

This is a repository copy of *Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/151936/>

Version: Accepted Version

Article:

Cappellini, Enrico, Welker, Frido, Pandolfi, Luca et al. (40 more authors) (2019) Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny. *Nature*. pp. 103-107. ISSN 0028-0836

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

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

1 Early Pleistocene enamel proteome sequences from Dmanisi 2 resolve *Stephanorhinus* phylogeny

3
4 Enrico Cappellini^{1,2,*}, Frido Welker^{2,3}, Luca Pandolfi⁴, Jazmín Ramos-Madrugal², Diana
5 Samodova⁵, Patrick L. Rüter⁵, Anna K. Fotakis², David Lyon⁵, J. Víctor Moreno-Mayar¹, Maia
6 Bukhsianidze⁶, Rosa Rakownikow Jersie-Christensen⁵, Meaghan Mackie^{2,5}, Aurélien
7 Ginolhac⁷, Reid Ferring⁸, Martha Tappen⁹, Eleftheria Palkopoulou¹⁰, Marc R. Dickinson¹¹,
8 Thomas W. Stafford Jr.¹², Yvonne L. Chan¹³, Anders Götherström¹⁴, Senthilvel KSS Nathan¹⁵,
9 Peter D. Heintzman^{16,17}, Joshua D. Kapp¹⁶, Irina Kirillova¹⁸, Yoshan Moodley¹⁹, Jordi
10 Agusti^{20,21}, Ralf-Dietrich Kahlke²², Gocha Kiladze⁶, Bienvenido Martínez–Navarro^{20,21,23},
11 Shanlin Liu^{2,24}, Marcela Sandoval Velasco², Mikkel-Holger S. Sinding^{2,25}, Christian D.
12 Kelstrup⁵, Morten E. Allentoft¹, Ludovic Orlando^{1,26}, Kirsty Penkman¹¹, Beth Shapiro^{16,27},
13 Lorenzo Rook⁴, Love Dalén¹³, M. Thomas P. Gilbert^{2,28}, Jesper V. Olsen^{5,*}, David
14 Lordkipanidze^{6,29}, Eske Willerslev^{1,30,31,32,*}

15
16 ¹ Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen,
17 Denmark.

18 ² Evolutionary Genomics Section, Globe Institute, University of Copenhagen, Denmark.

19 ³ Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology,
20 Germany.

21 ⁴ Dipartimento di Scienze della Terra, Università degli Studi di Firenze, Italy.

22 ⁵ Novo Nordisk Foundation Center for Protein Research, University of Copenhagen,
23 Denmark.

24 ⁶ Georgian National Museum, Tbilisi, Georgia.

25 ⁷ Life Sciences Research Unit, University of Luxembourg, Luxembourg.

26 ⁸ Department of Geography and Environment, University of North Texas, USA.

27 ⁹ Department of Anthropology, University of Minnesota, USA.

28 ¹⁰ Department of Genetics, Harvard Medical School, USA.

29 ¹¹ Department of Chemistry, University of York, UK.

30 ¹² Stafford Research LLC, Lafayette, USA.

31 ¹³ Department of Bioinformatics and Genetics, Swedish Museum of Natural History,
32 Stockholm, Sweden.

33 ¹⁴ Department of Archaeology and Classical Studies, Stockholm University, Stockholm,
34 Sweden.

35 ¹⁵ Sabah Wildlife Department, Kota Kinabalu, Malaysia.

36 ¹⁶ Department of Ecology and Evolutionary Biology, University of California Santa Cruz, USA.

37 ¹⁷ Tromsø University Museum, UiT - The Arctic University of Norway, Tromsø, Norway.

38 ¹⁸ National Alliance of Shidlovskiy "Ice Age", Moscow, Russia.

39 ¹⁹ Department of Zoology, University of Venda, Republic of South Africa.

40 ²⁰ Institut Català de Paleoecologia Humana i Evolució Social, Universitat Rovira i Virgili,
41 Spain.

42 ²¹ Institució Catalana de Recerca i Estudis Avançats (ICREA).

43 ²² Senckenberg Research Station of Quaternary Palaeontology, Weimar, Germany.

44 ²³ Departament d'Història i Geografia, Universitat Rovira i Virgili, Spain.

45 ²⁴ BGI Shenzhen, Shenzhen, China.

46 ²⁵ Greenland Institute of Natural Resources, Nuuk, Greenland.

47 ²⁶ Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, Université de
48 Toulouse, Université Paul Sabatier, France.
49 ²⁷ Howard Hughes Medical Institute, University of California Santa Cruz, USA.
50 ²⁸ University Museum, Norwegian University of Science and Technology, Norway.
51 ²⁹ Geology Department, Tbilisi State University, Georgia.
52 ³⁰ Department of Zoology, University of Cambridge, UK.
53 ³¹ Wellcome Trust Sanger Institute, Hinxton, UK.
54 ³² Danish Institute for Advanced Study, University of Southern Denmark, Odense, Denmark.
55
56 *Corresponding authors: E. Cappellini (ecappellini@bio.ku.dk), J.V. Olsen
57 (jesper.olsen@cpr.ku.dk), and E. Willerslev (ewillerslev@bio.ku.dk).

58 Ancient DNA (aDNA) sequencing has enabled reconstruction of speciation, migration, and
59 admixture events for extinct taxa¹. Outside the permafrost, however, irreversible aDNA
60 post-mortem degradation² has so far limited aDNA recovery to the past ~0.5 million years
61 (Ma)³. Contrarily, tandem mass spectrometry (MS) allowed sequencing ~1.5 million year
62 (Ma) old collagen type I (COL1)⁴ and suggested the presence of protein residues in
63 Cretaceous fossil remains⁵, though with limited phylogenetic use⁶. In the absence of
64 molecular evidence, the speciation of several Early and Middle Pleistocene extinct species
65 remain contentious. In this study, we address the phylogenetic relationships of the Eurasian
66 Pleistocene Rhinocerotidae⁷⁻⁹ using a ~1.77 Ma old dental enamel proteome of a
67 *Stephanorhinus* specimen from the Dmanisi archaeological site in Georgia (South
68 Caucasus)¹⁰. Molecular phylogenetic analyses place the Dmanisi *Stephanorhinus* as a sister
69 group to the woolly (*Coelodonta antiquitatis*) and Merck's rhinoceros (*S. kirchbergensis*)
70 clade. We show that *Coelodonta* evolved from an early *Stephanorhinus* lineage and that the
71 latter includes at least two distinct evolutionary lines. As such, the genus *Stephanorhinus* is
72 currently paraphyletic and its systematic revision is therefore needed. We demonstrate that
73 Early Pleistocene dental enamel proteome sequencing overcomes the limits of ancient
74 collagen- and aDNA-based phylogenetic inference. It also provides additional information
75 about the sex and taxonomic assignment of the specimens analysed. Dental enamel, the
76 hardest tissue in vertebrates¹¹, is highly abundant in the fossil record. Our findings reveal
77 that palaeoproteomic investigation of this material can push biomolecular investigation
78 further back into the Early Pleistocene.

79

80 Phylogenetic placement of extinct species increasingly relies on aDNA sequencing. Efforts to
81 improve the molecular tools underlying aDNA recovery have enabled the reconstruction of
82 ~0.4 Ma and ~0.7 Ma old DNA sequences from temperate deposits³ and subpolar regions¹²,
83 respectively. However, no aDNA data have so far been generated from species that became
84 extinct beyond this time range. In contrast, ancient proteins represent a more durable
85 source of genetic information, reported to survive, in eggshell, up to 3.8 Ma¹³. Ancient
86 protein sequences can carry taxonomic and phylogenetic information useful to trace the
87 evolutionary relationships between extant and extinct species^{14,15}. However, so far, the
88 recovery of ancient mammal proteins from sites too old or too warm to be compatible with
89 aDNA preservation is mostly limited to collagen type I (COL1). Being highly conserved¹⁶, this
90 protein is not an ideal phylogenetic marker. For example, regardless of endogeneity¹⁷,
91 collagen-based phylogenetic placement of Dinosauria in relation to extant Aves appears to
92 be unstable⁶. This suggests the exclusive use of COL1 in deep-time phylogenetics is
93 constraining. Here, we aimed at overcoming these limitations by testing whether dental
94 enamel can better preserve a richer set of ancient protein residues.

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

102 The geology of the Dmanisi deposits favours the preservation of faunal materials
103 (Supplementary Information: Extended Methods and Results), as the primary aeolian

104 deposits provide rapid burial in fine-grained, calcareous sediments. We studied 12 bone and
105 14 enamel+dentine samples from 23 specimens of large mammals from multiple excavation
106 units within stratum B1 (Fig. 1b, Extended Data Fig. 1, Extended Data Table 1, Table S3). This
107 is an ashfall deposit that contains faunal remains in different geomorphic contexts. All of
108 these are firmly dated between 1.85-1.76 Ma¹⁹. High-resolution tandem MS was used to
109 confidently sequence ancient protein residues from the set of faunal remains, after
110 digestion-based (protocols A and B), or digestion-free (protocol C), sample preparation
111 (Methods and Supplementary Information). Ancient DNA analysis was unsuccessfully
112 attempted on a subset of five bone and dentine specimens (Methods).

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

118 The small proteome^{20,21} of mature dental enamel consists of structural enamel
119 proteins, i.e. amelogenin (AMELX), enamelin (ENAM), amelotin (AMTN), and ameloblastin
120 (AMBN), and enamel-specific proteases secreted during amelogenesis, i.e. matrix
121 metalloproteinase-20 (MMP20) and kallikrein 4 (KLK4). The presence of non-specific
122 proteins, such as serum albumin (ALB) and collagen type I, has also been previously
123 reported in mature dental enamel²⁰ (Extended Data Table 2). The depth of coverage for
124 these proteins varied considerably across their sequence, with some positions covered by
125 over 1000 peptide spectrum matches (Extended Data Fig. 2). The high depth of coverage
126 also allows to identify multiple isoforms of AMELX (Extended Data Fig. 3).

127 Multiple lines of evidence support the authenticity and the endogenous origin of the
128 sequences recovered. Dental enamel proteins are extremely tissue-specific and confined to
129 the dental enamel mineral matrix²⁰. The amino acid composition of the intra-crystalline
130 protein fraction, measured by amino acid racemisation analysis, indicates that the dental
131 enamel behaves as a closed system, unaffected by amino acid and protein residues
132 exchange with the burial environment (Extended Data Fig. 4). The measured rate of
133 asparagine and glutamine deamidation, a spontaneous form of hydrolytic damage
134 consistently observed in ancient samples²², is particularly advanced. Deamidation in Dmanisi
135 enamel is higher than in the control enamel sample, supporting the antiquity of the
136 peptides recovered (Fig. 2a, Supplementary Information). Other forms of non-enzymatic
137 modifications are also present. Tyrosine (Y) experienced mono- and di-oxidation while
138 tryptophan (W) was extensively converted into multiple oxidation products (Fig. 2b,
139 Supplementary Information). Oxidative degradation of histidine (H) and conversion of
140 arginine (R) leading to ornithine accumulation were also observed (Supplementary
141 Information). These modifications are absent, or much less frequent, in the control sample.
142 Similarly, unlike in the control, the peptide length distribution in the Dmanisi dataset is
143 dominated by shorter fragments, generated by advanced, diagenetically-induced, terminal
144 hydrolysis²³ (Fig. 2c, d). Together all these independent lines of evidence clearly define the
145 substantial biomolecular damage affecting the proteomes retrieved and independently
146 support the authenticity of the sequences reconstructed. To demonstrate beyond
147 reasonable doubt the correct peptide sequence assignments of our MS2 spectra, we
148 performed manual validation of peptide-spectrum-matches, conducted fragment ion
149 intensity predictions, and generated synthetic peptides, for a range of phylogenetically

150 informative and phosphorylated peptides (Methods and Supplementary Information: Key
151 MS2 Spectra).

152 We confidently detect phosphorylation (Fig. 3, Extended Data Figs. 2, 5), a stable and
153 tightly *in vivo* regulated physiological post-translational modification (PTM) previously
154 detected in dental enamel proteins^{24,25}. Most of the phosphorylated sites we identified
155 belong to the S-x-E/phS motif, recognised by the secreted kinases of the Fam20C family,
156 which are involved in phosphorylation of extracellular proteins and regulation of
157 biomineralization²⁶. Spectra supporting the identification of serine phosphorylation were
158 validated manually and by comparison with MS2 obtained from synthetic peptides
159 (Supplementary Information), confirming the automated MaxQuant identifications.
160 Phosphorylated serine and threonine residues may be subjected to spontaneous
161 dephosphorylation. However, by complexing with the Ca²⁺ ions in the enamel
162 hydroxyapatite matrix, the peptide-bound phosphate groups can remain stable over
163 millennia, as recently observed in ancient bone²⁷. Previous studies demonstrated that, when
164 complexed with mineral matrix, ~3.8 Ma protein residues can be retrieved from sub-tropical
165 environments¹³. Limited availability of free water in the enamel matrix further reduces
166 spontaneous dephosphorylation via beta-elimination. Altogether, these observations
167 demonstrate that the heavily modified dental enamel proteome retrieved from the ~1.77
168 Ma old Dmanisi faunal material is endogenous and almost complete.

169 Next, we used the palaeoproteomic sequence information to improve taxonomic
170 assignment and achieve sex attribution for some of the Dmanisi faunal remains.
171 Phylogenetic analysis of the five largest enamel+dentine proteomes, and of a moderately
172 large bone proteome, allowed to confirm or improve the morphological identification of
173 their specimens of origin (Extended Data Fig. 6; Figs. S10-15). In addition, confident

174 identification of peptides specific for the isoform Y of amelogenin, coded on the non-
175 recombinant portion of the Y chromosome, indicates that four tooth specimens, namely
176 Dm.6/151.4.A4.12-16630 (*Pseudodama*), Dm.69/64.3.B1.53-16631 (Cervidae),
177 Dm.8/154.4.A4.22-16639 (Bovidae), and Dm.M6/7.II.296-16856 (Cervidae), belonged to
178 male individuals²¹ (Extended Data Fig. 7a-d).

179 An enamel+dentine fragment, from the lower molar of a *Stephanorhinus* ex gr.
180 *etruscus-hundsheimensis* (Dm.5/157-16635; Fig. 1c, Supplementary Information), returned
181 the highest proteomic sequence coverage, encompassing a total of 875 amino acids, across
182 987 peptides (6 proteins; Extended Data Fig. 2; Supplementary Information). Following
183 alignment of the enamel protein sequences retrieved from Dm.5/157-16635 against their
184 homologues from all the extant rhinoceros species, plus the extinct woolly rhinoceros
185 (*†Coelodonta antiquitatis*) and Merck's rhinoceros (*†Stephanorhinus kirchbergensis*),
186 phylogenetic reconstructions place the Dmanisi specimen closer to the extinct woolly and
187 Merck's rhinoceroses than to the extant Sumatran rhinoceros (*Dicerorhinus sumatrensis*), as
188 an early divergent sister lineage (Fig. 4; Extended Data Fig. 8).

189 Our phylogenetic reconstruction confidently recovers the expected differentiation of
190 the *Rhinoceros* genus from other genera considered, in agreement with previous cladistic²⁸
191 and genetic analyses²⁹ (Supplementary Information). This topology defines two-horned
192 rhinoceroses as monophyletic and the one-horned condition as plesiomorphic, as previously
193 proposed (Supplementary Information). We caution, however, that the higher-level
194 relationships we observe between the rhinoceros monophyletic clades might be affected by
195 demographic events, such as incomplete lineage sorting³⁰ and/or gene flow between
196 groups³¹, due to the limited number of markers considered. A confident and stable
197 reconstruction of the structure of the Rhinocerotidae family needs the strong support only

198 high-resolution whole-genome sequencing can provide. Regardless, the highly supported
199 placement of the Dmanisi rhinoceros in the (*Stephanorhinus*, Woolly, Sumatran) clade will
200 remain unaffected, should deeper phylogenetic relationships between the *Rhinoceros* genus
201 and other family members be revised (Extended Data Fig. 8).

202 The phylogenetic relationships of the genus *Stephanorhinus* within the family
203 Rhinocerotidae, as well as those of the several species recognized within this genus, are
204 contentious. *Stephanorhinus* was initially included in the extant South-East Asian genus
205 *Dicerorhinus* represented by the Sumatran rhinoceros species (*D. sumatrensis*)³². This
206 hypothesis has been rejected and, based on morphological data, *Stephanorhinus* has been
207 identified as a sister taxon of the woolly rhinoceros³³. Furthermore, ancient DNA analysis
208 supports a sister relationship between the woolly rhinoceros and *D. sumatrensis*^{7,34,35}.
209 As the *Stephanorhinus* ex gr. *etruscus-hundsheimensis* sequences from Dmanisi branch off
210 basal to the common ancestor of the woolly and Merck's rhinoceroses, these two species
211 most likely derived from an early *Stephanorhinus* lineage expanding eastward from western
212 Eurasia. Throughout the Plio-Pleistocene, *Coelodonta* adapted to continental and later to
213 cold-climate habitats in central Asia. Its earliest representative, *C. thibetana*, displayed some
214 clear *Stephanorhinus*-like anatomical features³³. The presence in eastern Europe and
215 Anatolia of the genus *Stephanorhinus*³⁵ is documented at least since the late Miocene, and
216 the Dmanisi specimen most likely represents an Early Pleistocene descendent of the
217 Western-Eurasian branch of this genus.

218 Ultimately, our phylogenetic reconstructions show that, as currently defined, the
219 genus *Stephanorhinus* is paraphyletic, in line with previous morphological and
220 palaeobiogeographical evidence (Supplementary Information). Accordingly, a systematic

221 revision of the genera *Stephanorhinus* and *Coelodonta*, as well as their closest relatives, is
222 needed.

223 In this study, we show that enamel proteome sequencing can overcome the time
224 limits of ancient DNA preservation and the reduced phylogenetic content of COL1
225 sequences. Given the abundance of teeth in the palaeontological record, the approach
226 presented here holds the potential to address a wide range of questions pertaining to the
227 Early and Middle Pleistocene evolutionary history of a large number of mammals, including
228 hominins, at least in temperate climates.

229 REFERENCES

230

- 231 1 Cappellini, E. *et al.* Ancient Biomolecules and Evolutionary Inference. *Annual Review*
 232 *of Biochemistry* **87**, 1029-1060, doi:10.1146/annurev-biochem-062917-012002
 233 (2018).
- 234 2 Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harbor*
 235 *Perspectives in Biology* **5**, a012567, doi:10.1101/cshperspect.a012567 (2013).
- 236 3 Meyer, M. *et al.* Nuclear DNA sequences from the Middle Pleistocene Sima de los
 237 Huesos hominins. *Nature* **531**, 504-507, doi:10.1038/nature17405 (2016).
- 238 4 Wadsworth, C. & Buckley, M. Proteome degradation in fossils: investigating the
 239 longevity of protein survival in ancient bone. *Rapid Communications in Mass*
 240 *Spectrometry* **28**, 605-615, doi:10.1002/rcm.6821 (2014).
- 241 5 Schweitzer, M. H. *et al.* Analyses of Soft Tissue from *Tyrannosaurus rex* Suggest the
 242 Presence of Protein. *Science* **316**, 277-280, doi:10.1126/science.1138709 (2007).
- 243 6 Schroeter, E. R. *et al.* Expansion for the *Brachylophosaurus canadensis* Collagen I
 244 Sequence and Additional Evidence of the Preservation of Cretaceous Protein. *Journal*
 245 *of Proteome Research* **16**, 920-932, doi:10.1021/acs.jproteome.6b00873 (2017).
- 246 7 Willerslev, E. *et al.* Analysis of complete mitochondrial genomes from extinct and
 247 extant rhinoceroses reveals lack of phylogenetic resolution. *BMC Evolutionary*
 248 *Biology* **9**, 95, doi:10.1186/1471-2148-9-95 (2009).
- 249 8 Welker, F. *et al.* Middle Pleistocene protein sequences from the rhinoceros genus
 250 *Stephanorhinus* and the phylogeny of extant and extinct Middle/Late Pleistocene
 251 *Rhinocerotidae*. *PeerJ* **5**, e3033, doi:10.7717/peerj.3033 (2017).
- 252 9 Kirillova, I. *et al.* Discovery of the skull of *Stephanorhinus kirchbergensis* (Jäger,
 253 1839) above the Arctic Circle. *Quaternary Research* **88**, 537-550,
 254 doi:10.1017/qua.2017.53 (2017).
- 255 10 Lordkipanidze, D. *et al.* A complete skull from Dmanisi, Georgia, and the evolutionary
 256 biology of early *Homo*. *Science* **342**, 326-331, doi:10.1126/science.1238484 (2013).
- 257 11 Eastoe, J. E. Organic Matrix of Tooth Enamel. *Nature* **187**, 411-412,
 258 doi:10.1038/187411b0 (1960).
- 259 12 Orlando, L. *et al.* Recalibrating *Equus* evolution using the genome sequence of an
 260 early Middle Pleistocene horse. *Nature* **499**, 74-78, doi:10.1038/nature12323 (2013).
- 261 13 Demarchi, B. *et al.* Protein sequences bound to mineral surfaces persist into deep
 262 time. *eLife* **5**, e17092, doi:10.7554/eLife.17092 (2016).
- 263 14 Welker, F. *et al.* Ancient proteins resolve the evolutionary history of Darwin's South
 264 American ungulates. *Nature* **522**, 81-84, doi:10.1038/nature14249 (2015).
- 265 15 Chen, F. *et al.* A late Middle Pleistocene Denisovan mandible from the Tibetan
 266 Plateau. *Nature* **569**, 409-412, doi:10.1038/s41586-019-1139-x (2019).
- 267 16 Nei, M. *Molecular evolutionary genetics*. Vol. 75 (Columbia University Press, 1987).
- 268 17 Buckley, M., Warwood, S., van Dongen, B., Kitchener, A. C. & Manning, P. L. A fossil
 269 protein chimera; difficulties in discriminating dinosaur peptide sequences from
 270 modern cross-contamination. *Proceedings of the Royal Society: Biological sciences*
 271 **284**, 20170544, doi:10.1098/rspb.2017.0544 (2017).

- 272 18 Gabunia, L. *et al.* Earliest Pleistocene hominid cranial remains from Dmanisi,
273 Republic of Georgia: taxonomy, geological setting, and age. *Science* **288**, 1019-1025,
274 doi:10.1126/science.288.5468.1019 (2000).
- 275 19 Ferring, R. *et al.* Earliest human occupations at Dmanisi (Georgian Caucasus) dated to
276 1.85-1.78 Ma. *Proceedings of the National Academy of Sciences of the United States*
277 *of America* **108**, 10432-10436, doi:10.1073/pnas.1106638108 (2011).
- 278 20 Castiblanco, G. A. *et al.* Identification of proteins from human permanent erupted
279 enamel. *European Journal of Oral Sciences* **123**, 390-395, doi:10.1111/eos.12214
280 (2015).
- 281 21 Stewart, N. A. *et al.* The identification of peptides by nanoLC-MS/MS from human
282 surface tooth enamel following a simple acid etch extraction. *RSC Advances* **6**,
283 61673-61679, doi:10.1039/c6ra05120k (2016).
- 284 22 van Doorn, N. L., Wilson, J., Hollund, H., Soressi, M. & Collins, M. J. Site-specific
285 deamidation of glutamine: a new marker of bone collagen deterioration. *Rapid*
286 *Communications in Mass Spectrometry* **26**, 2319-2327, doi:10.1002/rcm.6351 (2012).
- 287 23 Catak, S., Monard, G., Aviyente, V. & Ruiz-Lopez, M. F. Computational study on
288 nonenzymatic peptide bond cleavage at asparagine and aspartic acid. *J Phys Chem A*
289 **112**, 8752-8761, doi:10.1021/jp8015497 (2008).
- 290 24 Hunter, T. Why nature chose phosphate to modify proteins. *Philosophical*
291 *Transactions of the Royal Society B* **367**, 2513-2516, doi:10.1098/rstb.2012.0013
292 (2012).
- 293 25 Hu, J. C. C., Yamakoshi, Y., Yamakoshi, F., Krebsbach, P. H. & Simmer, J. P. Proteomics
294 and Genetics of Dental Enamel. *Cells Tissues Organs* **181**, 219-231,
295 doi:10.1159/000091383 (2005).
- 296 26 Tagliabracci, V. S. *et al.* Secreted kinase phosphorylates extracellular proteins that
297 regulate biomineralization. *Science* **336**, 1150-1153, doi:10.1126/science.1217817
298 (2012).
- 299 27 Cleland, T. P. Solid Digestion of Demineralized Bone as a Method to Access
300 Potentially Insoluble Proteins and Post-Translational Modifications. *Journal of*
301 *Proteome Research* **17**, 536-542, doi:10.1021/acs.jproteome.7b00670 (2018).
- 302 28 Antoine, P. O. *et al.* A revision of *Aceratherium blanfordi* Lydekker, 1884 (Mammalia:
303 Rhinocerotidae) from the Early Miocene of Pakistan: postcranials as a key. *Zoological*
304 *Journal of the Linnean Society* **160**, 139-194, doi:10.1111/j.1096-3642.2009.00597.x
305 (2010).
- 306 29 Steiner, C. C. & Ryder, O. A. Molecular phylogeny and evolution of the
307 Perissodactyla. *Zoological Journal of the Linnean Society* **163**, 1289-1303,
308 doi:10.1111/j.1096-3642.2011.00752.x (2011).
- 309 30 Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H. & Mailund, T. Incomplete
310 lineage sorting patterns among human, chimpanzee, and orangutan suggest recent
311 orangutan speciation and widespread selection. *Genome research* **21**, 349-356,
312 doi:10.1101/gr.114751.110 (2011).
- 313 31 Rieseberg, L. H. Evolution: replacing genes and traits through hybridization. *Current*
314 *Biology* **19**, R119-R122, doi:10.1016/j.cub.2008.12.016 (2009).
- 315 32 Guérin, C. Les rhinocéros (Mammalia, Perissodactyla) du Miocène terminal au
316 Pleistocène supérieur en Europe occidentale, comparaison avec les espèces
317 actuelles. *Documents du Laboratoire de Géologie de la Faculté des Sciences de Lyon*
318 **79**, 3-1183 (1980).

319 33 Deng, T. *et al.* Out of Tibet: pliocene woolly rhino suggests high-plateau origin of Ice
320 Age megaherbivores. *Science* **333**, 1285-1288, doi:10.1126/science.1206594 (2011).
321 34 Orlando, L. *et al.* Ancient DNA analysis reveals woolly rhino evolutionary
322 relationships. *Molecular Phylogenetics and Evolution* **28**, 485-499,
323 doi:10.1016/S1055-7903(03)00023-X (2003).
324 35 Yuan, J. *et al.* Ancient DNA sequences from *Coelodonta antiquitatis* in China reveal
325 its divergence and phylogeny. *Science China Earth Sciences* **57**, 388-396,
326 doi:10.1007/s11430-013-4702-6 (2014).
327

328 MAIN TEXT FIGURE LEGENDS

329 **Figure 1. Dmanisi location, stratigraphy, and *Stephanorhinus* specimen GNM Dm.5/157-**
330 **16635. a,** Geographic location of Dmanisi in the South Caucasus. The base map was
331 generated using public domain data from www.natureearthdata.com. **b,** Generalised
332 stratigraphic profile indicating origin and age of the analysed specimens. **c,** Isolated left
333 lower molar (m1 or m2) of *Stephanorhinus* ex gr. *etruscus-hundsheimensis*, from Dmanisi
334 (labial view). Scale bar: 1 cm.

335

336

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

348

349

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

354

355

356 **Figure 4. Phylogenetic relationships between the comparative enamel proteome dataset**
357 **and specimen Dm.5/157-16635 (*Stephanorhinus* ex gr. *etruscus-hundsheimensis*).**

358 Consensus tree from Bayesian inference on the concatenated alignment of six enamel
359 proteins, using *Homo sapiens* as an outgroup. For each bipartition, we show the posterior
360 probability obtained from the Bayesian inference. Additionally, for bipartitions where the
361 Bayesian and the Maximum-likelihood inference support are different, we show (right) the
362 support obtained in the latter. Scale indicates estimated branch lengths.

363

364

365 METHODS

366

367 **Dmanisi & sample selection**

368 Dmanisi is located about 65 km southwest of the capital city of Tbilisi in the Kvemo Kartli
369 region of Georgia, at an elevation of 910 meters above sea level (Lat: 41° 20' N, Lon: 44° 20'
370 E)^{10,18}. The 23 fossil specimens we analysed were retrieved from stratum B1, in excavation
371 blocks M17, M6, block 2, and area R11 (Extended Data Table 1, Extended Data Fig. 1).
372 Stratum B deposits date between 1.78 Ma and 1.76 Ma¹⁹. All the analysed specimens were
373 collected between 1984 and 2014 and their taxonomic identification was based on
374 traditional comparative anatomy.

375 After the sample preparation and data acquisition for all the Dmanisi specimens was
376 concluded, we applied the whole experimental procedure to a medieval ovicaprine
377 (sheep/goat) dental enamel+dentine specimen that was used as control. For this sample, we
378 used extraction protocol "C", and generated tandem MS data using a Q Exactive HF mass
379 spectrometer (Thermo Fisher Scientific). The data were searched against the goat
380 proteome, downloaded from the NCBI Reference Sequence Database (RefSeq) archive on
381 31st May 2017 (Supplementary Information). The ovicaprine specimen was found at the
382 "Hotel Skandinavia" site in the city of Århus, Denmark and stored at the Natural History
383 Museum of Denmark, Copenhagen.

384

385 **Biomolecular preservation**

386 We assessed the potential of ancient protein preservation prior to proteomic analysis by
387 measuring the extent of amino acid racemisation in a subset of samples (6/23)³⁶. Enamel
388 chips, with all dentine removed, were powdered, and two subsamples per specimen were

389 subject to analysis of their free (FAA) and total hydrolysable (THAA) amino acid fractions.
390 Samples were analysed in duplicate by RP-HPLC, with standards and blanks run alongside
391 each one of them (Supplementary Information). The D/L values of aspartic acid/asparagine,
392 glutamic acid/glutamine, phenylalanine and alanine (D/L Asx, Glx, Phe, Ala) were assessed
393 (Extended Data Fig. 4) to provide an overall estimate of intra-crystalline protein
394 decomposition (IcPD).

395

396 **PROTEOMICS**

397 All the sample preparation procedures for palaeoproteomic analysis were conducted in
398 laboratories dedicated to the analysis of ancient DNA and ancient proteins in clean rooms
399 fitted with filtered ventilation and positive pressure, in line with recent recommendations
400 for ancient protein analysis³⁷. A mock “extraction blank”, containing no starting material,
401 was prepared, processed and analysed together with each batch of ancient samples.

402

403 **Sample preparation**

404 The external surface of bone samples was gently removed, and the remaining material was
405 subsequently powdered. Enamel fragments, occasionally mixed with small amounts of
406 dentine, were removed from teeth with a cutting disc and subsequently crushed into a
407 rough powder. Ancient protein residues were extracted from approximately 180-220 mg of
408 mineralised material, unless otherwise specified, using three different extraction protocols,
409 hereafter referred to as “A”, “B” and “C” (Supplementary Information):

410

411 **EXTRACTION PROTOCOL A - FASP.** Tryptic peptides were generated using a filter-aided sample
412 preparation (FASP) approach³⁸, as previously performed on ancient samples³⁹.

413 **EXTRACTION PROTOCOL B - GuHCl SOLUTION AND DIGESTION.** Bone or enamel+dentine powder was
414 demineralised in 1 mL 0.5 M EDTA pH 8.0. After removal of the supernatant, all
415 demineralised pellets were re-suspended in a 300 μ L solution containing 2 M guanidine
416 hydrochloride (GuHCl, Thermo Scientific), 100 mM Tris pH 8.0, 20 mM 2-Chloroacetamide
417 (CAA), 10 mM Tris (2-carboxyethyl)phosphine (TCEP) in ultrapure H₂O^{40,41}. A total of 0.2 μ g
418 of mass spectrometry-grade rLysC (Promega P/N V1671) enzyme was added before the
419 samples were incubated for 3-4 hours at 37°C with agitation. Samples and negative controls
420 were subsequently diluted to 0.6 M GuHCl, and 0.8 μ g of mass spectrometry-grade Trypsin
421 (Promega P/N V5111) was added. Next, samples and negative controls were incubated
422 overnight under mechanical agitation at 37°C. On the following day, samples were acidified,
423 and the tryptic peptides were purified on C18 Stage-Tips, as previously described⁴².

424

425 **EXTRACTION PROTOCOL C - DIGESTION-FREE ACID DEMINERALISATION.** Dental enamel powder, with
426 possible trace amounts of dentine, was demineralised in 1.2 M HCl at room temperature,
427 after which the solubilised protein residues were directly cleaned and concentrated on
428 Stage-Tips, as described above. The sample prepared on Stage-Tip “#1217” was processed
429 with 10% TFA instead of 1.2 M HCl. All the other parameters and procedures were identical
430 to those used for all the other samples extracted with protocol “C”.

431

432 **Tandem mass spectrometry**

433 Different sets of samples (Supplementary Information §5.1, 5.2) were analysed by nanoflow
434 liquid chromatography coupled to tandem mass spectrometry (nanoLC-MS/MS) on an EASY-
435 nLC™ 1000 or 1200 system connected to a Q-Exactive, a Q-Exactive Plus, or to a Q-Exactive
436 HF (Thermo Scientific, Bremen, Germany) mass spectrometer. Before and after each MS/MS

437 run measuring ancient or extraction blank samples, two successive MS/MS runs were
438 included in the sample queue in order to prevent carryover contamination between the
439 samples. These consisted, first, of a MS/MS run ("MS/MS blank" run) with an injection
440 exclusively of the buffer used to re-suspend the samples (0.1% TFA, 5% ACN), followed by a
441 second MS/MS run ("MS/MS wash" run) with no injection.

442

443 **Data analysis**

444 Raw data files generated during MS/MS spectral acquisition were searched using
445 MaxQuant⁴³, version 1.5.3.30, and PEAKS⁴⁴, version 7.5. A two-stage peptide-spectrum
446 matching approach was adopted (Supplementary Information §5.3). Raw files were initially
447 searched against a target/reverse database of collagen and enamel proteins retrieved from
448 the UniProt and NCBI Reference Sequence Database (RefSeq) archives^{45,46}, taxonomically
449 restricted to mammalian species. A database of partial "COL1A1" and "COL1A2" sequences
450 from cervid species⁴⁷ was also included. The results from the preliminary analysis were used
451 for a first, provisional reconstruction of protein sequences (MaxQuant search 1, MQ1).

452 For specimens whose dataset resulted in a narrower, though not fully resolved,
453 initial taxonomic placement, a second MaxQuant search (MQ2) was performed using a new
454 protein database taxonomically restricted to the "order" taxonomic rank as determined
455 after MQ1. For the MQ2 matching of the MS/MS spectra from specimen Dm.5/157-16635,
456 partial sequences of serum albumin and enamel proteins from Sumatran (*Dicerorhinus*
457 *sumatrensis*), Javan (*Rhinoceros sondaicus*), Indian (*Rhinoceros unicornis*), woolly
458 (*Coelodonta antiquitatis*), Mercks (*Stephanorhinus kirchbergensis*), and Black rhinoceros
459 (*Diceros bicornis*), were also added to the protein database. All the protein sequences from

460 these species were reconstructed from draft genomes for each species (Dalen and Gilbert,
461 unpublished data, Supplementary Information).

462 For each MaxQuant and PEAKS search, enzymatic digestion was set to “unspecific”
463 and the following variable modifications were included: oxidation (M), deamidation (NQ), N-
464 term Pyro-Glu (Q), N-term Pyro-Glu (E), hydroxylation (P), phosphorylation (S). The error
465 tolerance was set to 5 ppm for the precursor and to 20 ppm, or 0.05 Da, for the fragment
466 ions in MaxQuant and PEAKS respectively. For searches of data generated from sample
467 fractions partially or exclusively digested with trypsin, another MaxQuant and PEAKS search
468 was conducted using the “enzyme” parameter set to “Trypsin/P”. Carbamidomethylation (C)
469 was set: (i) as a fixed modification, for searches of data generated from sets of sample
470 fractions exclusively digested with trypsin, or (ii) as a variable modification, for searches of
471 data generated from sets of sample fractions partially digested with trypsin. For searches of
472 data generated exclusively from undigested sample fractions, carbamidomethylation (C)
473 was not included as a modification, neither fixed nor variable.

474 The datasets re-analysed with MQ2 search, were also processed with the PEAKS
475 software using the entire workflow (PEAKS *de novo* to PEAKS SPIDER) in order to detect
476 hitherto unreported single amino acid polymorphisms (SAPs). Any amino acid substitution
477 detected by the “SPIDER” homology search algorithm was validated by repeating the
478 MaxQuant search (MQ3). In MQ3, the protein database used for MQ2 was modified to
479 include the amino acid substitutions detected by the “SPIDER” algorithm.

480

481 **Ancient protein sequence reconstruction**

482 The peptide sequences confidently identified by the MQ1, MQ2, MQ3 were aligned using
483 the software Geneious⁴⁸ (v. 5.4.4, substitution matrix BLOSUM62). The peptide sequences

484 confidently identified by the PEAKS searches were aligned using an in-house R-script. A
485 consensus sequence for each protein from each specimen was generated in FASTA format,
486 without filtering on depth of coverage. Amino acid positions that were not confidently
487 reconstructed were replaced by an "X". Novel SAPs discovered through PEAKS were only
488 accepted if these were further validated by repeating the MaxQuant search (MQ3). All
489 leucines were converted into isoleucines, as standard MS/MS cannot differentiate between
490 these two isobaric amino acids. For possible deamidated sites, we checked whether there
491 were positions in our reference sequence database where both Q and E or both N and D
492 occurred on the same position, and where we also had ancient sequences matching. For
493 sample Dm.5/157-16635, only one such position existed, and this was replaced by an "X" in
494 our consensus sequence. Based on parsimony, for other Q, E, N, and D positions we called
495 the amino acid present in the reference proteome, regardless of their phylogenetic
496 relevance. The output of the MQ2 and 3 searches was used to extend the coverage of the
497 ancient protein sequences initially identified in the MQ1 iteration. For specimen DM.5/157-
498 16335, all the experimentally identified peptides, as well as the respective best matching
499 MS/MS spectra covering the sites informative for Rhinocerotidae phylogenetic inference,
500 are provided as Supplementary Information ("Key MS-MS Spectra" file). All the reported
501 MS/MS spectra are annotated using the advanced annotation mode of MaxQuant. Selected
502 spectra matching to peptides covering phylogenetically informative amino acid positions
503 were manually inspected, validated and annotated by an experienced mass spectrometrists,
504 in all cases in full agreement with bioinformatic sequence assignment (Supplementary
505 Information, "Key MS-MS Spectra" file). We utilized MS²PIP fragment ion spectral intensity
506 prediction⁴⁹ (version: v20190107; model: HCD) to demonstrate that the experimentally
507 observed fragment ion intensities are highly correlated with the theoretical ones (Fig. S3).

508 Finally, we generated synthetic peptides for 19 selected peptides covering Rhinocerotidae
509 SAPs in DM.5/157-16635.

510

511 **Post translational modifications**

512 **DEAMIDATION.** After removal of likely contaminants, the extent of glutamine and asparagine
513 deamidation was estimated for individual specimens, by using the MaxQuant output files as
514 previously published⁴¹ (Supplementary Information).

515 **OTHER SPONTANEOUS CHEMICAL MODIFICATIONS.** Spontaneous post-translational modifications
516 (PTMs) associated with chemical protein damage were searched using the PEAKS PTM tool
517 and the dependent peptides search mode⁵⁰ in MaxQuant. In the PEAKS PTM search, all
518 modifications in the Unimod database were considered. The mass error was set to 5.0 ppm
519 and 0.5 Da for precursor and fragment, respectively. For PEAKS, the *de novo* ALC score was
520 set to a threshold of 15 % and the peptide hit threshold to 30. The results were filtered by
521 an FDR of 5 %, *de novo* ALC score of 50 %, and a protein hit threshold of ≥ 20 . The
522 MaxQuant dependent peptides search was carried out with the same search settings as
523 described above and with a dependent peptide FDR of 1 % and a mass bin size of 0.0065 Da.

524 **PHOSPHORYLATION.** Class I phosphorylation sites were selected with localisation probabilities
525 of ≥ 0.98 in the Phosph(ST)Sites MaxQuant output file. Sequence windows of ± 6 aa from all
526 identified sites were compared against a background file containing all non-phosphorylated
527 peptides using a linear kinase sequence motif enrichment analysis in IceLogo (version
528 1.3.8)⁵¹.

529

530 **PHYLOGENETIC ANALYSIS**

531 **Reference datasets**

532 We assembled a reference dataset consisting of publicly available protein sequences from
533 representative ungulate species belonging to the following families: Equidae,
534 Rhinocerotidae, Suidae and Bovidae (Supplementary Information §7 and §8). As Cervidae
535 and carnivores are absent from protein sequence databases to a various extent, we did not
536 attempt phylogenetic placement of samples from these taxa. Instead, we conducted our
537 phylogenetic analysis on the five best-performing enamel proteomes (Dm.5/154.2.A4.38-
538 16632), Dm.5/157-16635, Dm.5/154.1.B1.1-16638, Dm.8/154.4.A4.22-16639,
539 Dm.8/152.3.B1.2-16641) and the largest bone proteome (Dm.bXI.North.B1a.collection-
540 16658) we recovered (see Extended Data Table 2).

541 We extended this dataset with the protein sequences from extinct and extant
542 rhinoceros species including: the woolly rhinoceros († *Coelodonta antiquitatis*), the Merck's
543 rhinoceros († *Stephanorhinus kirchbergensis*), the Sumatran rhinoceros (*Dicerorhinus*
544 *sumatrensis*), the Javan rhinoceros (*Rhinoceros sondaicus*), the Indian rhinoceros
545 (*Rhinoceros unicornis*), and the Black rhinoceros (*Diceros bicornis*). Their corresponding
546 protein sequences were obtained following translation of high-throughput DNA sequencing
547 data, after filtering reads with mapping quality lower than 30 and nucleotides with base
548 quality lower than 20, and calling the majority rule consensus sequence using ANGSD⁵² For
549 the woolly and Merck's rhinoceroses we excluded the first and last five nucleotides of each
550 DNA fragment in order to minimize the effect of post-mortem ancient DNA damage⁵³. Each
551 consensus sequence was formatted as a separate blast nucleotide database. We then
552 performed a tblastn⁵⁴ alignment using the corresponding white rhinoceros sequence as a

553 query, favouring ungapped alignments in order to recover translated and spliced protein
554 sequences. Resulting alignments were processed using ProSplign algorithm from the NCBI
555 Eukaryotic Genome Annotation Pipeline⁵⁵ to recover the spliced alignments and translated
556 protein sequences.

557

558 **Construction of phylogenetic trees**

559 For each specimen, multiple sequence alignments for each protein were built using MAFFT⁵⁶
560 and concatenated onto a single alignment per specimen. These were inspected visually to
561 correct obvious alignment mistakes, and all the isoleucine residues were substituted with
562 leucine ones to account for indistinguishable isobaric amino acids at the positions where the
563 ancient protein carried one of such amino acids. Based on these alignments, we inferred the
564 phylogenetic relationship between the ancient samples and the species included in the
565 reference dataset by using three approaches: distance-based neighbour-joining, maximum
566 likelihood and Bayesian phylogenetic inference (Supplementary Information).

567 Neighbour-joining trees were built using the phangorn⁵⁷ R package, restricting to
568 sites covered in the ancient samples. Genetic distances were estimated using the JTT model,
569 considering pairwise deletions. We estimated bipartition support through a non-parametric
570 bootstrap procedure using 500 pseudoreplicates. We used PHyML 3.1⁵⁸ for maximum
571 likelihood inference based on the whole concatenated alignment. For likelihood
572 computation, we used the JTT substitution model with two additional parameters for
573 modelling rate heterogeneity and the proportion of invariant sites. Bipartition support was
574 estimated using a non-parametric bootstrap procedure with 500 replicates. Bayesian
575 phylogenetic inference was carried out using MrBayes 3.2.6⁵⁹ on each concatenated
576 alignment, partitioned per gene. While we chose the JTT substitution model in the two

577 approaches above, we allowed the Markov chain to sample parameters for the substitution
578 rates from a set of predetermined matrices, as well as the shape parameter of a gamma
579 distribution for modelling across-site rate variation and the proportion of invariable sites.
580 The MCMC algorithm was run with 4 chains for 5,000,000 cycles. Sampling was conducted
581 every 500 cycles and the first 25% were discarded as burn-in. Convergence was assessed
582 using Tracer v. 1.6.0, which estimated an ESS greater than 5,500 for each individual,
583 indicating reasonable convergence for all runs.

584

585 **ANCIENT DNA ANALYSIS**

586 The samples were processed using strict aDNA guidelines in a clean lab facility at the Natural
587 History Museum of Denmark, University of Copenhagen. DNA extraction was attempted on
588 five of the ancient animal samples (Supplementary Information §9, §13). Powdered samples
589 (120-140 mg) were extracted using a silica-in-solution method^{12,60}. To prepare the samples
590 for NGS sequencing, 20 µL of DNA extract was built into a blunt-end library using the
591 NEBNext DNA Sample Prep Master Mix Set 2 (E6070) with Illumina-specific adapters. The
592 libraries were PCR-amplified with inPE1.0 forward primers and custom-designed reverse
593 primers with a 6-nucleotide index⁶¹. Two extracts (MA399 and MA2481, from specimens
594 16859 and 16635 respectively) yielded detectable DNA concentrations (Table S9). The
595 libraries generated from specimen 16859 and 16635 were processed on different flow cells.
596 They were pooled with others for sequencing on an Illumina 2000 platform (MA399_L1,
597 MA399_L2), using 100bp single read chemistry, and on an Illumina 2500 platform
598 (MA2481_L1), using 81bp single read chemistry.

599 The data were base-called using the Illumina software CASAVA 1.8.2 and sequences
600 were demultiplexed with a requirement of a full match of the six nucleotide indexes that

601 were used. Raw reads were processed using the PALEOMIX pipeline following published
602 guidelines⁶², mapping against the cow nuclear genome (*Bos taurus* 4.6.1, accession
603 GCA_000003205.4), the cow mitochondrial genome (*Bos taurus*), the red deer
604 mitochondrial genome (*Cervus elaphus*, accession AB245427.2), and the human nuclear
605 genome (GRCh37/hg19), using BWA backtrack⁶³ v0.5.10 with the seed disabled. All other
606 parameters were set as default. PCR duplicates from mapped reads were removed using the
607 picard tool *MarkDuplicate* [<http://picard.sourceforge.net/>].

608

609 **SAMPLE Dm.5/157-16635 MORPHOLOGICAL MEASUREMENTS**

610 We followed the methodology introduced by Guérin³². The maximal length of the tooth is
611 measured with a digital calliper at the lingual side of the tooth and parallel to the occlusal
612 surface. All measurements are given in mm (Supplementary Information §3).

613

614

615 ACKNOWLEDGMENTS

616 EC and FW are supported by the VILLUM Fonden (grant number 17649) and by the
617 European Commission through a Marie Skłodowska Curie (MSC) Individual Fellowship (grant
618 number 795569). EW is supported by the Lundbeck Foundation, the Danish National
619 Research Foundation, the Carlsberg Foundation, KU2016, and the Wellcome Trust. EC, CK,
620 JVO, PR and DS are supported by the European Commission through the MSC European
621 Training Network “TEMPERA” (grant number 722606). MM and RRJ-C are supported by the
622 University of Copenhagen KU2016 (UCPH Excellence Programme) grant. MM is also
623 supported by the Danish National Research Foundation award PROTEIOS (DNRF128). Work
624 at the Novo Nordisk Foundation Center for Protein Research is funded in part by a donation
625 from the Novo Nordisk Foundation (grant number NNF14CC0001). MRD is supported by a
626 PhD DTA studentship from NERC and the Natural History Museum (NE/K500987/1 &
627 NE/L501761/1). KP is supported by the Leverhulme Trust (PLP -2012-116). LR and LP are
628 supported by the Italian Ministry for Foreign Affairs (MAECI, DGSP-VI). LP was also
629 supported by the EU-SYNTHESYS project (AT-TAF-2550, DE-TAF-3049, GB-TAF-2825, HU-TAF-
630 3593, ES-TAF-2997) funded by the European Commission. LD is supported by the Swedish
631 Research Council (grant number 2017-04647) and FORMAS (grant nr 2015-676). MTPG is
632 supported by ERC Consolidator Grant “EXTINCTION GENOMICS” (grant number 681396). LO
633 is supported by the ERC Consolidator Grant “PEGASUS” (grant agreement No 681605). BS, JK
634 and PDH are supported by the Gordon and Betty Moore foundation. BM-N is supported by
635 the Spanish Ministry of Sciences (grant number CGL2016-80975-P). The aDNA analysis was
636 carried out using the facilities of the University of Luxembourg, the Swedish Museum of
637 Natural History and UC Santa Cruz. The authors would like to acknowledge support from
638 Science for Life Laboratory, the National Genomics Infrastructure (Sweden), and UPPMAX

639 for providing assistance in massive parallel sequencing and computational infrastructure.
640 Research at Dmanisi is supported by the John Templeton Foundation, the Shota Rustaveli
641 Science Foundation, and the Alexander von Humboldt Fellowship research award. The
642 authors would also like to thank B. Triozzi and K. Murphy Gregersen for technical support.

643

644 AUTHOR CONTRIBUTIONS

645 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.
646 performed laboratory experiments. M.Bu., M.T., R.F., E.P., T.S., Y.L.C., A.Gö., S.KSS.N.,
647 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.,
648 and D.Lo., provided ancient samples or modern reference material. E.C., F.W., L.P., J.R.M.,
649 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
650 analyses and data interpretation. E.C., F.W., J.R.M., L.P. and E.W. wrote the manuscript with
651 contributions from all authors.

652

653 AUTHOR INFORMATION

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

655 The Authors declare no financial competing interests.

656 Correspondence and requests for material should be addressed to E.C.

657 (ecappellini@bio.ku.dk), J.V.O. (jesper.olsen@cpr.ku.dk) or E.W. (ewillerslev@bio.ku.dk).

658

659 METHODS REFERENCES

660 10 Lordkipanidze, D. *et al.* A complete skull from Dmanisi, Georgia, and the evolutionary
661 biology of early Homo. *Science* **342**, 326-331, doi:10.1126/science.1238484 (2013).
662 12 Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an
663 early Middle Pleistocene horse. *Nature* **499**, 74-78, doi:10.1038/nature12323 (2013).
664 18 Gabunia, L. *et al.* Earliest Pleistocene hominid cranial remains from Dmanisi,
665 Republic of Georgia: taxonomy, geological setting, and age. *Science* **288**, 1019-1025,
666 doi:10.1126/science.288.5468.1019 (2000).
667 19 Ferring, R. *et al.* Earliest human occupations at Dmanisi (Georgian Caucasus) dated to
668 1.85-1.78 Ma. *Proceedings of the National Academy of Sciences of the United States*
669 *of America* **108**, 10432-10436, doi:10.1073/pnas.1106638108 (2011).
670 32 Guérin, C. Les rhinocéros (Mammalia, Perissodactyla) du Miocène terminal au
671 Pleistocène supérieur en Europe occidentale, comparaison avec les espèces
672 actuelles. *Documents du Laboratoire de Géologie de la Faculté des Sciences de Lyon*
673 **79**, 3-1183 (1980).
674 36 Penkman, K. E. H., Kaufman, D. S., Maddy, D. & Collins, M. J. Closed-system
675 behaviour of the intra-crystalline fraction of amino acids in mollusc shells.
676 *Quaternary Geochronology* **3**, 2-25, doi:10.1016/j.quageo.2007.07.001 (2008).
677 37 Hendy, J. *et al.* A guide to ancient protein studies. *Nature Ecology & Evolution* **2**, 791-
678 799, doi:10.1038/s41559-018-0510-x (2018).
679 38 Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation
680 method for proteome analysis. *Nature Methods* **6**, 359-362,
681 doi:10.1038/nmeth.1322 (2009).
682 39 Cappellini, E. *et al.* Resolution of the type material of the Asian elephant, *Elephas*
683 *maximus* Linnaeus, 1758 (Proboscidea, Elephantidae). *Zoological Journal of the*
684 *Linnean Society* **170**, 222-232, doi:10.1111/zoj.12084 (2014).
685 40 Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated
686 proteomic-sample processing applied to copy-number estimation in eukaryotic cells.
687 *Nature Methods* **11**, 319-324, doi:10.1038/nmeth.2834 (2014).
688 41 Mackie, M. *et al.* Palaeoproteomic Profiling of Conservation Layers on a 14th Century
689 Italian Wall Painting. *Angewandte Chemie (International ed.)* **57**, 7369-7374,
690 doi:10.1002/anie.201713020 (2018).
691 42 Cappellini, E. *et al.* Proteomic analysis of a pleistocene mammoth femur reveals
692 more than one hundred ancient bone proteins. *Journal of Proteome Research* **11**,
693 917-926, doi:10.1021/pr200721u (2012).
694 43 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates,
695 individualized p.p.b.-range mass accuracies and proteome-wide protein
696 quantification. *Nature Biotechnology* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).
697 44 Zhang, J. *et al.* PEAKS DB: De novo sequencing assisted database search for sensitive
698 and accurate peptide identification. *Molecular and Cellular Proteomics* **11**,
699 M111.010587, doi:10.1074/mcp.M111.010587 (2012).
700 45 TheUniProtConsortium. UniProt: the universal protein knowledgebase. *Nucleic Acids*
701 *Research* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

702 46 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,
703 taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733-
704 D745, doi:10.1093/nar/gkv1189 (2016).

705 47 Welker, F. *et al.* Palaeoproteomic evidence identifies archaic hominins associated
706 with the Châtelperronian at the Grotte du Renne. *Proceedings of the National*
707 *Academy of Sciences* **113**, 11162-11167, doi:10.1073/pnas.1605834113 (2016).

708 48 Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software
709 platform for the organization and analysis of sequence data. *Bioinformatics* **28**,
710 1647-1649, doi:10.1093/bioinformatics/bts199 (2012).

711 49 Gabriels, R., Martens, L. & Degroeve, S. Updated MS2PIP web server delivers fast
712 and accurate MS2 peak intensity prediction for multiple fragmentation methods,
713 instruments and labeling techniques. *bioRxiv*, 544965, doi:10.1101/544965 (2019).

714 50 Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass
715 spectrometry-based shotgun proteomics. *Nature Protocols* **11**, 2301-2319,
716 doi:10.1038/nprot.2016.136 (2016).

717 51 Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved
718 visualization of protein consensus sequences by iceLogo. *Nature Methods* **6**, 786-
719 787, doi:10.1038/nmeth1109-786 (2009).

720 52 Korneliussen, T., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation
721 Sequencing Data. *BMC Bioinformatics* **15**, 356-356, doi:10.1186/s12859-014-0356-4
722 (2014).

723 53 Briggs, A. *et al.* Removal of deaminated cytosines and detection of in vivo
724 methylation in ancient DNA. *Nucleic Acids Research* **38**, e87,
725 doi:10.1093/nar/gkp1163 (2010).

726 54 Altschul, S. F. *et al.* Gapped BLAST and PSI- BLAST: a new generation of protein
727 database search programs. *Nucleic Acids Research* **25**, 3389-3402 (1997).

728 55 SeaUrchinGenomeSequencingConsortium. The Genome of the Sea Urchin
729 *Strongylocentrotus purpuratus*. *Science* **314**, 941-952 (2006).

730 56 Katoh, K. & Frith, M. C. Adding unaligned sequences into an existing alignment using
731 MAFFT and LAST. *Bioinformatics* **28**, 3144-3146, doi:10.1093/bioinformatics/bts578
732 (2012).

733 57 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593,
734 doi:10.1093/bioinformatics/btq706 (2011).

735 58 Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood
736 Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-
737 321, doi:10.1093/sysbio/syq010 (2010).

738 59 Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model
739 Choice Across a Large Model Space. *Systematic Biology* **61**, 539-542,
740 doi:10.1093/sysbio/sys029 (2012).

741 60 Rohland, N. & Hofreiter, M. Comparison and optimization of ancient DNA extraction.
742 *BioTechniques* **42**, 343-352, doi:10.2144/000112383 (2007).

743 61 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly
744 multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*,
745 doi:10.1101/pdb.prot5448 (2010).

746 62 Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP
747 detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature*
748 *Protocols* **9**, 1056-1082, doi:10.1038/nprot.2014.063 (2014).

749 63 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows– Wheeler
750 transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
751 64 Dickinson, M. L., A.; Penkman, K. A new method for enamel amino acid racemization
752 dating: a closed system approach. *Quaternary Geochronology* **50**, 29-46,
753 doi:10.1016/j.quageo.2018.11.005 (2019).
754
755

756 DATA AVAILABILITY

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

762

763

764 CODE AVAILABILITY

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

767

768

769

770 SUPPLEMENTARY INFORMATION

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

772

773 EXTENDED DATA LEGENDS

774

775 **Extended Data Table 1. Genome and proteome survival in 23 Dmanisi fossil fauna**

776 **specimens.** For each specimen, the Centre for GeoGenetics (CGG) reference number and
777 the Georgian National Museum (GNM) specimen field number are reported. *or the
778 narrowest possible taxonomic identification achievable using comparative anatomy
779 methods. †Only collagens survive. B = Bone, D = Dentine, E = Enamel. Extractions of enamel
780 might include some residual dentine. Accordingly, both tissues are either listed separately
781 (○D, ●E, in case of no collagen preservation), or together (●E+D, in case of collagen
782 preservation). Open circles (○) indicate no molecular preservation; (●) closed circles indicate
783 molecular preservation.

784

785

786 **Extended Data Table 2. Proteome composition and coverage.** Aggregated data from

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

795

796 **Extended Data Figure 1. Generalized stratigraphic profiles for Dmanisi, indicating**

797 **specimen origins. a,** Type section of the Dmanisi M5 Excavation block. **b,** Stratigraphic
798 profile of excavation area M6. M6 preserves a larger gully associated with the pipe-gully
799 phase of stratigraphic-geomorphic development in Stratum B1. The thickness of Stratum B1
800 gully fill extends to the basalt surface, but includes “rip-ups” of Strata A1 and A2, showing
801 that B1 deposits post-date Stratum A. **c,** Stratigraphic section of excavation area M17. Here,
802 Stratum B1 was deposited after erosion of Stratum A deposits. The stratigraphic position of
803 the *Stephanorhinus* sample Dm.5/157-16635 is highlighted with a red diamond. The
804 Masavara basalt is ca. 50 cm below the base of the shown profile. **d,** Northern section of
805 Block 2. Following collapse of a pipe and erosion to the basalt, the deeper part of this area
806 was filled with local gully fill of Stratum B1/x/y/z. Note the uniform burial of all Stratum B1
807 deposits by Strata B2-B4. Sampled specimens are indicated by CGG five-digit numbers. See
808 Extended Data Table 1 for both CGG and GNM specimen numbers.

809

810

811 **Extended Data Fig. 2. Proteomic sequence coverage for specimen Dm.5/157-16635**

812 **(*Stephanorhinus*).** **a, c, e, g, i, j,** PSM sequence coverage of proteins AMBN, ENAM, AMELX,

813 AMTN, MMP20 and ALB, respectively. Annotations include: “amino acid position, amino
814 acid called in that position (number of PSMs/peptides covering that position)” for the
815 phylogenetically informative SAPs within Rhinocerotidae. **b, d, f, h**, Frequency (%) of
816 phosphorylated (green) and non-phosphorylated (red) PSMs per amino acid position for
817 AMBN, ENAM, AMELX and AMTN, respectively. Numbers within the bars provide the PSM
818 counts. **k**, Violinplot of PSM coverage distribution for all covered sites (n=693) and those of
819 phylogenetic relevance (SAPs, n=30). The boxplots define the range of the data, with
820 whiskers extending to 1.5 the interquartile range, 25th and 75th percentiles (boxes), and
821 medians (dots). All panels based on MQ results only. Supplementary File “Key MS-MS
822 Spectra” contains spectral examples and fragment ion series alignments for each of the
823 marked SAPs.

824
825

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

835
836

837 **Extended Data Figure 4. Amino Acid Racemisation.** Extent of intra-crystalline racemization
838 in enamel for the free amino acid (FAA, x-axis) fraction and the total hydrolysable amino
839 acids (THAA, y-axis) fraction for four amino acids (Asx, Glx, Ala and Phe). Note differences in
840 axis scale. Intra-crystalline data from Proboscidea enamel from a range of UK sites⁶⁴ has
841 been shown for comparison (black crosses). Both taxa from Dmanisi and the UK exhibit a
842 similar relationship between FAA and THAA racemization and R² values have been
843 calculated based on a polynomial relationship (order = 2, all >0.93).

844
845

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

851
852

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

856

857

858 **Extended Data Figure 7. Amelogenin Y-specific matches. a)** Specimen Dm.6/151.4.A4.12-
859 16630 (*Pseudodama*). **b)** Specimen Dm.69/64.3.B1.53-16631 (Cervidae). **c)** Specimen
860 Dm.8/154.4.A4.22-16639 (Bovidae). **d)** Specimen Dm.M6/7.II.296-16856 (Cervidae). Note
861 the presence of deamidated glutamine (deQ) and asparagine (deN), oxidated methionine
862 (oxM), and phosphorylated serine (phS).

863

864

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

873