recerca i Estudis Avançats (ICREA).
[22] Senckenberg Research Station of Quaternary Palaeontology, Weimar, Germany.
[23] Departament d'Història i Geografia, Universitat Rovira i Virgili, Spain.
[24] BGI Shenzhen, Shenzen, China.
[25] Greenland Institute of Natural Resources, Nuuk, Greenland.

[26] Laboratoire d’Anthropobiologie Moléculaire et d’Imagerie de Synthèse, Université de Toulouse, Université Paul Sabatier, France.
[27] Howard Hughes Medical Institute, University of California Santa Cruz, USA.
[28] University Museum, Norwegian University of Science and Technology, Norway.
[29] Geology Department, Tbilisi State University, Georgia.
[30] Department of Zoology, University of Cambridge, UK.
[31] Wellcome Trust Sanger Institute, Hinxton, UK.
[32] Danish Institute for Advanced Study, University of Southern Denmark, Odense, Denmark.

*Corresponding authors: E. Cappellini (ecappellini@bio.ku.dk), J.V. Olsen (jesper.olsen@cpr.ku.dk), and E. Willerslev (ewillerslev@bio.ku.dk).

Ancient DNA (aDNA) sequencing has enabled reconstruction of speciation, migration, and admixture events for extinct taxa[1]. Outside the permafrost, however, irreversible aDNA post-mortem degradation[2] has so far limited aDNA recovery to the past ~0.5 million years (Ma)[3]. Contrarily, tandem mass spectrometry (MS) allowed sequencing ~1.5 million year (Ma) old collagen type I (COL1)[4] and suggested the presence of protein residues in Cretaceous fossil remains[5], though with limited phylogenetic use[6]. In the absence of molecular evidence, the speciation of several Early and Middle Pleistocene extinct species remain contentious. In this study, we address the phylogenetic relationships of the Eurasian Pleistocene Rhinocerotidae[7-9] using a ~1.77 Ma old dental enamel proteome of a *Stephanorhinus* specimen from the Dmanisi archaeological site in Georgia (South Caucasus)[10]. Molecular phylogenetic analyses place the Dmanisi *Stephanorhinus* as a sister group to the woolly (*Coelodonta antiquitatis*) and Merck's rhinoceros (*S. kirchbergensis*) clade. We show that *Coelodonta* evolved from an early *Stephanorhinus* lineage and that the latter includes at least two distinct evolutionary lines. As such, the genus *Stephanorhinus* is currently paraphyletic and its systematic revision is therefore needed. We demonstrate that Early Pleistocene dental enamel proteome sequencing overcomes the limits of ancient collagen- and aDNA-based phylogenetic inference. It also provides additional information about the sex and taxonomic assignment of the specimens analysed. Dental enamel, the hardest tissue in vertebrates[11], is highly abundant in the fossil record. Our findings reveal that palaeoproteomic investigation of this material can push biomolecular investigation further back into the Early Pleistocene.

Phylogenetic placement of extinct species increasingly relies on aDNA sequencing. Efforts to improve the molecular tools underlying aDNA recovery have enabled the reconstruction of ~0.4 Ma and ~0.7 Ma old DNA sequences from temperate deposits[3] and subpolar regions[12], respectively. However, no aDNA data have so far been generated from species that became extinct beyond this time range. In contrast, ancient proteins represent a more durable source of genetic information, reported to survive, in eggshell, up to 3.8 Ma[13]. Ancient protein sequences can carry taxonomic and phylogenetic information useful to trace the evolutionary relationships between extant and extinct species[14,15]. However, so far, the recovery of ancient mammal proteins from sites too old or too warm to be compatible with aDNA preservation is mostly limited to collagen type I (COL1). Being highly conserved[16], this protein is not an ideal phylogenetic marker. For example, regardless of endogeneity[17], collagen-based phylogenetic placement of Dinosauria in relation to extant Aves appears to be unstable[6]. This suggests the exclusive use of COL1 in deep-time phylogenetics is constraining. Here, we aimed at overcoming these limitations by testing whether dental enamel can better preserve a richer set of ancient protein residues.

Dated to ~1.77 Ma by a combination of ^{40}Ar/^{39}Ar dating, paleomagnetism and biozonation[18,19], the archaeological site of Dmanisi (Georgia, South Caucasus; Fig. 1a) represents a context currently considered outside the scope of aDNA recovery. This site has been excavated since 1983, resulting in the discovery, along with stone tools and contemporaneous fauna (Table S1), of almost one hundred hominin fossils, including five skulls representing the *georgicus* paleodeme within *Homo erectus*[10]. These are the earliest fossils of the genus *Homo* outside Africa.

The geology of the Dmanisi deposits favours the preservation of faunal materials (Supplementary Information: Extended Methods and Results), as the primary aeolian deposits provide rapid burial in fine-grained, calcareous sediments. We studied 12 bone and 14 enamel+dentine samples from 23 specimens of large mammals from multiple excavation units within stratum B1 (Fig. 1b, Extended Data Fig. 1, Extended Data Table 1, Table S3). This is an ashfall deposit that contains faunal remains in different geomorphic contexts. All of these are firmly dated between 1.85-1.76 Ma[19]. High-resolution tandem MS was used to confidently sequence ancient protein residues from the set of faunal remains, after digestion-based (protocols A and B), or digestion-free (protocol C), sample preparation (Methods and Supplementary Information). Ancient DNA analysis was unsuccessfully attempted on a subset of five bone and dentine specimens (Methods).

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

The small proteome[20,21] of mature dental enamel consists of structural enamel proteins, i.e. amelogenin (AMELX), enamelin (ENAM), amelotin (AMTN), and ameloblastin (AMBN), and enamel-specific proteases secreted during amelogenesis, i.e. matrix metalloproteinase-20 (MMP20) and kallikrein 4 (KLK4). The presence of non-specific proteins, such as serum albumin (ALB) and collagen type I, has also been previously reported in mature dental enamel[20] (Extended Data Table 2). The depth of coverage for these proteins varied considerably across their sequence, with some positions covered by over 1000 peptide spectrum matches (Extended Data Fig. 2). The high depth of coverage also allows to identify multiple isoforms of AMELX (Extended Data Fig. 3).

Multiple lines of evidence support the authenticity and the endogenous origin of the sequences recovered. Dental enamel proteins are extremely tissue-specific and confined to the dental enamel mineral matrix[20]. The amino acid composition of the intra-crystalline protein fraction, measured by amino acid racemisation analysis, indicates that the dental enamel behaves as a closed system, unaffected by amino acid and protein residues exchange with the burial environment (Extended Data Fig. 4). The measured rate of asparagine and glutamine deamidation, a spontaneous form of hydrolytic damage consistently observed in ancient samples[22], is particularly advanced. Deamidation in Dmanisi enamel is higher than in the control enamel sample, supporting the antiquity of the peptides recovered (Fig. 2a, Supplementary Information). Other forms of non-enzymatic modifications are also present. Tyrosine (Y) experienced mono- and di-oxidation while tryptophan (W) was extensively converted into multiple oxidation products (Fig. 2b, Supplementary Information). Oxidative degradation of histidine (H) and conversion of arginine (R) leading to ornithine accumulation were also observed (Supplementary Information). These modifications are absent, or much less frequent, in the control sample. Similarly, unlike in the control, the peptide length distribution in the Dmanisi dataset is dominated by shorter fragments, generated by advanced, diagenetically-induced, terminal hydrolysis[23] (Fig. 2c, d). Together all these independent lines of evidence clearly define the substantial biomolecular damage affecting the proteomes retrieved and independently support the authenticity of the sequences reconstructed. To demonstrate beyond reasonable doubt the correct peptide sequence assignments of our MS2 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 and Supplementary Information: Key MS2 Spectra).

We confidently detect phosphorylation (Fig. 3, Extended Data Figs. 2, 5), a stable and tightly *in vivo* regulated physiological post-translational modification (PTM) previously detected in dental enamel proteins[24,25]. Most of the phosphorylated sites we identified belong to the S-x-E/phS motif, recognised by the secreted kinases of the Fam20C family, which are involved in phosphorylation of extracellular proteins and regulation of biomineralization[26]. Spectra supporting the identification of serine phosphorylation were validated manually and by comparison with MS2 obtained from synthetic peptides (Supplementary Information), confirming the automated MaxQuant identifications. 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 in ancient bone[27]. Previous studies demonstrated that, when complexed with mineral matrix, ~3.8 Ma protein residues can be retrieved from sub-tropical environments[13]. Limited availability of free water in the enamel matrix further reduces spontaneous dephosphorylation via beta-elimination. Altogether, these observations demonstrate that the heavily modified dental enamel proteome retrieved from the ~1.77 Ma old Dmanisi faunal material is endogenous and almost complete.

Next, we used the palaeoproteomic sequence information to improve taxonomic assignment and achieve sex attribution for some of the Dmanisi faunal remains. Phylogenetic analysis of the five largest enamel+dentine proteomes, and of a moderately large bone proteome, allowed to confirm or improve the morphological identification of their specimens of origin (Extended Data Fig. 6; Figs. S10-15). In addition, confident identification of peptides specific for the isoform Y of amelogenin, coded on the non-recombinant portion of the Y chromosome, indicates that four tooth specimens, namely Dm.6/151.4.A4.12-16630 (*Pseudodama*), 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 the lower molar of a *Stephanorhinus* ex gr. *etruscus-hundsheimensis* (Dm.5/157-16635; Fig. 1c, Supplementary Information), returned the highest proteomic sequence coverage, encompassing a total of 875 amino acids, across 987 peptides (6 proteins; Extended Data Fig. 2; Supplementary Information). Following alignment of the enamel protein sequences retrieved from Dm.5/157-16635 against their homologues from all the extant rhinoceros species, plus the extinct woolly rhinoceros (†*Coelodonta antiquitatis*) and Merck's rhinoceros (†*Stephanorhinus kirchbergensis*), phylogenetic reconstructions place the Dmanisi specimen closer to the extinct woolly and Merck's 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 recovers the expected differentiation of the *Rhinoceros* genus from other genera considered, in agreement with previous cladistic[28] and genetic analyses[29] (Supplementary Information). This topology defines two-horned rhinoceroses as monophyletic and the one-horned condition as plesiomorphic, as previously proposed (Supplementary Information). We caution, however, that the higher-level relationships 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], due to the limited number of markers considered. A confident and stable reconstruction of the structure of the Rhinocerotidae family needs the strong support only high-resolution whole-genome sequencing can provide. Regardless, the highly supported placement of the Dmanisi rhinoceros in the (*Stephanorhinus*, Woolly, Sumatran) clade will remain unaffected, should 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 South-East Asian genus *Dicerorhinus* represented by the Sumatran rhinoceros species (*D. sumatrensis*)[32]. This hypothesis has been rejected and, based on morphological data, *Stephanorhinus* has been identified as a sister taxon of the woolly rhinoceros[33]. Furthermore, ancient DNA analysis 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 and Merck's rhinoceroses, these two species most likely derived from an early *Stephanorhinus* lineage expanding eastward from western Eurasia. Throughout the Plio-Pleistocene, *Coelodonta* adapted to continental and later to cold-climate habitats in central Asia. Its earliest representative, *C. thibetana,* displayed some clear *Stephanorhinus*-like anatomical features[33]. The presence in eastern Europe and Anatolia of the genus *Stephanorhinus*[35] is documented at least since the late Miocene, and the Dmanisi specimen most likely represents an Early Pleistocene descendent of the Western-Eurasian branch of this genus.

Ultimately, our phylogenetic reconstructions show that, as currently defined, the genus *Stephanorhinus* is paraphyletic, in line with previous morphological and palaeobiogeographical 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 enamel proteome sequencing can overcome the time limits of ancient DNA preservation and 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 pertaining to the Early and Middle Pleistocene evolutionary history of a large number of mammals, including hominins, at least in temperate climates.

REFERENCES

1 Cappellini, E. *et al.* Ancient Biomolecules and Evolutionary Inference. *Annual Review of Biochemistry* **87**, 1029-1060, doi:10.1146/annurev-biochem-062917-012002 (2018).

2 Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology* **5**, a012567, doi:10.1101/cshperspect.a012567 (2013).

3 Meyer, M. *et al.* Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531**, 504-507, doi:10.1038/nature17405 (2016).

4 Wadsworth, C. & Buckley, M. Proteome degradation in fossils: investigating the longevity of protein survival in ancient bone. *Rapid Communications in Mass Spectrometry* **28**, 605-615, doi:10.1002/rcm.6821 (2014).

5 Schweitzer, M. H. *et al.* Analyses of Soft Tissue from *Tyrannosaurus rex* Suggest the Presence of Protein. *Science* **316**, 277-280, doi:10.1126/science.1138709 (2007).

6 Schroeter, E. R. *et al.* Expansion for the Brachylophosaurus canadensis Collagen I Sequence and Additional Evidence of the Preservation of Cretaceous Protein. *Journal of Proteome Research* **16**, 920-932, doi:10.1021/acs.jproteome.6b00873 (2017).

7 Willerslev, E. *et al.* Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. *BMC Evolutionary Biology* **9**, 95, doi:10.1186/1471-2148-9-95 (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, doi:10.7717/peerj.3033 (2017).

9 Kirillova, I. *et al.* Discovery of the skull of Stephanorhinus kirchbergensis (Jäger, 1839) above the Arctic Circle. *Quaternary Research* **88**, 537-550, doi:10.1017/qua.2017.53 (2017).

10 Lordkipanidze, D. *et al.* A complete skull from Dmanisi, Georgia, and the evolutionary biology of early Homo. *Science* **342**, 326-331, doi:10.1126/science.1238484 (2013).

11 Eastoe, J. E. Organic Matrix of Tooth Enamel. *Nature* **187**, 411-412, doi:10.1038/187411b0 (1960).

12 Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74-78, doi:10.1038/nature12323 (2013).

13 Demarchi, B. *et al.* Protein sequences bound to mineral surfaces persist into deep time. *eLife* **5**, e17092, doi:10.7554/eLife.17092 (2016).

14 Welker, F. *et al.* Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature* **522**, 81-84, doi:10.1038/nature14249 (2015).

15 Chen, F. *et al.* A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature* **569**, 409-412, doi:10.1038/s41586-019-1139-x (2019).

16 Nei, M. *Molecular evolutionary genetics*. Vol. 75 (Columbia University 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. *Proceedings of the Royal Society: Biological sciences* **284**, 20170544, doi:10.1098/rspb.2017.0544 (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, doi:10.1126/science.288.5468.1019 (2000).

19 Ferring, R. *et al.* Earliest human occupations at Dmanisi (Georgian Caucasus) dated to 1.85-1.78 Ma. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10432-10436, doi:10.1073/pnas.1106638108 (2011).

20 Castiblanco, G. A. *et al.* Identification of proteins from human permanent erupted enamel. *European Journal of Oral Sciences* **123**, 390-395, doi:10.1111/eos.12214 (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, doi:10.1039/c6ra05120k (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 Communications in Mass Spectrometry* **26**, 2319-2327, doi:10.1002/rcm.6351 (2012).

23 Catak, S., Monard, G., Aviyente, V. & Ruiz-Lopez, M. F. Computational study on nonenzymatic peptide bond cleavage at asparagine and aspartic acid. *J Phys Chem A* **112**, 8752-8761, doi:10.1021/jp8015497 (2008).

24 Hunter, T. Why nature chose phosphate to modify proteins. *Philosophical Transactions of the Royal Society B* **367**, 2513-2516, doi:10.1098/rstb.2012.0013 (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, doi:10.1159/000091383 (2005).

26 Tagliabracci, V. S. *et al.* Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. *Science* **336**, 1150-1153, doi:10.1126/science.1217817 (2012).

27 Cleland, T. P. Solid Digestion of Demineralized Bone as a Method to Access Potentially Insoluble Proteins and Post-Translational Modifications. *Journal of Proteome Research* **17**, 536-542, doi:10.1021/acs.jproteome.7b00670 (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. *Zoological Journal of the Linnean Society* **160**, 139-194, doi:10.1111/j.1096-3642.2009.00597.x (2010).

29 Steiner, C. C. & Ryder, O. A. Molecular phylogeny and evolution of the Perissodactyla. *Zoological Journal of the Linnean Society* **163**, 1289-1303, doi:10.1111/j.1096-3642.2011.00752.x (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 research* **21**, 349-356, doi:10.1101/gr.114751.110 (2011).

31 Rieseberg, L. H. Evolution: replacing genes and traits through hybridization. *Current Biology* **19**, R119-R122, doi:10.1016/j.cub.2008.12.016 (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* **79**, 3-1183 (1980).

33 Deng, T. *et al.* Out of Tibet: pliocene woolly rhino suggests high-plateau origin of Ice Age megaherbivores. *Science* **333**, 1285-1288, doi:10.1126/science.1206594 (2011).

34 Orlando, L. *et al.* Ancient DNA analysis reveals woolly rhino evolutionary relationships. *Molecular Phylogenetics and Evolution* **28**, 485-499, doi:10.1016/S1055-7903(03)00023-X (2003).

35 Yuan, J. *et al.* Ancient DNA sequences from Coelodonta antiquitatis in China reveal its divergence and phylogeny. *Science China Earth Sciences* **57**, 388-396, doi:10.1007/s11430-013-4702-6 (2014).

MAIN TEXT FIGURE LEGENDS

Figure 1. Dmanisi location, stratigraphy, and *Stephanorhinus* specimen GNM Dm.5/157-16635. a, Geographic location of Dmanisi in the South Caucasus. The base map was generated using public domain data from www.naturalearthdata.com. **b,** Generalised stratigraphic profile indicating origin and age of the analysed specimens. **c,** Isolated left lower molar (m1 or m2) of *Stephanorhinus* ex gr. *etruscus-hundsheimensis*, from Dmanisi (labial view). Scale bar: 1 cm.

Figure 2. Enamel proteome degradation. a, Deamidation of asparagine (N) and glutamine (Q). Violin plots based on 1000 bootstrap replicates. The boxplots define the range of the data, with whiskers extending to 1.5 the interquartile range, 25th and 75th percentiles (boxes), and medians (dots). Tissue source (B = Bone, D = Dentine, E = Enamel) and the number of peptides used for the calculation are shown at the bottom. **b,** Extent of tryptophan (W) oxidation leading to several diagenetic products, measured as relative spectral counts. **c,** Alignment of peptides (positions 124-137, Enamelin) retrieved by digestion-free acid demineralisation from Pleistocene *Stephanorhinus* ex gr. *etruscus-hundsheimensis* specimen (GNM Dm.5/157-16635). **d,** Barplot of peptide length distribution of specimen Dm.5/157-16635 and Medieval (CTRL) undigested ovicaprine dental enamel proteomes.

Figure 3. Sequence motif analysis of ancient enamel proteome phosphorylation. Indicated is the overrepresentation of specific amino acids within six positions N- and C-terminal of the phosphorylated amino acids (position 0). See Extended Data Figure 5 for MS2 examples of both S-x-E and S-x-phS phosphorylated motifs.

Figure 4. Phylogenetic relationships between the comparative enamel proteome dataset and specimen Dm.5/157-16635 (*Stephanorhinus* ex gr. *etruscus-hundsheimensis*). 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. Additionally, for bipartitions where the Bayesian and the Maximum-likelihood inference support are different, we show (right) the support obtained in the latter. Scale indicates estimated branch lengths.

METHODS

Dmanisi & sample selection

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

After the sample preparation and data acquisition for all the Dmanisi specimens was concluded, we applied the whole experimental procedure to a medieval ovicaprine (sheep/goat) dental enamel+dentine specimen that was used as control. For this sample, we used extraction protocol "C", and generated tandem MS 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 31st May 2017 (Supplementary Information). The ovicaprine specimen was found at the "Hotel Skandinavia" site in the city of Århus, Denmark and stored at the Natural History Museum of Denmark, Copenhagen.

Biomolecular preservation

We assessed the potential of ancient protein preservation prior to proteomic analysis by measuring the extent of amino acid racemisation in a subset of samples (6/23)[36]. Enamel chips, with all dentine removed, were powdered, and two subsamples per specimen were subject to analysis of their free (FAA) and total hydrolysable (THAA) amino acid fractions. Samples were analysed in duplicate by RP-HPLC, with standards and blanks run alongside each one of them (Supplementary Information). The D/L values of aspartic acid/asparagine, glutamic acid/glutamine, phenylalanine and alanine (D/L Asx, Glx, Phe, Ala) were assessed (Extended Data Fig. 4) to provide an overall estimate of intra-crystalline protein decomposition (IcPD).

PROTEOMICS

All the sample preparation procedures for palaeoproteomic analysis were conducted in laboratories dedicated to the analysis of ancient DNA and ancient proteins in clean rooms fitted with filtered ventilation and positive pressure, in line 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. Ancient protein residues were extracted from approximately 180-220 mg of mineralised material, unless otherwise specified, using three different extraction protocols, hereafter referred to as "A", "B" and "C" (Supplementary Information):

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

Extraction protocol B - GuHCl solution and digestion. Bone or enamel+dentine powder was demineralised in 1 mL 0.5 M EDTA pH 8.0. After removal of the supernatant, all demineralised pellets were re-suspended 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 H_2O[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 hours 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 demineralisation. Dental enamel powder, with possible trace amounts of dentine, was demineralised in 1.2 M HCl at room temperature, after which the solubilised protein residues were directly cleaned and concentrated on Stage-Tips, as described above. The sample prepared on Stage-Tip "#1217" was processed with 10% TFA instead of 1.2 M HCl. All the other parameters and procedures were identical to those used for all the other samples extracted with protocol "C".

Tandem mass spectrometry

Different sets of samples (Supplementary Information §5.1, 5.2) were analysed by nanoflow liquid chromatography coupled to tandem mass spectrometry (nanoLC-MS/MS) on an EASY-nLC™ 1000 or 1200 system connected to a Q-Exactive, a Q-Exactive Plus, or to a Q-Exactive HF (Thermo Scientific, Bremen, Germany) 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 in order to prevent carryover contamination between the samples. These consisted, first, of a MS/MS run ("MS/MS blank" run) with an injection exclusively of the buffer used to re-suspend the samples (0.1% TFA, 5% ACN), followed by a second MS/MS run ("MS/MS wash" run) with no injection.

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 §5.3). Raw files were initially searched against a target/reverse database of collagen and enamel proteins retrieved from the UniProt and NCBI Reference Sequence Database (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 whose dataset resulted in a narrower, though 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 (*Dicerorhinus sumatrensis*), Javan (*Rhinoceros sondaicus*), Indian (*Rhinoceros unicornis*), woolly (*Coelodonta antiquitatis*), Mercks (*Stephanorhinus kirchbergensis*), and Black rhinoceros (*Diceros bicornis*), were also added to the protein database. All the protein sequences from these species were reconstructed from draft genomes for each species (Dalen and Gilbert, unpublished data, Supplementary Information).

For each MaxQuant and PEAKS search, enzymatic digestion was set to "unspecific" and the following variable modifications were included: oxidation (M), deamidation (NQ), N-term Pyro-Glu (Q), N-term Pyro-Glu (E), hydroxylation (P), phosphorylation (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 (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 (C) was not included as a modification, neither fixed nor variable.

The datasets re-analysed with MQ2 search, were also processed with the PEAKS software using the entire workflow (PEAKS *de novo* to PEAKS SPIDER) in order to detect hitherto unreported single amino acid polymorphisms (SAPs). 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.

Ancient protein sequence reconstruction

The peptide sequences confidently identified by the MQ1, MQ2, 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 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”. Novel SAPs discovered through PEAKS were only accepted if these were further validated by repeating the MaxQuant search (MQ3). All leucines were converted into isoleucines, 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 where both Q and E or both N and D occurred on the same position, and where we also had ancient sequences matching. For sample 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 their phylogenetic relevance. The output of the MQ2 and 3 searches was used to extend the coverage of the ancient protein sequences initially identified in the MQ1 iteration. For specimen DM.5/157-16335, all 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 Information (“Key MS-MS Spectra” file). All the reported MS/MS spectra are annotated using the advanced annotation mode of MaxQuant. Selected spectra matching to peptides covering 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 Information, “Key MS-MS Spectra” file). We utilized MS^{2}PIP fragment ion spectral intensity prediction[49] (version: v20190107; model: HCD) to demonstrate that the experimentally observed fragment ion intensities are highly correlated with the theoretical ones (Fig. S3).

Finally, we generated synthetic peptides for 19 selected peptides covering Rhinocerotidae SAPs 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, by 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* ALC score was set to a threshold of 15 % and the peptide hit threshold to 30. The results were filtered by an FDR of 5 %, *de novo* ALC score of 50 %, and a protein hit threshold of ≥ 20. The MaxQuant dependent peptides search was carried out with the same search settings as described above and with a dependent peptide FDR of 1 % and a mass bin size of 0.0065 Da.

PHOSPHORYLATION. Class I phosphorylation sites were selected with localisation probabilities of ≥0.98 in the Phosph(ST)Sites MaxQuant output file. Sequence windows of ±6 aa from all identified sites were compared against a background file containing all non-phosphorylated 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 consisting of publicly available protein sequences from representative ungulate species belonging to the following families: Equidae, Rhinocerotidae, Suidae and Bovidae (Supplementary Information §7 and §8). As Cervidae and carnivores are absent from protein sequence databases to a various extent, 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, Dm.8/152.3.B1.2-16641) and the largest bone proteome (Dm.bXI.North.B1a.collection-16658) we recovered (see Extended Data Table 2).

We extended this dataset with the protein sequences from extinct and extant rhinoceros species including: the woolly rhinoceros († *Coelodonta antiquitatis*), the Merck's rhinoceros († *Stephanorhinus kirchbergensis*), the Sumatran rhinoceros (*Dicerorhinus sumatrensis*), the Javan rhinoceros (*Rhinoceros sondaicus*), the Indian rhinoceros (*Rhinoceros unicornis*), and the Black rhinoceros (*Diceros 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 and Merck's rhinoceroses we excluded the first and last five nucleotides of each DNA fragment in order to minimize the effect of post-mortem ancient DNA damage[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 in order to recover translated and spliced protein sequences. 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 the isoleucine residues were substituted with leucine ones to account for indistinguishable isobaric amino acids at the positions where the ancient protein carried one of such amino acids. Based on these alignments, we inferred the phylogenetic relationship between the ancient samples and the species included in the reference dataset by 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 based on 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. While 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 MCMC algorithm was run with 4 chains for 5,000,000 cycles. Sampling was conducted every 500 cycles and the first 25% were discarded as burn-in. Convergence was assessed using Tracer v. 1.6.0, which estimated an ESS greater than 5,500 for each individual, indicating reasonable convergence for all runs.

ANCIENT DNA ANALYSIS

The samples were processed using strict aDNA guidelines in a clean lab facility at the Natural History Museum of Denmark, University of Copenhagen. DNA extraction was attempted on five of the ancient animal samples (Supplementary Information §9, §13). Powdered samples (120-140 mg) were extracted using a silica-in-solution method[12,60]. To prepare the samples for NGS 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 6-nucleotide index[61]. Two extracts (MA399 and MA2481, from specimens 16859 and 16635 respectively) yielded detectable DNA concentrations (Table S9). 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, MA399_L2), using 100bp single read chemistry, and on an Illumina 2500 platform (MA2481_L1), using 81bp 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] v0.5.10 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/].

SAMPLE Dm.5/157-16635 MORPHOLOGICAL MEASUREMENTS

We followed the methodology introduced by Guérin[32]. The maximal length of the tooth is 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 §3).

ACKNOWLEDGMENTS

EC and FW 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). EW is supported by the Lundbeck Foundation, the Danish National Research Foundation, the Carlsberg Foundation, KU2016, and the Wellcome Trust. EC, CK, JVO, PR and DS are supported by the European Commission through the MSC European Training Network “TEMPERA” (grant number 722606). MM and RRJ-C are supported by the University of Copenhagen KU2016 (UCPH Excellence Programme) grant. MM 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). MRD is supported by a PhD DTA studentship from NERC and the Natural History Museum (NE/K500987/1 & NE/L501761/1). KP is supported by the Leverhulme Trust (PLP -2012-116). LR and LP are supported by the Italian Ministry for Foreign Affairs (MAECI, DGSP-VI). LP was also supported by the EU-SYNTHESYS project (AT-TAF-2550, DE-TAF-3049, GB-TAF-2825, HU-TAF-3593, ES-TAF-2997) funded by the European Commission. LD is supported by the Swedish Research Council (grant number 2017-04647) and FORMAS (grant nr 2015-676). MTPG is supported by ERC Consolidator Grant “EXTINCTION GENOMICS” (grant number 681396). LO is supported by the ERC Consolidator Grant “PEGASUS” (grant agreement No 681605). BS, JK and PDH are supported by the Gordon and Betty Moore foundation. BM-N is supported by the Spanish Ministry of Sciences (grant number CGL2016-80975-P). The aDNA analysis was carried out using the facilities of the University of Luxembourg, the Swedish Museum of Natural History and UC Santa Cruz. The authors would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure (Sweden), and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure. Research at Dmanisi is supported by the John Templeton Foundation, the Shota Rustaveli Science Foundation, and the Alexander von Humboldt Fellowship research award. The authors would also like to thank B. Triozzi and K. Murphy Gregersen for technical support.

AUTHOR CONTRIBUTIONS

E.C., D.Lo., 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.Bu., M.T., R.F., E.P., T.S., Y.L.C., A.Gö., S.KSS.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.Lo., provided ancient samples or modern reference material. E.C., F.W., L.P., J.R.M., D.Ly, V.J.M.M., D.S., C.D.K., A.Gi., 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.

AUTHOR INFORMATION

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

The Authors declare no financial competing interests.

Correspondence and requests for material should be addressed to E.C. (ecappellini@bio.ku.dk), J.V.O. (jesper.olsen@cpr.ku.dk) or E.W. (ewillerslev@bio.ku.dk).

METHODS REFERENCES

10 Lordkipanidze, D. *et al.* A complete skull from Dmanisi, Georgia, and the evolutionary biology of early Homo. *Science* **342**, 326-331, doi:10.1126/science.1238484 (2013).

12 Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74-78, doi:10.1038/nature12323 (2013).

18 Gabunia, L. *et al.* Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: taxonomy, geological setting, and age. *Science* **288**, 1019-1025, doi:10.1126/science.288.5468.1019 (2000).

19 Ferring, R. *et al.* Earliest human occupations at Dmanisi (Georgian Caucasus) dated to 1.85-1.78 Ma. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10432-10436, doi:10.1073/pnas.1106638108 (2011).

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* **79**, 3-1183 (1980).

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. *Quaternary Geochronology* **3**, 2-25, doi:10.1016/j.quageo.2007.07.001 (2008).

37 Hendy, J. *et al.* A guide to ancient protein studies. *Nature Ecology & Evolution* **2**, 791-799, doi:10.1038/s41559-018-0510-x (2018).

38 Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nature Methods* **6**, 359-362, doi:10.1038/nmeth.1322 (2009).

39 Cappellini, E. *et al.* Resolution of the type material of the Asian elephant, Elephas maximus Linnaeus, 1758 (Proboscidea, Elephantidae. *Zoological Journal of the Linnean Society* **170**, 222-232, doi:10.1111/zoj.12084 (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. *Nature Methods* **11**, 319-324, doi:10.1038/nmeth.2834 (2014).

41 Mackie, M. *et al.* Palaeoproteomic Profiling of Conservation Layers on a 14th Century Italian Wall Painting. *Angewandte Chemie (International ed.)* **57**, 7369-7374, doi:10.1002/anie.201713020 (2018).

42 Cappellini, E. *et al.* Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins. *Journal of Proteome Research* **11**, 917-926, doi:10.1021/pr200721u (2012).

43 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).

44 Zhang, J. *et al.* PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular and Cellular Proteomics* **11**, M111.010587, doi:10.1074/mcp.M111.010587 (2012).

45 TheUniProtConsortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

46 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733-D745, doi:10.1093/nar/gkv1189 (2016).

47 Welker, F. *et al.* Palaeoproteomic evidence identifies archaic hominins associated with the Châtelperronian at the Grotte du Renne. *Proceedings of the National Academy of Sciences* **113**, 11162-11167, doi:10.1073/pnas.1605834113 (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, doi:10.1093/bioinformatics/bts199 (2012).

49 Gabriels, R., Martens, L. & Degroeve, S. Updated MS2PIP web server delivers fast and accurate MS2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *bioRxiv*, 544965, doi:10.1101/544965 (2019).

50 Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **11**, 2301-2319, doi:10.1038/nprot.2016.136 (2016).

51 Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved visualization of protein consensus sequences by iceLogo. *Nature Methods* **6**, 786-787, doi:10.1038/nmeth1109-786 (2009).

52 Korneliussen, T., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356-356, doi:10.1186/s12859-014-0356-4 (2014).

53 Briggs, A. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research* **38**, e87, doi:10.1093/nar/gkp1163 (2010).

54 Altschul, S. F. *et al.* Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402 (1997).

55 SeaUrchinGenomeSequencingConsortium. 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, doi:10.1093/bioinformatics/bts578 (2012).

57 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593, doi:10.1093/bioinformatics/btq706 (2011).

58 Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).

59 Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology* **61**, 539-542, doi:10.1093/sysbio/sys029 (2012).

60 Rohland, N. & Hofreiter, M. Comparison and optimization of ancient DNA extraction. *BioTechniques* **42**, 343-352, doi:10.2144/000112383 (2007).

61 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, doi:10.1101/pdb.prot5448 (2010).

62 Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols* **9**, 1056-1082, doi:10.1038/nprot.2014.063 (2014).

63 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows– Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

64 Dickinson, M. L., A.; Penkman, K. A new method for enamel amino acid racemization dating: a closed system approach. *Quaternary Geochronology* **50**, 29-46, doi:10.1016/j.quageo.2018.11.005 (2019).

DATA AVAILABILITY

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

SUPPLEMENTARY INFORMATION

Supplementary information is available in the online version of the paper.

EXTENDED DATA LEGENDS

Extended Data Table 1. Genome and proteome survival in 23 Dmanisi fossil fauna specimens. For each specimen, the Centre for GeoGenetics (CGG) reference number and the Georgian National Museum (GNM) specimen field number are reported. *or the narrowest possible taxonomic identification achievable using comparative anatomy methods. †Only collagens survive. B = Bone, D = Dentine, E = Enamel. Extractions of enamel might include some residual dentine. Accordingly, both tissues are either listed separately (○D, ●E, in case of no collagen preservation), or together (●E+D, in case of collagen preservation). Open circles (○) indicate no molecular preservation; (●) closed circles indicate molecular preservation.

Extended Data Table 2. Proteome composition and coverage. Aggregated data from different extraction methods and/or tissues from the same specimen. In those cells reporting two values separated by the "|" symbol, the first value refers to MaxQuant (MQ) searches performed selecting unspecific digestion, while the second value refers to MQ searches performed selecting trypsin digestion. For those cells including one value only, it refers to MQ searches performed selecting unspecific digestion. Final amino acid coverage, incorporating both MQ and PEAKS searches, is reported in the last column. *supporting all peptides. See Extended Data Table 1 for tissue sources per specimen and both CGG and GNM specimen numbers.

Extended Data Figure 1. Generalized stratigraphic profiles for Dmanisi, indicating specimen origins. 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 Stratum B1 gully fill extends to the basalt surface, but includes "rip-ups" of Strata A1 and A2, showing that B1 deposits post-date Stratum A. **c,** Stratigraphic section of excavation area M17. Here, Stratum B1 was deposited after erosion of Stratum A deposits. The stratigraphic position of the *Stephanorhinus* sample Dm.5/157-16635 is highlighted with a red diamond. The Masavara basalt is ca. 50 cm below the base of the shown profile. **d,** Northern section of Block 2. Following collapse of a pipe and erosion to the basalt, the deeper part of this area was filled with local gully fill of Stratum B1/x/y/z. Note the uniform burial of all Stratum B1 deposits by Strata B2-B4. Sampled specimens are indicated by CGG five-digit numbers. See Extended Data Table 1 for both CGG and GNM specimen numbers.

Extended Data Fig. 2. Proteomic sequence coverage for specimen Dm.5/157-16635 (*Stephanorhinus*). a, c, e, g, i, j, PSM sequence coverage of proteins AMBN, ENAM, AMELX, AMTN, MMP20 and ALB, respectively. Annotations include: "amino acid position, amino acid called in that position (number of PSMs/peptides covering that position)" for the phylogenetically informative SAPs within Rhinocerotidae. **b, d, f, h,** Frequency (%) of phosphorylated (green) and non-phosphorylated (red) PSMs per amino acid position for AMBN, ENAM, AMELX and AMTN, respectively. Numbers within the bars provide the PSM counts. **k**, Violinplot of PSM coverage distribution for all covered sites (n=693) and those of phylogenetic relevance (SAPs, n=30). The boxplots define the range of the data, with whiskers extending to 1.5 the interquartile range, 25th and 75th percentiles (boxes), and medians (dots). All panels based on MQ results only. Supplementary File "Key MS-MS Spectra" contains spectral examples and fragment ion series alignments for each of the marked SAPs.

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

Extended Data Figure 4. Amino Acid Racemisation. Extent of intra-crystalline racemization in enamel for the free amino acid (FAA, x-axis) fraction and the total hydrolysable amino acids (THAA, y-axis) fraction for four amino acids (Asx, Glx, Ala and Phe). Note differences in axis scale. Intra-crystalline data from Proboscidea enamel from a range of UK sites[64] has been shown for comparison (black crosses). Both taxa from Dmanisi and the UK exhibit a similar relationship between FAA and THAA racemization and R^2 values have been calculated based on a polynomial relationship (order = 2, all >0.93).

Extended Data Figure 5. Ancient enamel proteome phosphorylation. Annotated spectra including phosphorylated serine (phS). **a**, Phosphorylation in the S-x-E motif (AMEL). **b**, Phosphorylation in the S-x-phS motif (AMBN). Phosphorylation was independently observed in all three separate analyses of Dm.5/157-16635, including multiple spectra and peptides (see Extended Data Fig. 2).

Extended Data Figure 6. Phylogenetic relationships between the comparative reference dataset and specimen Dm.bXI-16857. Consensus tree from Bayesian inference. The posterior probability of each bipartition is shown as a percentage to the left of each node.

Extended Data Figure 7. Amelogenin Y-specific matches. a) Specimen Dm.6/151.4.A4.12-16630 (*Pseudodama*). **b)** Specimen Dm.69/64.3.B1.53-16631 (Cervidae). **c)** Specimen Dm.8/154.4.A4.22-16639 (Bovidae). **d)** Specimen Dm.M6/7.II.296-16856 (Cervidae). Note the presence of deamidated glutamine (deQ) and asparagine (deN), oxidated methionine (oxM), and phosphorylated serine (phS).

Extended Data Figure 8. Effect of the missingness in the tree topology. a, Maximum-likelihood phylogeny obtained using PhyML and the protein alignment excluding the ancient Dmanisi rhinoceros Dm.5/157-16635. **b,** Topologies obtained from 100 random replicates of the Woolly rhinoceros (*Coelodonta antiquitatis*). In each replicate the amount of missing sites was similar to the one observed in the Dm.5/157-16635 specimen (72.4% missingness). The percentage shown for each topology indicates the number of replicates in which that particular topology was recovered. **c,** Similar to **b**, but for the Javan rhinoceros (*Rhinoceros sondaicus*). **d,** Similar to **b**, but for the black rhinoceros (*Diceros bicornis*).