

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

Enrico Cappellini¹, Frido Welker^{2,3}, Luca Pandolfi⁴, Jazmin Ramos Madrigal², Anna K. Fotakis², David Lyon⁵, Victor J. Moreno Mayar¹, Maia Bukhsianidze⁶, Rosa Rakownikow Jersie-Christensen⁵, Meaghan Mackie^{2,5}, Aurélien Ginolhac⁷, Reid Ferring⁸, Martha Tappen⁹, Eleftheria Palkopoulou¹⁰, Diana Samodova⁵, Patrick L. Rütter⁵, Marc R. Dickinson¹¹, Tom Stafford¹², Yvonne L. Chan¹³, Anders Götherström¹⁴, Senthilvel KSS Nathan¹⁵, Peter D. Heintzman¹⁶, Joshua D. Kapp¹⁶, Irina Kirillova¹⁷, Yoshan Moodley¹⁸, Jordi Agustí^{19,20}, Ralf-Dietrich Kahlke²¹, Gocha Kiladze⁶, Bienvenido Martínez–Navarro^{19,20,22}, Shanlin Liu^{1,23}, Marcela Sandoval Velasco², Mikkel-Holger S. Sinding^{1,24}, Christian D. Kelstrup⁵, Morten E. Allentoft¹, Anders Krogh²⁵, Ludovic Orlando^{26,1}, Kirsty Penkman¹¹, Beth Shapiro¹⁶, Lorenzo Rook⁴, Love Dalén¹³, M. Thomas P. Gilbert^{1,27}, Jesper V. Olsen⁵, David Lordkipanidze⁶, Eske Willerslev^{1,28,29}

¹ Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Denmark.

² Natural History Museum of Denmark, University of Copenhagen, Denmark.

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

⁴ Dipartimento di Scienze della Terra, University of 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 National Alliance of Shidlovskiy "Ice Age", Moscow, Russia.

¹⁸ Department of Zoology, University of Venda, South Africa.

¹⁹ Institut Català de Paleoecologia Humana i Evolució Social, Universitat Rovira i Virgili, Spain.

²⁰ Institució Catalana de Recerca i Estudis Avançats (ICREA)

²¹ Department of Quaternary Palaeontology, Senckenberg Research Institute, Germany.

²² Departament d'Història i Geografia, Universitat Rovira i Virgili, Spain.

²³ BGI Shenzhen, Shenzhen, China

²⁴ Greenland Institute of Natural Resources, Nuuk, Greenland.

²⁵ The Bioinformatics Centre, Department of Biology, University of Copenhagen, Denmark.

²⁶ Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, Université de Toulouse, Université Paul Sabatier, France.

45 ²⁷ University Museum, Norwegian University of Science and Technology, Norway.

46 ²⁸ Department of Zoology, University of Cambridge, UK.

47 ²⁹ Wellcome Trust Sanger Institute, Hinxton, UK.

48

49

50 Correspondence and requests for material should be addressed to E.C. (ecappellini@snm.ku.dk) or

51 E.W. (ewillerslev@snm.ku.dk).

52

53 ABSTRACT

54

55 Ancient DNA (aDNA) sequencing has enabled unprecedented reconstruction of speciation,
56 migration, and admixture events for extinct taxa¹. Outside the permafrost, however, irreversible
57 aDNA post-mortem degradation² has so far limited aDNA recovery within the ~0.5 million years (Ma)
58 time range³. Tandem mass spectrometry (MS)-based collagen type I (COL1) sequencing provides
59 direct access to older biomolecular information⁴, though with limited phylogenetic use. In the
60 absence of molecular evidence, the speciation of several Early and Middle Pleistocene extinct
61 species remain contentious. In this study, we address the phylogenetic relationships of the Eurasian
62 Pleistocene Rhinocerotidae⁵⁻⁷ using ~1.77 million years (Ma) old dental enamel proteome sequences
63 of a *Stephanorhinus* specimen from the Dmanisi archaeological site in Georgia (South Caucasus)⁸.
64 Molecular phylogenetic analyses place the Dmanisi *Stephanorhinus* as a sister group to the woolly
65 (*Coelodonta antiquitatis*) and Merck's rhinoceros (*S. kirchbergensis*) clade. We show that
66 *Coelodonta* evolved from an early *Stephanorhinus* lineage and that this genus includes at least two
67 distinct evolutionary lines. As such, the genus *Stephanorhinus* is currently paraphyletic and its
68 systematic revision is therefore needed. We demonstrate that Early Pleistocene dental enamel
69 proteome sequencing overcomes the limits of ancient collagen- and aDNA-based phylogenetic
70 inference, and also provides additional information about the sex and taxonomic assignment of the
71 specimens analysed. Dental enamel, the hardest tissue in vertebrates, is highly abundant in the fossil
72 record. Our findings reveal that palaeoproteomic investigation of this material can push
73 biomolecular investigation further back into the Early Pleistocene.

74

75

76 MAIN TEXT

77

78 Phylogenetic placement of extinct species increasingly relies on aDNA sequencing. Relentless
79 efforts to improve the molecular tools underlying aDNA recovery have enabled the reconstruction
80 of ~0.4 Ma and ~0.7 Ma old DNA sequences from temperate deposits⁹ and subpolar regions¹⁰
81 respectively. However, no aDNA data have so far been generated from species that became
82 extinct beyond this time range. In contrast, ancient proteins represent a more durable source of
83 genetic information, reported to survive, in eggshell, up to 3.8 Ma¹¹. Ancient protein sequences
84 can carry taxonomic and phylogenetic information useful to trace the evolutionary relationships
85 between extant and extinct species^{12,13}. However, so far, the recovery of ancient mammal proteins
86 from sites too old or too warm to be compatible with aDNA preservation is mostly limited to
87 collagen type I (COL1). Being highly conserved¹⁴, this protein is not an ideal marker. For example,
88 regardless of endogeneity¹⁵, collagen-based phylogenetic placement of Dinosauria in relation to
89 extant Aves appears to be unstable¹⁶. This suggests the exclusive use of COL1 in deep-time
90 phylogenetics is constraining. Here, we aimed at overcoming these limitations by testing whether
91 dental enamel, the hardest tissue in vertebrates¹⁷, can better preserve a richer set of ancient
92 protein residues. This material, very abundant in the fossil record, would provide unprecedented
93 access to biomolecular and phylogenetic data from Early Pleistocene animal remains.

94 Dated to ~1.77 Ma by a combination of Ar/Ar dating, paleomagnetism and biozonation^{18,19},
95 the archaeological site of Dmanisi (Georgia, South Caucasus; Fig 1a) represents a context currently
96 considered outside the scope of aDNA recovery. This site has been excavated since 1983, resulting
97 in the discovery, along with stone tools and contemporaneous fauna, of almost one hundred
98 hominin fossils, including five skulls representing the *georgicus* paleodeme within *Homo erectus*⁸.
99 These are the earliest fossils of the first *Homo* species leaving Africa.

100 The geology of the Dmanisi deposits provides an ideal context for the preservation of
101 faunal materials. The primary deposits at Dmanisi are aeolian, providing for rapid, gentle burial in
102 fine-grained, calcareous sediments. We collected 23 bone, dentine, and dental enamel specimens
103 of large mammals (Tab. 1) from multiple excavation units within stratum B1 (Fig. 1b, Fig. 2, Tab. 1).
104 This is an ashfall deposit that contains thousands of faunal remains, as well as all hominin fossils,
105 in different geomorphic contexts including pipes, shallow gullies and carnivore dens. All of these

106 are firmly dated between 1.85-1.76 Ma¹⁸. High-resolution tandem MS was used to confidently
107 sequence ancient protein residues from the set of faunal remains, after digestion-free
108 demineralisation in acid (see Methods). Ancient DNA analysis was unsuccessfully attempted on a
109 subset of five bone and dentine specimens (see Methods).

110 While the recovery of proteins from bone and dentine specimens was sporadic and limited
111 to collagen fragments, the analysis of dental enamel consistently returned sequences from most
112 of its proteome, with occasional detection of multiple isoforms of the same protein²⁰ (Tab. 2, Fig.
113 3). The small proteome²¹ of mature dental enamel consists of structural enamel proteins, i.e.
114 amelogenin (*AMELX*), enamelin (*ENAM*), amelotin (*AMTN*), and ameloblastin (*AMBN*), and enamel-
115 specific proteases secreted during amelogenesis, i.e. matrix metalloproteinase-20 (*MMP20*) and
116 kallikrein 4 (*KLK4*). The presence of non-specific proteins, such as serum albumin (*ALB*), has also
117 been previously reported in mature dental enamel^{21,22} (Tab. 2).

118 Multiple lines of evidence support the authenticity and the endogenous origin of the
119 sequences recovered. There is full correspondence between the source material and the
120 composition of the proteome recovered. Dental enamel proteins are extremely tissue-specific and
121 confined to the dental enamel mineral matrix²¹. The amino acid composition of the intra-
122 crystalline protein fraction, measured by chiral amino acid racemisation analysis, indicates that the
123 dental enamel has behaved as a closed system, unaffected by amino acid and protein residues
124 exchange with the burial environment (Fig. 4). The measured rate of asparagine and glutamine
125 deamidation, a spontaneous form of hydrolytic damage consistently observed in ancient
126 samples²³, is particularly high, in some cases close to 100%, in full agreement with the age of the
127 specimens investigated. (Fig. 2a). Other forms of non-enzymatic modifications are also present.
128 Tyrosine (Y) experienced mono- and di-oxidation while tryptophan (W) was extensively converted
129 into multiple oxidation products. (Fig. 5b). Oxidative degradation of histidine (H) and conversion of
130 arginine (R) leading to ornithine accumulation were also observed. These modifications are
131 absent, or much less frequent, in a medieval ovicaprine dental enamel control sample, further
132 confirming the authenticity of the sequences reconstructed. Similarly, unlike in the control, the
133 peptide length distribution in the Dmanisi dataset is dominated by short overlapping fragments,
134 generated by advanced, diagenetically-induced, terminal hydrolysis (Fig. 5c and d).

135 Lastly, we confidently detect phosphorylation (Fig. 6 and Fig. 7), a tightly regulated
136 physiological post-translational modification (PTM) occurring *in vivo*. Recently observed in ancient
137 bone²⁴, phosphorylation is known to be a stable PTM²⁵ present in dental enamel proteins^{26,27}.
138 Altogether, these observations demonstrate, beyond reasonable doubt, that the heavily
139 diagenetically modified dental enamel proteome retrieved from the ~1.77 Ma old Dmanisi faunal
140 material is endogenous and almost complete.

141 Next, we used the palaeoproteomic sequence information to improve taxonomic
142 assignment and achieve sex attribution for some of the Dmanisi faunal remains. For example, the
143 bone specimen 16857, described morphologically as an “undetermined herbivore”, could be
144 assigned to the Bovidae family based on COL1 sequences (Fig. 8). In addition, confident
145 identification of peptides specific for the isoform Y of amelogenin, coded on the non-recombinant
146 portion of the Y chromosome, indicates that four tooth specimens, namely 16630, 16631, 16639,
147 and 16856, belonged to male individuals²² (Fig. 9a-d).

148 An enamel fragment, from the lower molar of a *Stephanorhinus* ex gr. *etruscus-*
149 *hundsheimensis* (16635, Fig. 1c), returned the highest proteomic sequence coverage,
150 encompassing a total of 875 amino acids, across 987 peptides (6 proteins). Following alignment of
151 the enamel protein sequences retrieved from 16635 against their homologues from all the extant
152 rhinoceros species, plus the extinct woolly rhinoceros (*†Coelodonta antiquitatis*) and Merck’s
153 rhinoceros (*†Stephanorhinus kirchbergensis*), phylogenetic reconstructions place the Dmanisi
154 specimen closer to the extinct woolly and Merck’s rhinoceroses than to the extant Sumatran
155 rhinoceros (*Dicerorhinus sumatrensis*), as an early divergent sister lineage (Fig. 10).

156 Our phylogenetic reconstruction confidently recovers the expected differentiation of the
157 *Rhinoceros* genus from other genera considered, in agreement with previous cladistic²⁸ and
158 genetic analyses²⁹. This topology defines two-horned rhinoceroses as monophyletic and the one-
159 horned condition as plesiomorphic, as previously proposed³⁰. We caution, however, that the
160 higher-level relationships we observe between the rhinoceros monophyletic clades might be
161 affected by demographic events, such as incomplete lineage sorting³¹ and/or gene flow between
162 groups³², due to the limited number of markers considered. A previous phylogenetic
163 reconstruction, based on two collagen (*COL1α1* and *COL1α2*) partial amino acid sequences,
164 supported a different topology, with the African clade representing an outgroup to Asian

165 rhinoceros species⁶. Most probably, a confident and stable reconstruction of the structure of the
166 Rhinocerotidae family needs the strong support only high-resolution whole-genome sequencing
167 can provide. Regardless, the highly supported placement of the Dmanisi rhinoceros in the
168 (*Stephanorhinus*, Woolly, Sumatran) clade will likely remain unaffected, should deeper
169 phylogenetic relationships between the *Rhinoceros* genus and other family members be revised.

170 The phylogenetic relationships of the genus *Stephanorhinus* within the family
171 Rhinocerotidae, as well as those of the several species recognized within this genus, are
172 contentious. *Stephanorhinus* was initially included in the extant South-East Asian genus
173 *Dicerorhinus* represented by the Sumatran rhinoceros species (*D. sumatrensis*)³³. This hypothesis
174 has been rejected and, based on morphological data, *Stephanorhinus* has been identified as a
175 sister taxon of the woolly rhinoceros³⁴. Furthermore, ancient DNA analysis supports a sister
176 relationship between the woolly rhinoceros and *D. sumatrensis*^{5,35,36}.
177 Recently, MS-based sequencing of collagen type I from a Middle Pleistocene European
178 *Stephanorhinus* sp. specimen, ~320 ka (thousand years) old, was not able to resolve the
179 relationships between *Stephanorhinus*, *Coelodonta* and *Dicerorhinus*⁶. Instead, the complete
180 mitochondrial sequence of a terminal, 45-70 ka old, Siberian *S. kirchbergensis* specimen placed
181 this species closer to *Coelodonta*, with *D. sumatrensis* as a sister branch⁷. Our results confirm the
182 latter reconstruction. As the *Stephanorhinus* ex gr. *etruscus-hundsheimensis* sequences from
183 Dmanisi branch off basal to the common ancestor of the woolly and Merck's rhinoceroses, these
184 two species most likely derived from an early *Stephanorhinus* lineage expanding eastward from
185 western Eurasia. Throughout the Plio-Pleistocene, *Coelodonta* adapted to continental and later
186 cold-climate habitats in central Asia. Its earliest representative, *C. thibetana*, displayed some clear
187 *Stephanorhinus*-like anatomical features³⁴. The presence in eastern Europe and Anatolia of the
188 genus *Stephanorhinus*³⁵ is documented at least since the late Miocene, and the Dmanisi specimen
189 most likely represents an Early Pleistocene descendent of the Western-Eurasian branch of this
190 genus.

191 Ultimately, our phylogenetic reconstructions show that, as currently defined, the genus
192 *Stephanorhinus* is paraphyletic, in line with previous conclusions³⁷ based on morphological
193 characters and the palaeobiogeographic fossil distribution. Accordingly, a systematic revision of
194 the genera *Stephanorhinus* and *Coelodonta*, as well as their closest relatives, is needed.

195 In this study, we show that enamel proteome sequencing can overcome the time limits of
196 ancient DNA preservation and the reduced phylogenetic content of COL1 sequences. Dental
197 enamel proteomic sequences can be used to study evolutionary process that occurred in the Early
198 Pleistocene. This posits dental enamel as the material of choice for deep-time palaeoproteomic
199 analysis. Given the abundance of teeth in the palaeontological record, the approach presented
200 here holds the potential to address a wide range of questions pertaining to the Early and Middle
201 Pleistocene evolutionary history of a large number of mammals, including hominins, at least in
202 temperate climates.
203

204 **METHODS**

205

206 **Dmanisi & sample selection**

207 Dmanisi is located about 65 km southwest of the capital city of Tbilisi in the Kvemo Kartli region of
208 Georgia, at an elevation of 910 m MSL (Lat: 41° 20' N, Lon: 44° 20' E)^{8,19}. The 23 fossil specimens
209 we analysed were retrieved from stratum B1, in excavation blocks M17, M6, block 2, and area R11
210 (Tab. 1 and Fig. 2). Stratum B deposits date between 1.78 Ma and 1.76 Ma¹⁸. All the analysed
211 specimens were collected between 1984 and 2014 and their taxonomic identification was based
212 on traditional comparative anatomy.

213 After the sample preparation and data acquisition for all the Dmanisi specimens was
214 concluded, we applied the whole experimental procedure to a medieval ovicaprine (sheep/goat)
215 dental enamel specimen that was used as control. For this sample, we used extraction protocol
216 “C”, and generated tandem MS data using a Q Exactive HF mass spectrometer (Thermo Fisher
217 Scientific). The data were searched against the goat proteome, downloaded from the NCBI
218 Reference Sequence Database (RefSeq) archive³⁸ on 31st May 2017. The ovicaprine specimen was
219 found at the “Hotel Skandinavia” site in the city of Århus, Denmark and was stored at the Natural
220 History Museum of Denmark.

221

222 **Biomolecular preservation**

223 We assessed the potential of ancient protein preservation prior to proteomic analysis by
224 measuring the extent of amino acid racemisation in a subset of samples (6/23)³⁹. Enamel chips
225 were powdered, and two subsamples per specimen were subject to analysis of their free (FAA)
226 and total hydrolysable (THAA) amino acid fractions. Samples were analysed in duplicate by RP-
227 HPLC, with standards and blanks run alongside each one of them. The D/L values of aspartic
228 acid/asparagine, glutamic acid/glutamine, phenylalanine and alanine (D/L Asx, Glx, Phe, Ala) were
229 assessed (Fig. 4) to provide an overall estimate of intra-crystalline protein decomposition (IcPD).

230

231 **PROTEOMICS**

232 All the sample preparation procedures for palaeoproteomic analysis were conducted in
233 laboratories dedicated to the analysis of ancient DNA and ancient proteins in clean rooms fitted

234 with filtered ventilation and positive pressure, in line with recent recommendations for ancient
235 protein analysis⁴⁰. A mock “extraction blank”, containing no starting material, was prepared,
236 processed and analysed together with each batch of ancient samples.

237

238 **Sample preparation**

239 The external surface of bone and dentine samples was gently removed, and the remaining
240 material was subsequently powdered. Enamel fragments, occasionally mixed with small amounts
241 of dentine, were removed from teeth with a cutting disc and subsequently crushed into a rough
242 powder. Ancient protein residues were extracted from approximately 180-220 mg of mineralised
243 material, unless otherwise specified, using three different extraction protocols, hereafter referred
244 to as “A”, “B” and “C”:

245

246 **EXTRACTION PROTOCOL A - FASP.** Tryptic peptides were generated using a filter-aided sample
247 preparation (FASP) approach⁴¹, as previously performed on ancient samples⁴².

248

249 **EXTRACTION PROTOCOL B - GuHCl SOLUTION AND DIGESTION.** Bone or dentine powder was demineralised
250 in 1 mL 0.5 M EDTA pH 8.0. After removal of the supernatant, all demineralised pellets were re-
251 suspended in a 300 µL solution containing 2 M guanidine hydrochloride (GuHCl, Thermo
252 Scientific), 100 mM Tris pH 8.0, 20 mM 2-Chloroacetamide (CAA), 10 mM Tris (2-
253 carboxyethyl)phosphine (TCEP) in ultrapure H₂O^{43,44}. A total of 0.2 µg of mass spectrometry-grade
254 rLysC (Promega P/N V1671) enzyme was added before the samples were incubated for 3-4 hours
255 at 37°C with agitation. Samples and negative controls were subsequently diluted to 0.6 M GuHCl,
256 and 0.8 µg of mass spectrometry-grade Trypsin (Promega P/N V5111) was added. The entire
257 amount of extracted proteins was digested. Next, samples and negative controls were incubated
258 overnight under mechanical agitation at 37°C. On the following day, samples were acidified, and
259 the tryptic peptides were immobilised on Stage-Tips, as previously described⁴⁵.

260

261 **EXTRACTION PROTOCOL C - DIGESTION-FREE ACID DEMINERALISATION.** Dental enamel powder was
262 demineralised in 1.2 M HCl at room temperature, after which the solubilised protein residues were
263 directly cleaned and concentrated on Stage-Tips, as described above. The sample prepared on

264 Stage-Tip “#1217” was processed with 10% TFA instead of 1.2 M HCl. All the other parameters and
265 procedures were identical to those used for all the other samples extracted with protocol “C”.

266

267 **Tandem mass spectrometry**

268 Different sets of samples were analysed by nanoflow liquid chromatography coupled to tandem
269 mass spectrometry (nanoLC-MS/MS) on an EASY-nLC™ 1000 or 1200 system connected to a Q-
270 Exactive, a Q-Exactive Plus, or to a Q-Exactive HF (Thermo Scientific, Bremen, Germany) mass
271 spectrometer. Before and after each MS/MS run measuring ancient or extraction blank samples,
272 two successive MS/MS run were included in the sample queue in order to prevent carryover
273 contamination between the samples. These consisted, first, of a MS/MS run (“MS/MS blank” run)
274 with an injection exclusively of the buffer used to re-suspend the samples (0.1% TFA, 5% ACN),
275 followed by a second MS/MS run (“MS/MS wash” run) with no injection.

276

277 **Data analysis**

278 Raw data files generated during MS/MS spectral acquisition were searched using MaxQuant⁴⁶,
279 version 1.5.3.30, and PEAKS⁴⁷, version 7.5. A two-stage peptide-spectrum matching approach was
280 adopted. Raw files were initially searched against a target/reverse database of collagen and
281 enamel proteins retrieved from the UniProt and NCBI Reference Sequence Database (RefSeq)
282 archives^{38,48}, taxonomically restricted to mammalian species. A database of partial “COL1A1” and
283 “COL1A2” sequences from cervid species¹³ was also included. The results from the preliminary
284 analysis were used for a first, provisional reconstruction of protein sequences.

285 For specimens whose dataset resulted in a narrower, though not fully resolved, initial
286 taxonomic placement, a second MaxQuant search (MQ2) was performed using a new protein
287 database taxonomically restricted to the “order” taxonomic rank as determined after MQ1. For
288 the MQ2 matching of the MS/MS spectra from specimen 16635, partial sequences of serum
289 albumin and enamel proteins from Sumatran (*Dicerorhinus sumatrensis*), Javan (*Rhinoceros*
290 *sondaicus*), Indian (*Rhinoceros unicornis*), woolly (*Coelodonta antiquitatis*), Mercks
291 (*Stephanorhinus kirchbergensis*), and Black rhinoceros (*Diceros bicornis*), were also added to the
292 protein database. All the protein sequences from these species were reconstructed from draft
293 genomes for each species (Dalen and Gilbert, unpublished data).

294 For each MaxQuant and PEAKS search, enzymatic digestion was set to “unspecific” and the
295 following variable modifications were included: oxidation (M), deamidation (NQ), N-term Pyro-Glu
296 (Q), N-term Pyro-Glu (E), hydroxylation (P), phosphorylation (S). The error tolerance was set to 5
297 ppm for the precursor and to 20 ppm, or 0.05 Da, for the fragment ions in MaxQuant and PEAKS
298 respectively. For searches of data generated from sample fractions partially or exclusively digested
299 with trypsin, another MaxQuant and PEAKS search was conducted using the “enzyme” parameter
300 set to “Trypsin/P”. Carbamidomethylation (C) was set: (i) as a fixed modification, for searches of
301 data generated from sets of sample fractions exclusively digested with trypsin, or (ii) as a variable
302 modification, for searches of data generated from sets of sample fractions partially digested with
303 trypsin. For searches of data generated exclusively from undigested sample fractions,
304 carbamidomethylation (C) was not included as a modification, neither fixed nor variable.

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

311

312 **Ancient protein sequence reconstruction**

313 The peptide sequences confidently identified by the MQ1, MQ2, MQ3 were aligned using the
314 software Geneious⁴⁹ (v. 5.4.4, substitution matrix BLOSUM62, gap open penalty 12 and gap
315 extension penalty). The peptide sequences confidently identified by the PEAKS searches were
316 aligned using an in-house R-script. A consensus sequence for each protein from each specimen
317 was generated in FASTA format, without filtering on depth of coverage. Amino acid positions that
318 were not confidently reconstructed were replaced by an “X”. We took into account variable
319 leucine/isoleucine, glutamine/glutamic acid, and asparagine/aspartic acid positions through
320 manual interpretation of possible conflicting positions (leucine/isoleucine) and replacement of
321 possibly deamidated positions into “X” for phylogenetically informative sites. The output of the
322 MQ2 and 3 peptide-spectrum matching was used to extend the coverage of the ancient protein
323 sequences initially identified in the MQ1 iteration.

324

325 **Post translational modifications**

326 **DEAMIDATION.** After removal of likely contaminants, the extent of glutamine and asparagine
327 deamidation was estimated for individual specimens, by using the MaxQuant output files as
328 previously published⁴⁴.

329 **OTHER SPONTANEOUS CHEMICAL MODIFICATIONS.** Spontaneous post-translational modifications (PTMs)
330 associated with chemical protein damage were searched using the PEAKS PTM tool and the
331 dependent peptides search mode⁵⁰ in MaxQuant. In the PEAKS PTM search, all modifications in
332 the Unimod database were considered. The mass error was set to 5.0 ppm and 0.5 Da for
333 precursor and fragment, respectively. For PEAKS, the *de novo* ALC score was set to a threshold of
334 15 % and the peptide hit threshold to 30. The results were filtered by an FDR of 5 %, *de novo* ALC
335 score of 50 %, and a protein hit threshold of ≥ 20 . The MaxQuant dependent peptides search was
336 carried out with the same search settings as described above and with a dependent peptide FDR
337 of 1 % and a mass bin size of 0.0065 Da. For validation purposes, up to 10 discovered modifications
338 were specified as variable modifications and re-searched with MaxQuant. The peptide FDR was
339 manually adjusted to 5 % on PSM level and the PTMs were semi-quantified by relative spectral
340 counting.

341 **PHOSPHORYLATION.** Class I phosphorylation sites were selected with localisation probabilities of
342 ≥ 0.98 in the Phosph(ST)Sites MaxQuant output file. Sequence windows of ± 6 aa from all identified
343 sites were compared against a background file containing all non-phosphorylated peptides using a
344 linear kinase sequence motif enrichment analysis in IceLogo⁵¹.

345

346 **PHYLOGENETIC ANALYSIS**

347 **Reference datasets**

348 We assembled a reference dataset consisting of publicly available protein sequences from
349 representative ungulate species belonging to the following families: Equidae, Rhinocerotidae,
350 Suidae and Bovidae. We extended this dataset with the protein sequences from extinct and extant
351 rhinoceros species including: the woolly rhinoceros (*†Coelodonta antiquitatis*), the Merck's
352 rhinoceros (*†Stephanorhinus kirchbergensis*), the Sumatran rhinoceros (*Dicerorhinus sumatrensis*),
353 the Javan rhinoceros (*Rhinoceros sondaicus*), the Indian rhinoceros (*Rhinoceros unicornis*), and the

354 Black rhinoceros (*Diceros bicornis*). Their corresponding protein sequences were obtained
355 following translation of high-throughput DNA sequencing data, after filtering reads with mapping
356 quality lower than 30 and nucleotides with base quality lower than 20, and calling the majority
357 rule consensus sequence using ANGSD⁵². For the woolly and Merck's rhinoceroses we excluded the
358 first and last five nucleotides of each DNA fragment in order to minimize the effect of post-
359 mortem ancient DNA damage⁵³. Each consensus sequence was formatted as a separate blast
360 nucleotide database. We then performed a tblastn⁵⁴ alignment using the corresponding white
361 rhinoceros sequence as a query, favouring ungapped alignments in order to recover translated
362 and spliced protein sequences. Resulting alignments were processed using ProSplign algorithm
363 from the NCBI Eukaryotic Genome Annotation Pipeline⁵⁵ to recover the spliced alignments and
364 translated protein sequences.

365

366 **Construction of phylogenetic trees**

367 For each specimen, multiple sequence alignments for each protein were built using mafft⁵⁶ and
368 concatenated onto a single alignment per specimen. These were inspected visually to correct
369 obvious alignment mistakes, and all the isoleucine residues were substituted with leucine ones to
370 account for indistinguishable isobaric amino acids at the positions where the ancient protein
371 carried one of such amino acids. Based on these alignments, we inferred the phylogenetic
372 relationship between the ancient samples and the species included in the reference dataset by
373 using three approaches: distance-based neighbour-joining, maximum likelihood and Bayesian
374 phylogenetic inference.

375 Neighbour-joining trees were built using the phangorn⁵⁷ R package, restricting to sites
376 covered in the ancient samples. Genetic distances were estimated using the JTT model,
377 considering pairwise deletions. We estimated bipartition support through a non-parametric
378 bootstrap procedure using 500 pseudoreplicates. We used PHyML 3.1⁵⁸ for maximum likelihood
379 inference based on the whole concatenated alignment. For likelihood computation, we used the
380 JTT substitution model with two additional parameters for modelling rate heterogeneity and the
381 proportion of invariant sites. Bipartition support was estimated using a non-parametric bootstrap
382 procedure with 500 replicates. Bayesian phylogenetic inference was carried out using MrBayes
383 3.2.6⁵⁹ on each concatenated alignment, partitioned per gene. While we chose the JTT

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

391

392 **ANCIENT DNA ANALYSIS**

393 The samples were processed using strict aDNA guidelines in a clean lab facility at the Centre for
394 GeoGenetics, Natural History Museum of Denmark, University of Copenhagen. DNA extraction was
395 attempted on five of the ancient animal samples. Powdered samples (120-140 mg) were extracted
396 using a silica-in-solution method^{10,60}. To prepare the samples for NGS sequencing, 20 μ L of DNA
397 extract was built into a blunt-end library using the NEBNext DNA Sample Prep Master Mix Set 2
398 (E6070) with Illumina-specific adapters. The libraries were PCR-amplified with inPE1.0 forward
399 primers and custom-designed reverse primers with a 6-nucleotide index⁶¹. Two extracts (MA399
400 and MA2481, from specimens 16859 and 16635 respectively) yielded detectable DNA
401 concentrations. These extracts were used to construct three individual index-barcoded libraries
402 (MA399_L1, MA399_L2, MA2481_L1) whose amplification required a total of 30 PCR cycles in a 2-
403 round setup (12 cycles with total library + 18 cycles with a 5 μ L library aliquot from the first
404 amplification). The libraries generated from specimen 16859 and 16635 were processed on
405 different flow cells. They were pooled with others for sequencing on an Illumina 2000 platform
406 (MA399_L1, MA399_L2) using 100bp single read chemistry and on an Illumina 2500 platform
407 (MA2481_L1) using 81bp single read chemistry.

408 The data were base-called using the Illumina software CASAVA 1.8.2 and sequences were
409 demultiplexed with a requirement of a full match of the six nucleotide indexes that were used.
410 Raw reads were processed using the PALEOMIX pipeline following published guidelines⁶², mapping
411 against the cow nuclear genome (*Bos taurus* 4.6.1, accession GCA_000003205.4), the cow
412 mitochondrial genome (*Bos taurus*), the red deer mitochondrial genome (*Cervus elaphus*,
413 accession AB245427.2), and the human nuclear genome (GRCh37/hg19), using BWA backtrack⁶³

414 v0.5.10 with the seed disabled. All other parameters were set as default. PCR duplicates from
415 mapped reads were removed using the picard tool *MarkDuplicate*
416 [<http://picard.sourceforge.net/>].

417

418 **SAMPLE 16635 MORPHOLOGICAL MEASUREMENTS**

419 We followed the methodology introduced by Guérin³³. The maximal length of the tooth is
420 measured with a digital calliper at the lingual side of the tooth and parallel to the occlusal surface.
421 All measurements are given in mm.

422

423 **DATA DEPOSITION**

424 All the mass spectrometry proteomics data have been deposited in the ProteomeXchange
425 Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with
426 the data set identifier PXD011008.

427

428

429 **References**

430

- 431 1 Cappellini, E. *et al.* Ancient Biomolecules and Evolutionary Inference. *Annual Review of*
432 *Biochemistry* **87**, 1029-1060, doi:10.1146/annurev-biochem-062917-012002 (2018).
- 433 2 Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harbor Perspectives in*
434 *Biology* **5**, a012567, doi:10.1101/cshperspect.a012567 (2013).
- 435 3 Meyer, M. *et al.* Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos
436 hominins. *Nature* **531**, 504-507, doi:10.1038/nature17405 (2016).
- 437 4 Wadsworth, C. & Buckley, M. Proteome degradation in fossils: investigating the longevity
438 of protein survival in ancient bone. *Rapid Communications in Mass Spectrometry* **28**, 605-
439 615, doi:10.1002/rcm.6821 (2014).
- 440 5 Willerslev, E. *et al.* Analysis of complete mitochondrial genomes from extinct and extant
441 rhinoceroses reveals lack of phylogenetic resolution. *BMC Evolutionary Biology* **9**, 95,
442 doi:10.1186/1471-2148-9-95 (2009).
- 443 6 Welker, F. *et al.* Middle Pleistocene protein sequences from the rhinoceros genus and the
444 phylogeny of extant and extinct Middle/Late Pleistocene Rhinocerotidae. *PeerJ* **5**, e3033,
445 doi:10.7717/peerj.3033 (2017).
- 446 7 Kirillova, I. *et al.* Discovery of the skull of *Stephanorhinus kirchbergensis* (Jäger, 1839)
447 above the Arctic Circle. *Quaternary Research* **88**, 537-550, doi:10.1017/qua.2017.53 (2017).
- 448 8 Lordkipanidze, D. *et al.* A complete skull from Dmanisi, Georgia, and the evolutionary
449 biology of early Homo. *Science* **342**, 326-331, doi:10.1126/science.1238484 (2013).
- 450 9 Valdiosera, C. *et al.* Typing single polymorphic nucleotides in mitochondrial DNA as a way
451 to access Middle Pleistocene DNA. *Biology Letters* **2**, 601-603, doi:10.1098/rsbl.2006.0515
452 (2006).

- 453 10 Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early
454 Middle Pleistocene horse. *Nature* **499**, 74-78, doi:10.1038/nature12323 (2013).
- 455 11 Demarchi, B. *et al.* Protein sequences bound to mineral surfaces persist into deep time.
456 *eLife* **5**, e17092, doi:10.7554/eLife.17092 (2016).
- 457 12 Welker, F. *et al.* Ancient proteins resolve the evolutionary history of Darwin's South
458 American ungulates. *Nature* **522**, 81-84, doi:10.1038/nature14249 (2015).
- 459 13 Welker, F. *et al.* Palaeoproteomic evidence identifies archaic hominins associated with the
460 Châtelperronian at the Grotte du Renne. *Proceedings of the National Academy of Sciences*
461 **113**, 11162-11167, doi:10.1073/pnas.1605834113 (2016).
- 462 14 Nei, M. *Molecular evolutionary genetics*. Vol. 75 (Columbia University Press, 1987).
- 463 15 Buckley, M., Warwood, S., van Dongen, B., Kitchener, A. C. & Manning, P. L. A fossil protein
464 chimera; difficulties in discriminating dinosaur peptide sequences from modern cross-
465 contamination. *Proceedings of the Royal Society: Biological sciences* **284**, 20170544,
466 doi:10.1098/rspb.2017.0544 (2017).
- 467 16 Schroeter, E. R. *et al.* Expansion for the Brachylophosaurus canadensis Collagen I Sequence
468 and Additional Evidence of the Preservation of Cretaceous Protein. *Journal of Proteome*
469 *Research* **16**, 920-932, doi:10.1021/acs.jproteome.6b00873 (2017).
- 470 17 Eastoe, J. E. Organic Matrix of Tooth Enamel. *Nature* **187**, 411-412, doi:10.1038/187411b0
471 (1960).
- 472 18 Ferring, R. *et al.* Earliest human occupations at Dmanisi (Georgian Caucasus) dated to 1.85-
473 1.78 Ma. *Proceedings of the National Academy of Sciences of the United States of America*
474 **108**, 10432-10436, doi:10.1073/pnas.1106638108 (2011).
- 475 19 Gabunia, L. *et al.* Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of
476 Georgia: taxonomy, geological setting, and age. *Science* **288**, 1019-1025,
477 doi:10.1126/science.288.5468.1019 (2000).
- 478 20 Gibson, C. W. *et al.* Identification of the leucine-rich amelogenin peptide (LRAP) as the
479 translation product of an alternatively spliced transcript. *Biochemical and biophysical*
480 *research communications* **174**, 1306, doi:10.1016/0006-291X(91)91564-S (1991).
- 481 21 Castiblanco, G. A. *et al.* Identification of proteins from human permanent erupted enamel.
482 *European Journal of Oral Sciences* **123**, 390-395, doi:10.1111/eos.12214 (2015).
- 483 22 Stewart, N. A. *et al.* The identification of peptides by nanoLC-MS/MS from human surface
484 tooth enamel following a simple acid etch extraction. *RSC Advances* **6**, 61673-61679,
485 doi:10.1039/c6ra05120k (2016).
- 486 23 van Doorn, N. L., Wilson, J., Hollund, H., Soressi, M. & Collins, M. J. Site-specific
487 deamidation of glutamine: a new marker of bone collagen deterioration. *Rapid*
488 *Communications in Mass Spectrometry* **26**, 2319-2327, doi:10.1002/rcm.6351 (2012).
- 489 24 Cleland, T. P. Solid Digestion of Demineralized Bone as a Method to Access Potentially
490 Insoluble Proteins and Post-Translational Modifications. *Journal of Proteome Research* **17**,
491 536-542, doi:10.1021/acs.jproteome.7b00670 (2018).
- 492 25 Hunter, T. Why nature chose phosphate to modify proteins. *Philosophical Transactions of*
493 *the Royal Society B* **367**, 2513-2516, doi:10.1098/rstb.2012.0013 (2012).
- 494 26 Tagliabracci, V. S. *et al.* Secreted kinase phosphorylates extracellular proteins that regulate
495 biomineralization. *Science* **336**, 1150-1153, doi:10.1126/science.1217817 (2012).

- 496 27 Lasa-Benito, M., Marin, O., Meggio, F. & Pinna, L. A. Golgi apparatus mammary gland
497 casein kinase: monitoring by a specific peptide substrate and definition of specificity
498 determinants. *FEBS Letters* **382**, 149-152, doi:10.1016/0014-5793(96)00136-6 (1996).
- 499 28 Antoine, P. O. *et al.* A revision of *Aceratherium blanfordi* Lydekker, 1884 (Mammalia:
500 Rhinocerotidae) from the Early Miocene of Pakistan: postcranials as a key. *Zoological*
501 *Journal of the Linnean Society* **160**, 139-194, doi:10.1111/j.1096-3642.2009.00597.x (2010).
- 502 29 Steiner, C. C. & Ryder, O. A. Molecular phylogeny and evolution of the Perissodactyla.
503 *Zoological Journal of the Linnean Society* **163**, 1289-1303, doi:10.1111/j.1096-
504 3642.2011.00752.x (2011).
- 505 30 Loose, H. Pleistocene Rhinocerotidae of W. Europe with reference to the recent two-
506 horned species of Africa and S. E. Asia. *Scripta Geologica* **33**, 1-59 (1975).
- 507 31 Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H. & Mailund, T. Incomplete lineage
508 sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan
509 speciation and widespread selection. *Genome research* **21**, 349-356,
510 doi:10.1101/gr.114751.110 (2011).
- 511 32 Rieseberg, L. H. Evolution: replacing genes and traits through hybridization. *Current Biology*
512 **19**, R119-R122, doi:10.1016/j.cub.2008.12.016 (2009).
- 513 33 Guérin, C. Les rhinocéros (Mammalia, Perissodactyla) du Miocène terminal au Pleistocène
514 supérieur en Europe occidentale, comparaison avec les espèces actuelles. *Documents du*
515 *Laboratoire de Géologie de la Faculté des Sciences de Lyon* **79**, 3-1183 (1980).
- 516 34 Deng, T. *et al.* Out of Tibet: pliocene woolly rhino suggests high-plateau origin of Ice Age
517 megaherbivores. *Science* **333**, 1285-1288, doi:10.1126/science.1206594 (2011).
- 518 35 Orlando, L. *et al.* Ancient DNA analysis reveals woolly rhino evolutionary relationships.
519 *Molecular Phylogenetics and Evolution* **28**, 485-499, doi:10.1016/S1055-7903(03)00023-X
520 (2003).
- 521 36 Yuan, J. *et al.* Ancient DNA sequences from *Coelodonta antiquitatis* in China reveal its
522 divergence and phylogeny. *Science China Earth Sciences* **57**, 388-396, doi:10.1007/s11430-
523 013-4702-6 (2014).
- 524 37 Heissig, K. in *The Miocene Land Mammals of Europe* (eds G.E. Rössner & K Heissig) 175-188
525 (Friedrich Pfeil, 1999).
- 526 38 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,
527 taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733-D745,
528 doi:10.1093/nar/gkv1189 (2016).
- 529 39 Penkman, K. E. H., Kaufman, D. S., Maddy, D. & Collins, M. J. Closed-system behaviour of
530 the intra-crystalline fraction of amino acids in mollusc shells. *Quaternary Geochronology* **3**,
531 2-25, doi:10.1016/j.quageo.2007.07.001 (2008).
- 532 40 Hendy, J. *et al.* A guide to ancient protein studies. *Nature Ecology & Evolution* **2**, 791-799,
533 doi:10.1038/s41559-018-0510-x (2018).
- 534 41 Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation
535 method for proteome analysis. *Nature Methods* **6**, 359-362, doi:10.1038/nmeth.1322
536 (2009).
- 537 42 Cappellini, E. *et al.* Resolution of the type material of the Asian elephant, *Elephas maximus*
538 Linnaeus, 1758 (Proboscidea, Elephantidae). *Zoological Journal of the Linnean Society* **170**,
539 222-232, doi:10.1111/zoj.12084 (2014).

- 540 43 Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated
541 proteomic-sample processing applied to copy-number estimation in eukaryotic cells.
542 *Nature Methods* **11**, 319-324, doi:10.1038/nmeth.2834 (2014).
- 543 44 Mackie, M. *et al.* Palaeoproteomic Profiling of Conservation Layers on a 14th Century
544 Italian Wall Painting. *Angewandte Chemie (International ed.)* **57**, 7369-7374,
545 doi:10.1002/anie.201713020 (2018).
- 546 45 Cappellini, E. *et al.* Proteomic analysis of a pleistocene mammoth femur reveals more than
547 one hundred ancient bone proteins. *Journal of Proteome Research* **11**, 917-926,
548 doi:10.1021/pr200721u (2012).
- 549 46 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized
550 p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature*
551 *Biotechnology* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).
- 552 47 Zhang, J. *et al.* PEAKS DB: De novo sequencing assisted database search for sensitive and
553 accurate peptide identification. *Molecular and Cellular Proteomics* **11**, M111.010587,
554 doi:10.1074/mcp.M111.010587 (2012).
- 555 48 The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids*
556 *Research* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).
- 557 49 Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform
558 for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649,
559 doi:10.1093/bioinformatics/bts199 (2012).
- 560 50 Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass
561 spectrometry-based shotgun proteomics. *Nature Protocols* **11**, 2301-2319,
562 doi:10.1038/nprot.2016.136 (2016).
- 563 51 Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved
564 visualization of protein consensus sequences by iceLogo. *Nature Methods* **6**, 786-787,
565 doi:10.1038/nmeth1109-786 (2009).
- 566 52 Korneliussen, T., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation
567 Sequencing Data. *BMC Bioinformatics* **15**, 356-356, doi:10.1186/s12859-014-0356-4 (2014).
- 568 53 Briggs, A. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in
569 ancient DNA. *Nucleic Acids Research* **38**, e87, doi:10.1093/nar/gkp1163 (2010).
- 570 54 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database
571 search programs. *Nucleic Acids Research* **25**, 3389-3402 (1997).
- 572 55 Sea Urchin Genome Sequencing Consortium. The Genome of the Sea Urchin
573 *Strongylocentrotus purpuratus*. *Science* **314**, 941-952 (2006).
- 574 56 Katoh, K. & Frith, M. C. Adding unaligned sequences into an existing alignment using
575 MAFFT and LAST. *Bioinformatics* **28**, 3144-3146, doi:10.1093/bioinformatics/bts578 (2012).
- 576 57 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593,
577 doi:10.1093/bioinformatics/btq706 (2011).
- 578 58 Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood
579 Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321,
580 doi:10.1093/sysbio/syq010 (2010).
- 581 59 Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice
582 Across a Large Model Space. *Systematic Biology* **61**, 539-542, doi:10.1093/sysbio/sys029
583 (2012).

- 584 60 Rohland, N. & Hofreiter, M. Comparison and optimization of ancient DNA extraction.
585 *BioTechniques* **42**, 343-352, doi:10.2144/000112383 (2007).
- 586 61 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed
587 target capture and sequencing. *Cold Spring Harbor Protocols*, doi:10.1101/pdb.prot5448
588 (2010).
- 589 62 Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP detection and
590 phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols* **9**, 1056-1082,
591 doi:10.1038/nprot.2014.063 (2014).
- 592 63 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows– Wheeler
593 transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
- 594

595

596 **ACKNOWLEDGEMENTS**

597

598 We would like to thank, Kristian Murphy Gregersen, for providing the medieval control specimen,
599 Marcus Anders Krag for the photographs used in Fig. 1c, Fedor Shidlovskiy for providing access to
600 the Merck’s rhino sample, Beatrice Triozzi for technical help, Ashot Margaryan and Shyam
601 Gopalakrishnan for their precious comments during data interpretation. EC and FW are supported
602 by VILLUM Fonden (grant number 17649). EC, CK, JVO, PR and DS are supported by the Marie
603 Skłodowska Curie European Training Network “TEMPERA” (grant number 722606). MM is
604 supported by the University of Copenhagen KU2016 (UCPH Excellence Programme) grant and by
605 the Danish National Research Foundation award PROTEIOS (DNRF128). Work at the Novo Nordisk
606 Foundation Center for Protein Research is funded in part by a generous donation from the Novo
607 Nordisk Foundation (Grant number NNF14CC0001). MTPG is supported by ERC Consolidator Grant
608 “EXTINCTION GENOMICS” (grant number 681396). LP was supported by the EU-SYNTHESYS project
609 (AT-TAF-2550, DE-TAF-3049, GB-TAF-2825, HU-TAF-3593, ES-TAF-2997) funded by the European
610 Commission. LO is supported by the ERC Consolidator Grant “PEGASUS” (grant agreement No
611 681605). BM-N is supported by the Spanish Ministry of Sciences (grant number CGL2016-80975-P).
612 BS, JK and PH are supported by the Gordon and Betty Moore foundation. The aDNA analysis was
613 carried out using the HPC facilities of the University of Luxembourg.

614

615

616 **AUTHOR CONTRIBUTIONS**

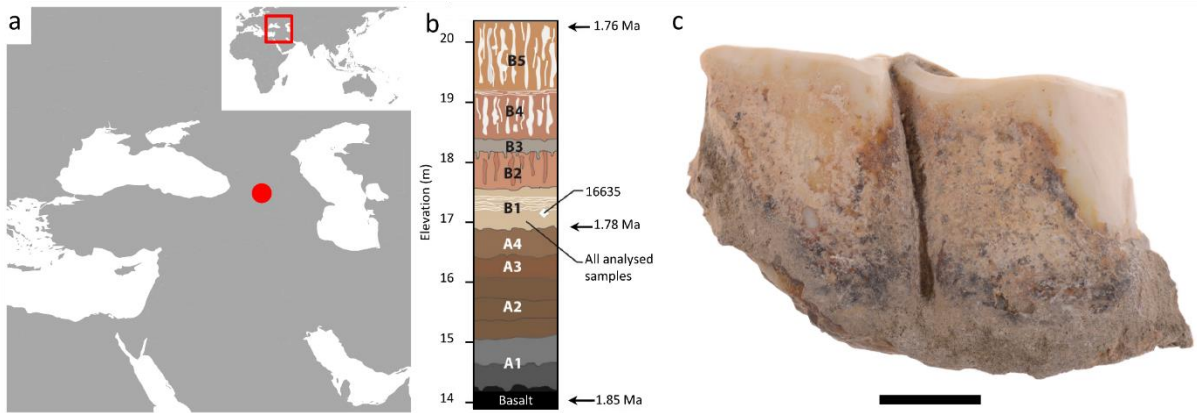
617

618 E.C., D.Lo., and E.W. designed the study. A.K.F., M.M., R.R.J.-C., M.E.A., M.D., K.P., and E.C.
619 performed laboratory experiments. M.Bu., M.T., R.F., E.P., T.S., Y.L.C., A.Gö., S.N., P.H., J.K., I.K.,
620 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
621 ancient samples or modern reference material. E.C., F.W., L.P., J.R.M., D.Ly, V.J.M.M., A.K., D.S.,
622 C.K., A.Gi., L.O., L.R., J.V.O., P.R., M.D., and K.P. performed analyses and data interpretation. E.C.,
623 F.W., J.R.M., L.P. and E.W. wrote the manuscript with contributions of all authors.
624

625 **FIGURES**

626

627



628

629

630

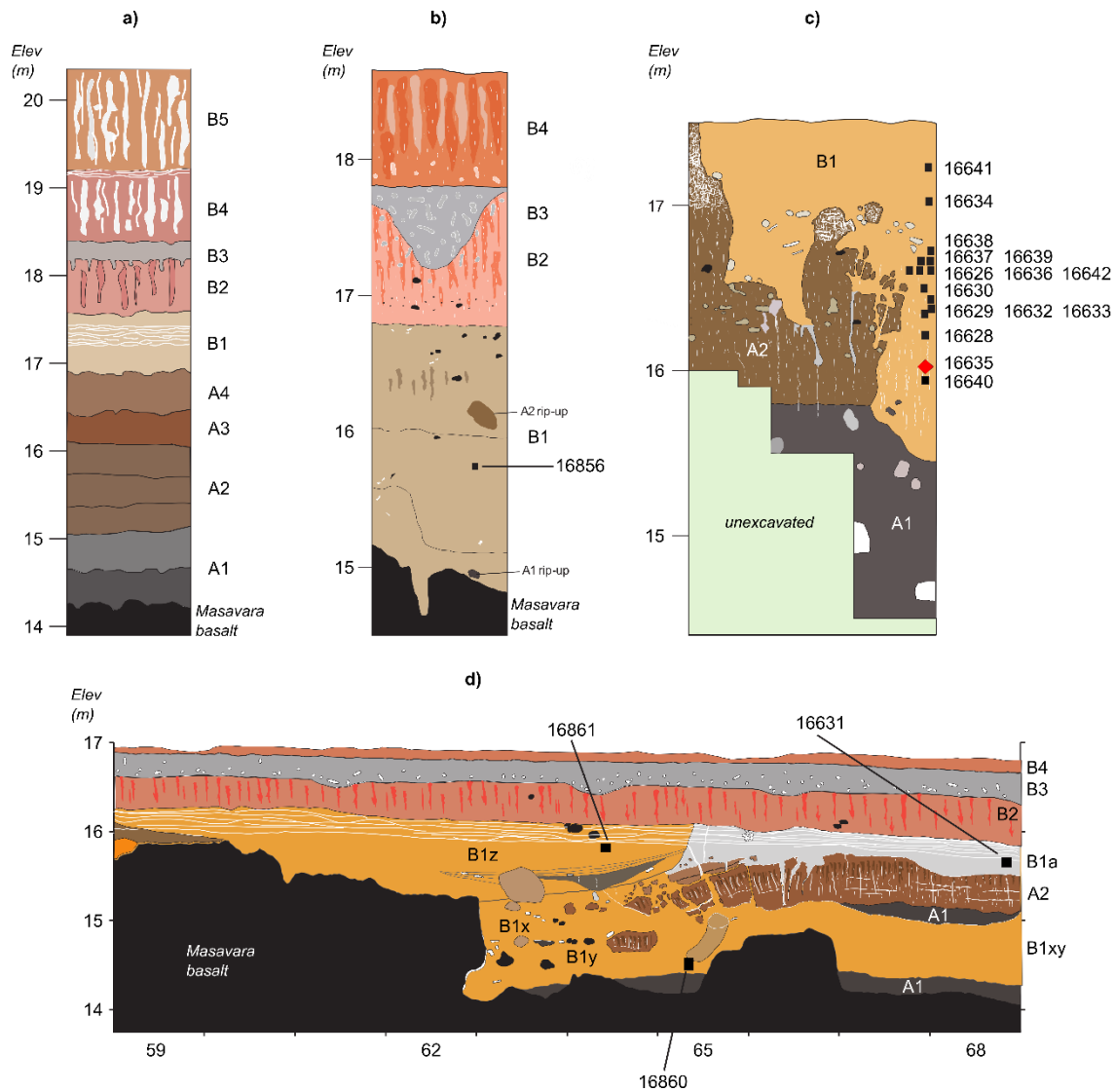
631

632

633

634

Figure 1. Dmanisi location, stratigraphy, and rhinoceros sample 16635. a) Geographic location of Dmanisi in the South Caucasus. **b)** Generalized stratigraphic profile indicating origin of the analysed specimens, recovered in layer B1 and dated to between 1.76 and 1.78 Ma. **c)** Isolated left lower molar (m1 or m2; GNM Dm.5/157-16635) of *Stephanorhinus* ex gr. *etruscus-hundsheimensis*, from Dmanisi (labial view). Scale bar: 1 cm.



635
 636 **Figure 2. Generalized stratigraphic profiles for Dmanisi, indicating sample origins.** a) Type section
 637 of Dmanisi in the M5 Excavation block. b) Stratigraphic profile of excavation area M6. M6 preserves
 638 a larger gully associated with the pipe-gully phase of stratigraphic-geomorphic development in
 639 Stratum B1. The thickness of Stratum B1 gully fill extends to the basalt surface, but includes “rip-
 640 ups” of Strata A1 and A2, showing that B1 deposits post-date Stratum A. c) Stratigraphic section of
 641 excavation area M17. Here, Stratum B1 was deposited after erosion of Stratum A deposits. The
 642 stratigraphic position of the *Stephanorhinus* sample 16635 is highlighted with a red diamond. The
 643 Masavara basalt is ca. 50 cm below the base of the shown profile. d) Northern section of Block 2.
 644 Following collapse of a pipe and erosion to the basalt, the deeper part of this area was filled with
 645 local gully fill of Stratum B1/x/y/z. Note the uniform burial of all Stratum B1 deposits by Strata B2-
 646 B4. Sampled specimens are indicated by five-digit numbers (Tab. 1). Note differences in y-axis for
 647 elevation. Five additional samples were studied from excavation area R11, stratigraphic unit B1, not
 648 shown in a stratigraphic profiles here.

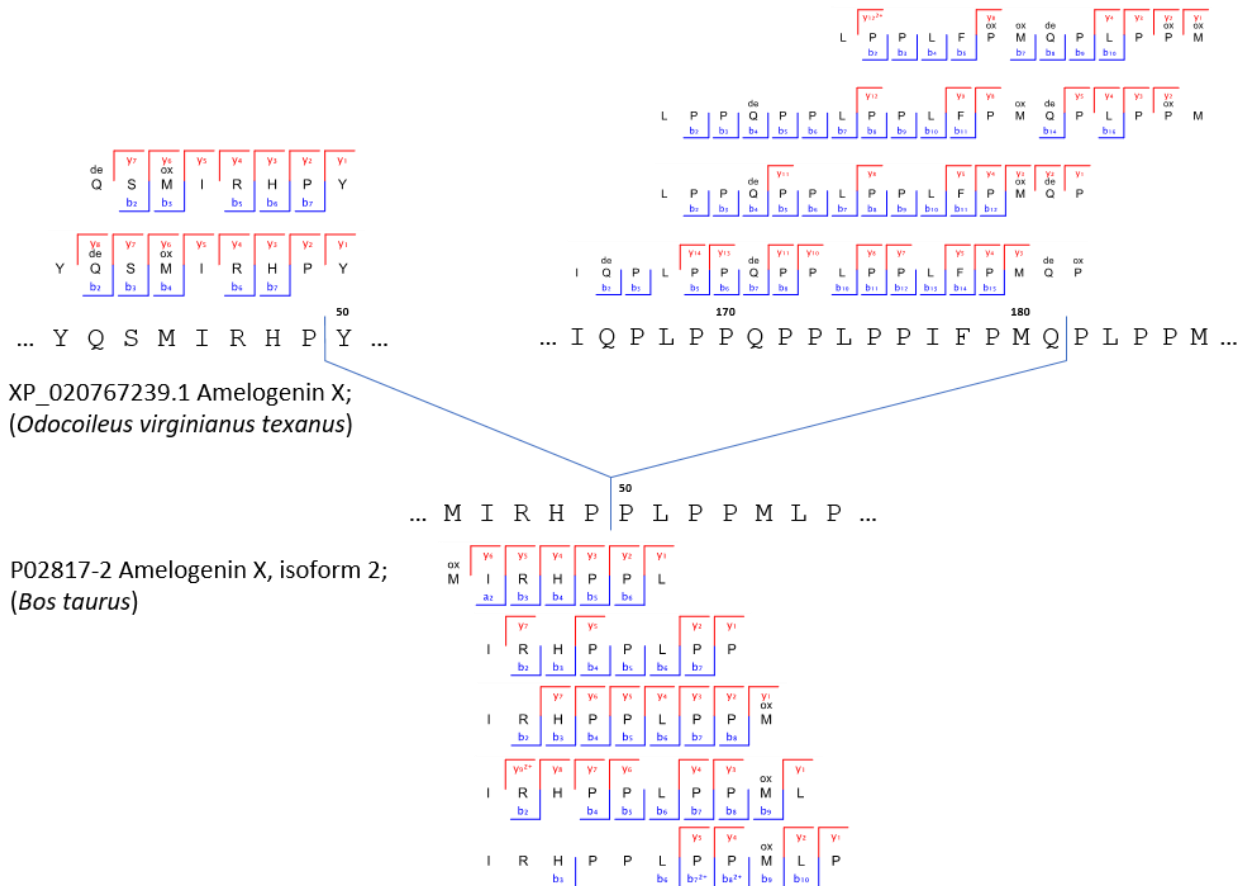
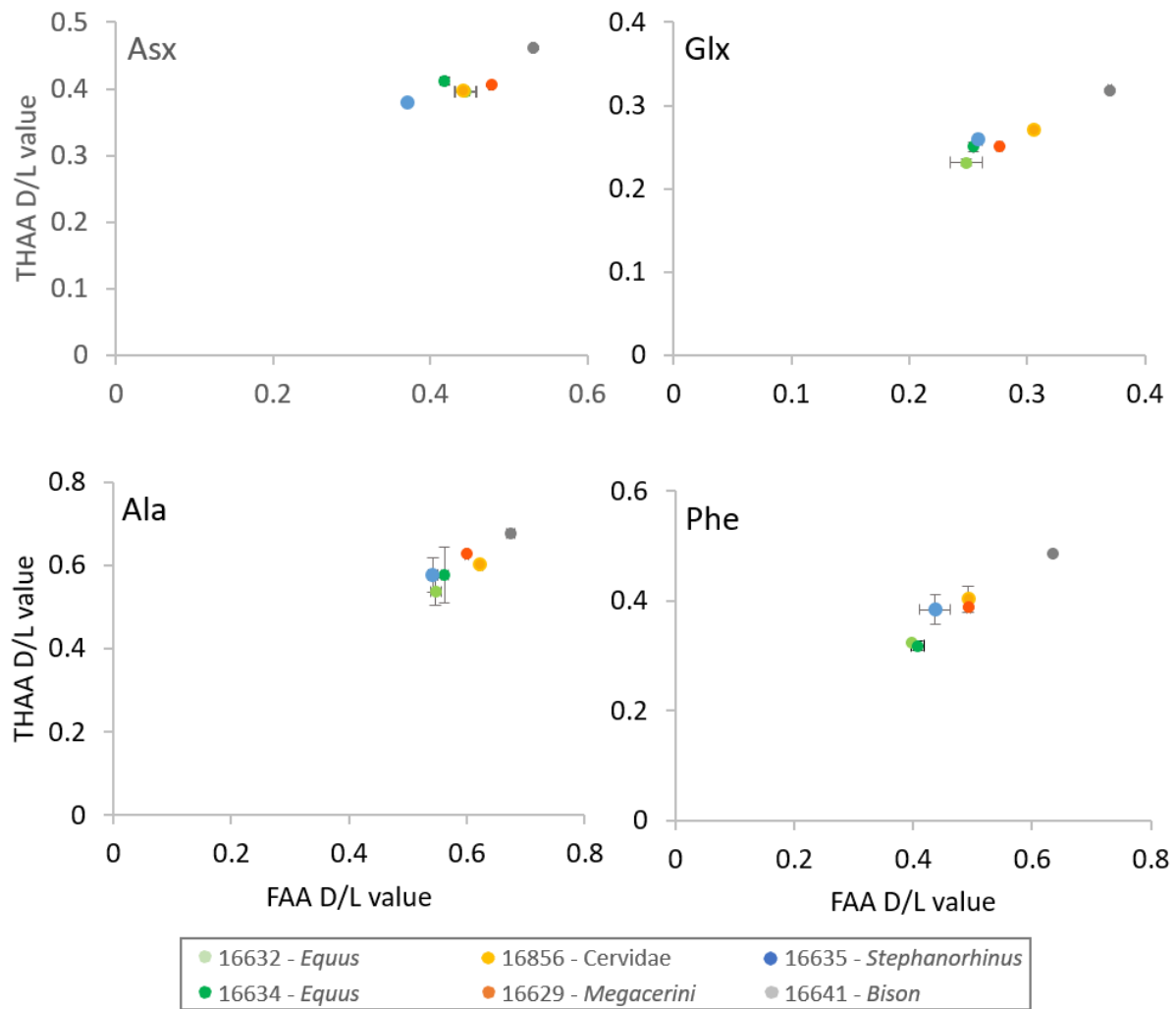


Figure 3. Peptide and ion fragment coverage of amelogenin X (AMELX) isoforms 1 and 2 from specimen 16856 (Cervidae). Peptides specific for amelogenin X (AMELX) isoforms 1 and 2 appear in the upper and lower parts of the figure respectively. No amelogenin X isoform 2 is currently reported in public databases 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 alternative Amelogenin X isoform from the translation product of an alternatively spliced transcript.



659

660

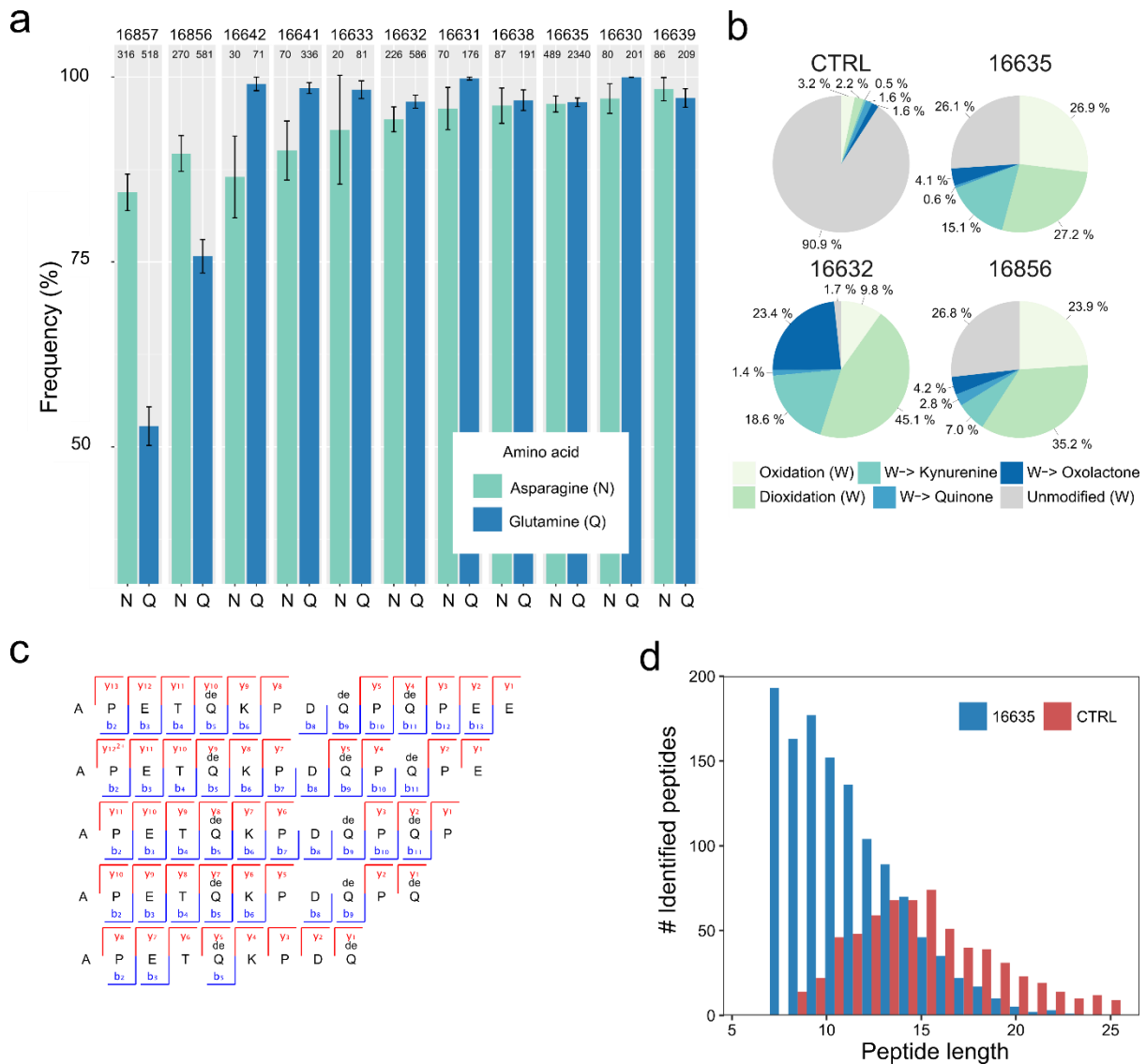
661 **Figure 4. Amino Acid Racemisation.** Extent of intra-crystalline racemization for four amino acids

662 (Asx, Glx, Ala and Phe). Error bars indicate one standard deviation based on preparative replicates

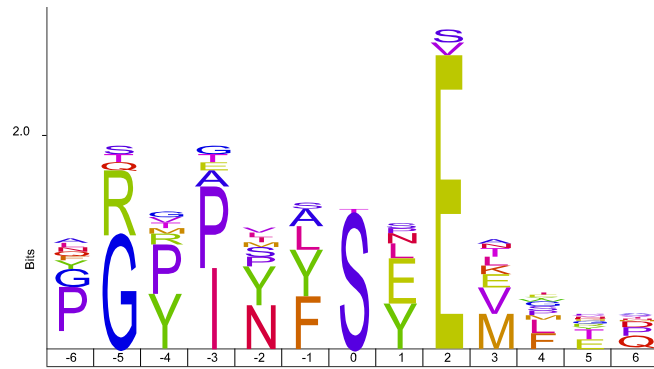
663 (n=2). "Free" amino acids (FAA) on the x-axis, "total hydrolysable" amino acids (THAA) on the y-axis.

664

664



665
 666 **Figure 5. Enamel proteome degradation.** **a)** Deamidation of asparagine (N) and glutamine (Q) amino
 667 acids. Error bars indicate confidence interval around 1000 bootstrap replicates. Numeric sample
 668 identifiers are shown at the very top, while the number of peptides used for the calculation are
 669 indicated for each bar. **b)** Extent of tryptophan (W) oxidation leading to several diagenetic products,
 670 measured as relative spectral counts. **c)** Peptide alignment (positions 124-137, enamelins) for acid
 671 demineralisation without enzymatic digestion. **d)** Barplot of peptide length distribution of
 672 Pleistocene *Stephanorhinus* ex gr. *etruscus-hundsheimensis* (16635) and Medieval (CTRL)
 673 undigested ovicaprine dental enamel proteomes, extracted and analysed in an identical manner.
 674



675

676

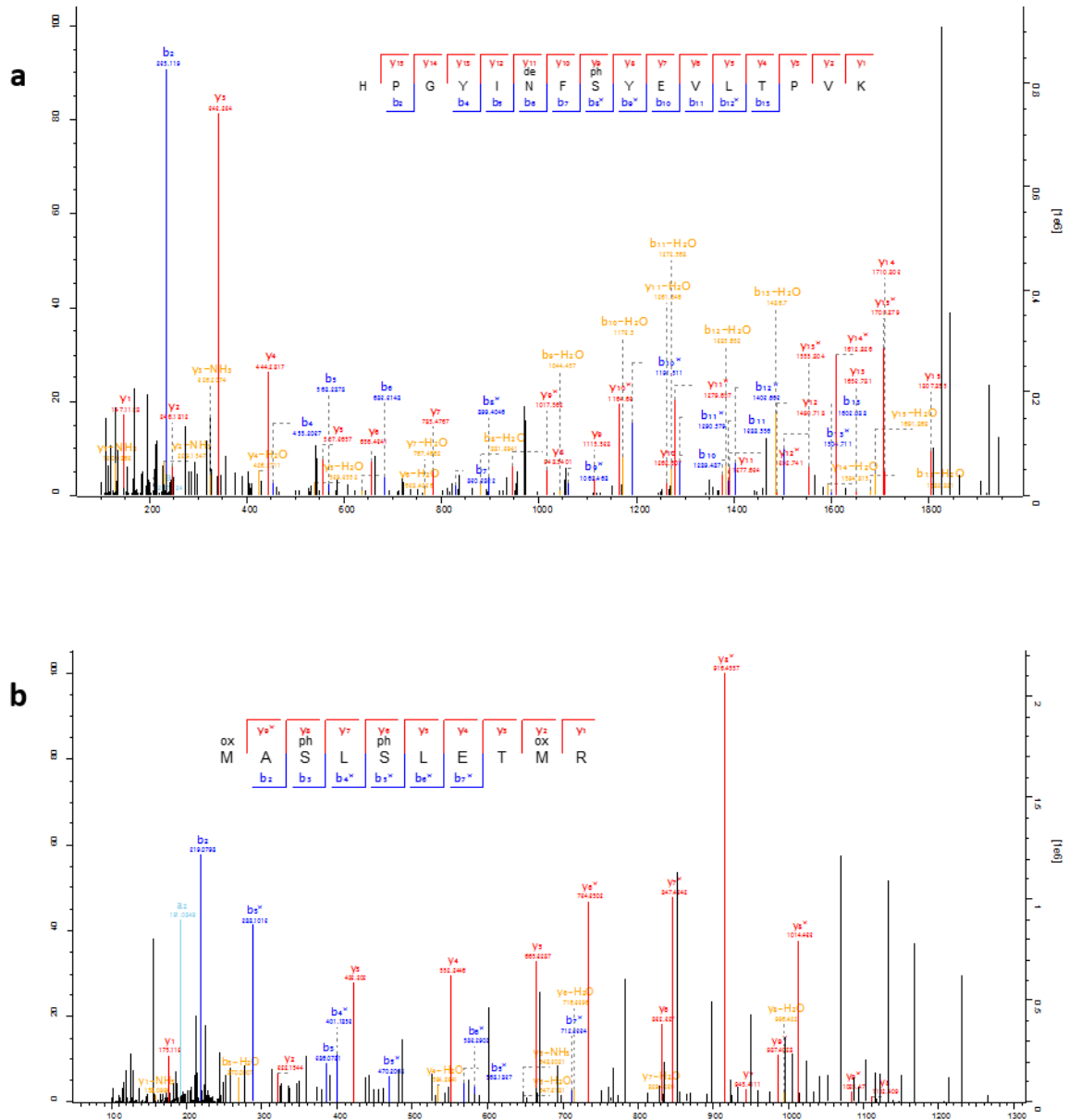
677

678

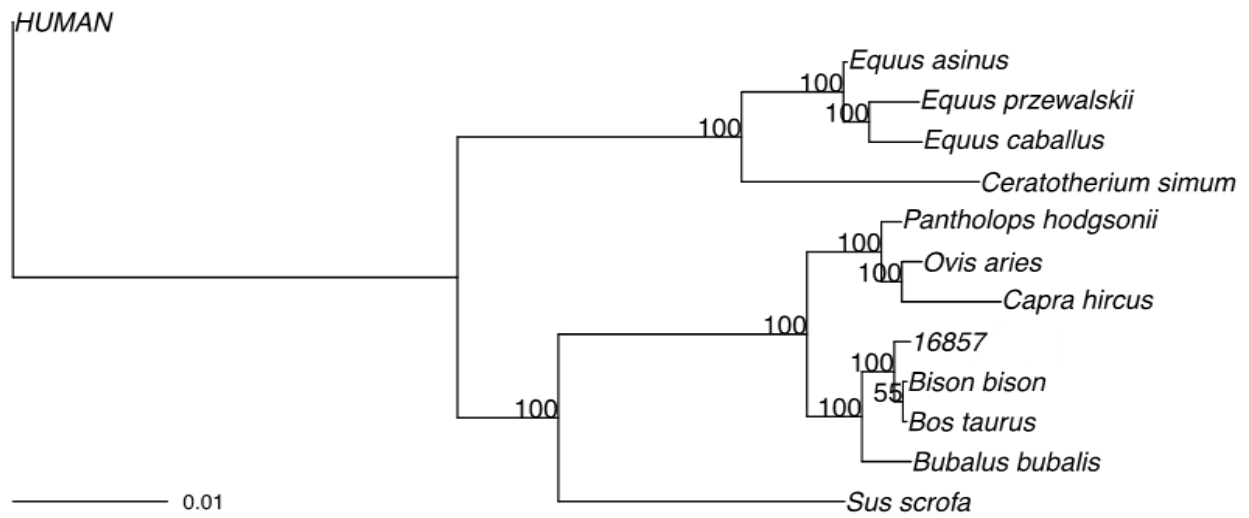
679

680

Figure 6. Sequence motif analysis of ancient enamel proteome phosphorylation. The identified S-x-E/phS motif is recognised by the secreted kinases of the Fam20C family, which are dedicated to the phosphorylation of extracellular proteins and involved in regulation of biomineralization²⁶. See Fig. 7 for spectral examples of both S-x-E and S-x-phS phosphorylated motifs.

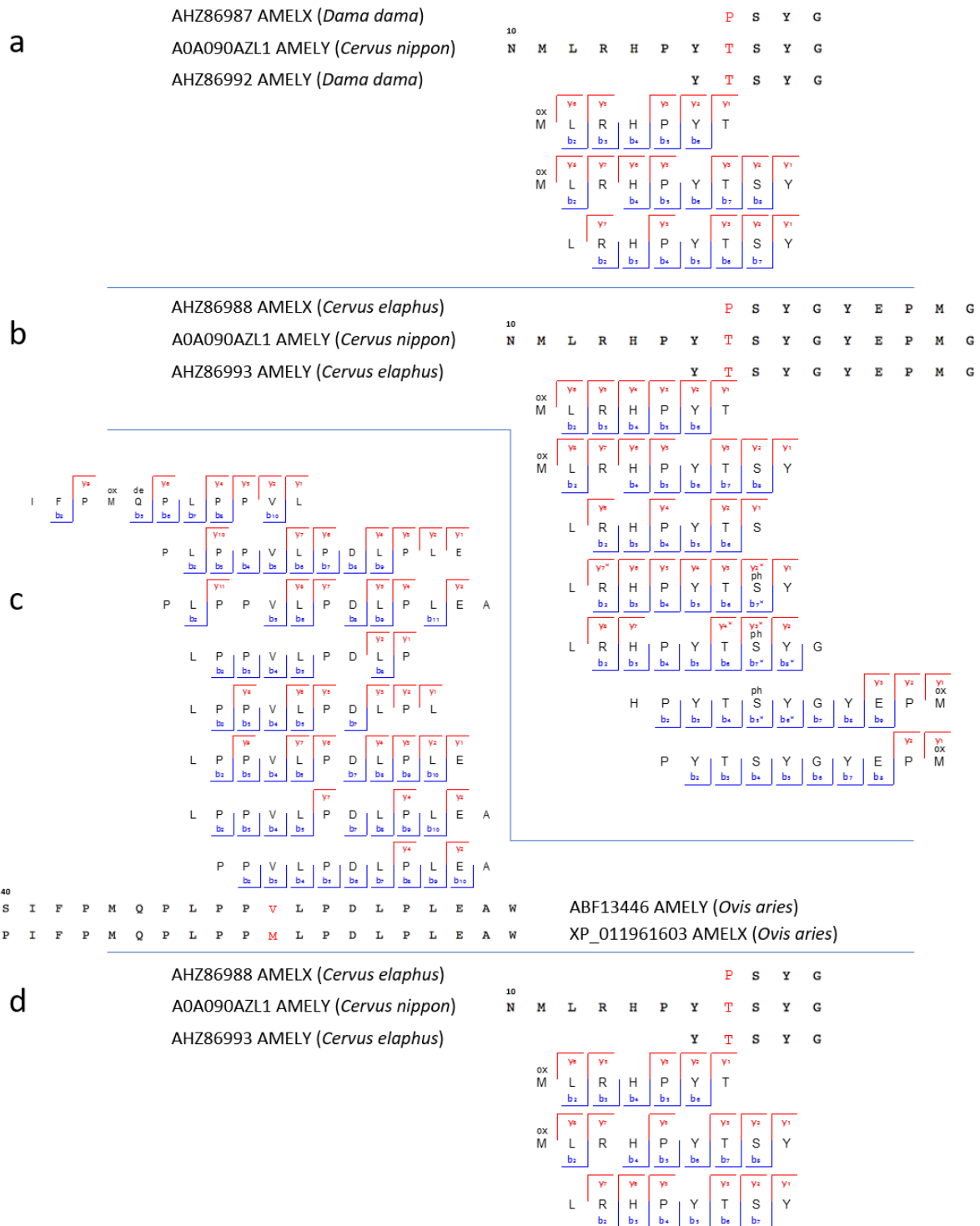


681
 682 **Figure 7. Ancient enamel proteome phosphorylation.** Annotated example spectra including
 683 phosphorylated serines (phS) in the S-x-E motif (**a**; AMEL), and in the S-x-phS motif (**b**; AMBN), as
 684 well as deamidated asparagine (deN). Icelogo analysis of all phosphorylated amino acids indicates
 685 the majority derive from Fam20C kinase activity with a specificity for the phosphorylation of S-x-E
 686 or S-x-phS motifs (see Fig. 6).
 687



688
689
690
691
692
693

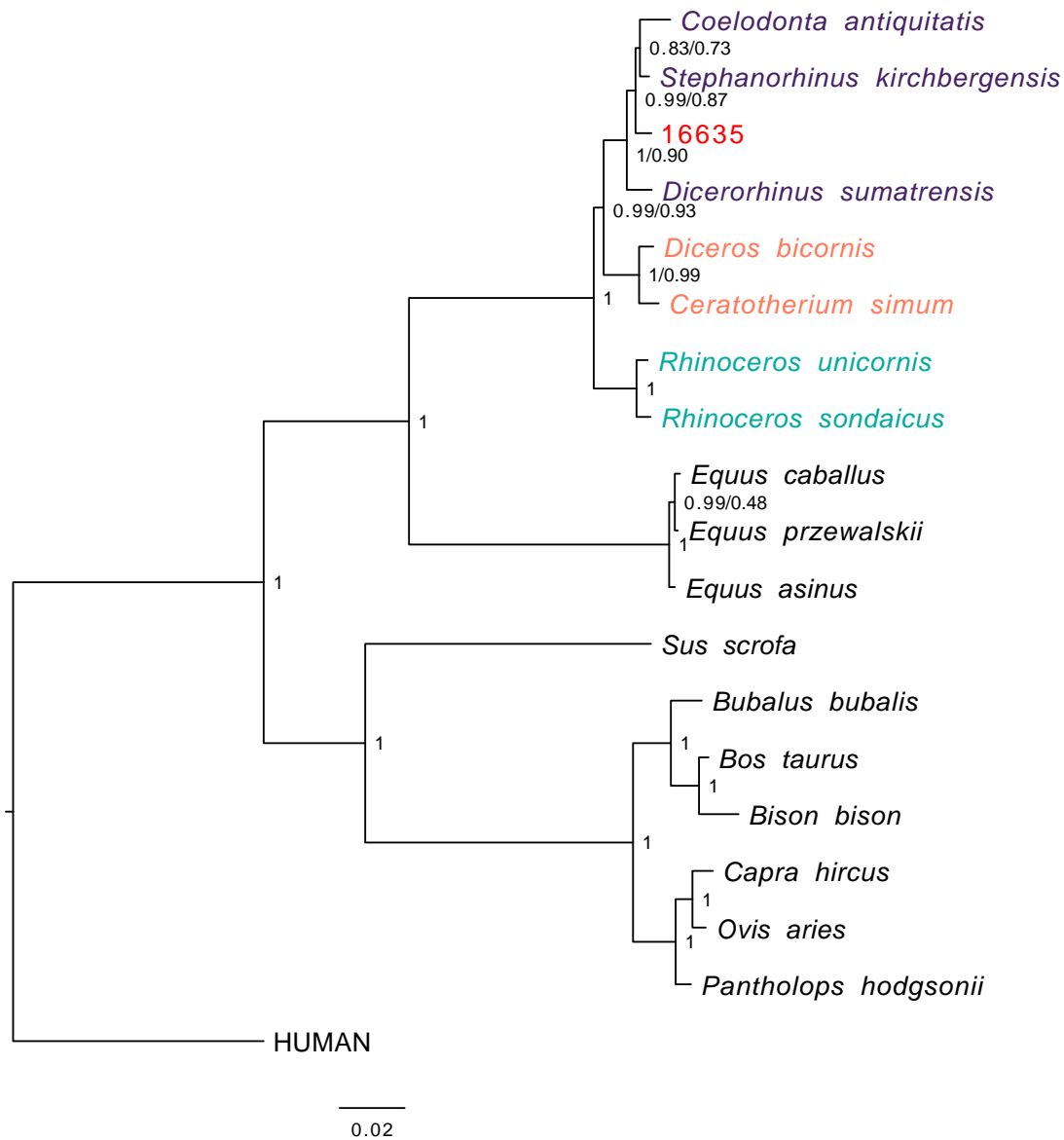
Figure 8. Phylogenetic relationships between the comparative reference dataset and sample 16857. Consensus tree from Bayesian inference. The posterior probability of each bipartition is shown as a percentage to the left of each node. For all panels, we show a scale for estimated branch lengths.



694
695
696
697
698
699

Figure 9. Amelogenin Y-specific matches. a) Sample 16630, Cervidae. **b)** Sample 16631, Cervidae. **c)** Sample 16639, Bovidae. **d)** Sample 16856, Cervidae. Note the presence of deamidated glutamines (deQ) and asparagines (deN), oxidated methionines (oxM), and phosphorylated serines (phS) in several of the indicated γ - and b -ion series.

700



701

702

703

704

705

706

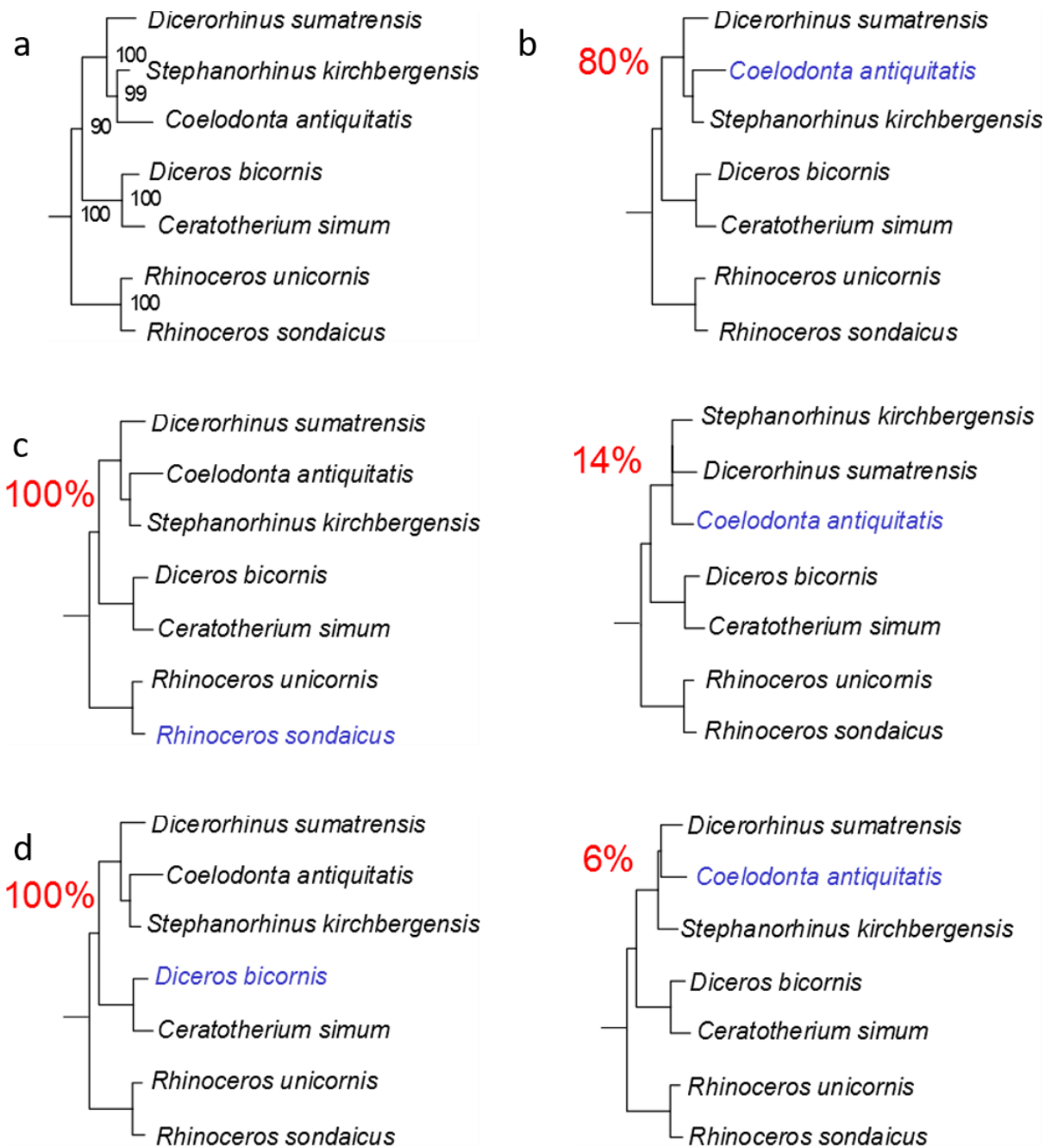
707

708

709

710

Figure 10. Phylogenetic relationships between the comparative enamel proteome dataset and specimen 16635 (*Stephanorhinus ex gr. etruscus-hundsheimensis*). Consensus tree from Bayesian inference on the concatenated alignment of six enamel proteins and 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. Colours indicate the three main rhinoceros clades: Sumatran-extinct (purple), African (orange) and Indian-Javan (green), as well as the specimen 16635 (red).



711
712
713
714
715
716
717
718
719

Figure 11. Effect of the missingness in the tree topology. a) Maximum-likelihood phylogeny obtained using PhyML and the protein alignment excluding the ancient Dmanisi rhinoceros. **b)** Topologies obtained from 100 random replicates of the Woolly rhinoceros (*Coelodonta antiquitatis*). Each replicate was added a similar amount of missing sites as in the Dmanisi sample (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*).

720
721
722

TABLES

n.	CGG reference number	GNM specimen field number	Year of finding	Anatomical identification	Order	Family	Species*
1	CGG_1_016486	Dm.bXI.sqA6.V...	1984	P4 sin.	Camivora	Canidae	<i>Canis etruscus</i>
2	CGG_1_016626	Dm.6/154.2/4.A4.17	2014	tibia sin.	Artiodactyla	Indet.	Indet.
3	CGG_1_016628	Dm.7/154.2.A2.27	2014	mc III&IV dex.	Artiodactyla	Cervidae	Tribe Megacerini
4	CGG_1_016629	Dm.5/154.3.A4.32	2014	hemimandible sin. with dp2, dp3, dp4, m1	Artiodactyla	Cervidae	Tribe Megacerini
5	CGG_1_016630	Dm.6/151.4.A4.12	2014	hemimandible dex. with p2-m3	Artiodactyla	Cervidae	<i>Pseudodama nestii</i>
6	CGG_1_016631	Dm.69/64.3.B1.53	2014	maxilla sin. with P3	Artiodactyla	Cervidae	Tribe Megacerini
7	CGG_1_016632	Dm.5/154.2.A4.38	2014	i3 dex.	Perissodactyla	Equidae	<i>Equus stenoris</i>
8	CGG_1_016633	Dm.5/153.3.A2.33	2014	mc III & mc II sin.	Perissodactyla	Equidae	<i>Equus stenoris</i>
9	CGG_1_016634	Dm.7/151.2.B1/A4.1	2014	m/1 or m/2 dex.	Perissodactyla	Equidae	<i>Equus stenoris</i>
10	CGG_1_016635	Dm.5/157.profile cleaning	2014	m/1 sin.	Perissodactyla	Rhinocerotidae	<i>Stephanorhinus</i> sp.
11	CGG_1_016636	Dm.6/153.1.A4.13	2014	tibia dex.	Perissodactyla	Rhinocerotidae	Rhinocerotini indet.
12	CGG_1_016637	Dm.7/154.2.A4.8	2014	mt III&IV sin.	Artiodactyla	Bovidae	Tribe Ovisovini? Nemorhaedini?
13	CGG_1_016638	Dm.5/154.1.B1.1	2014	hemimandible dex. with p3-m3	Artiodactyla	Bovidae	Tribe Nemorhaedini
14	CGG_1_016639	Dm.8/154.4.A4.22	2014	maxilla dex. with P2-M2	Artiodactyla	Bovidae	Tribe Ovisovini? Nemorhaedini?
15	CGG_1_016640	Dm.6/151.2.A4.97	2014	mt III&IV sin.	Artiodactyla	Bovidae	<i>Bison georgicus</i>
16	CGG_1_016641	Dm.8/152.3.B1.2	2014	m3 dex.	Artiodactyla	Bovidae	<i>Bison georgicus</i>
17	CGG_1_016642	Dm.8/153.4.A4.5	2014	hemimandible sin. with p1-m2	Camivora	Canidae	<i>Canis etruscus</i>
18	CGG_1_016856	Dm.M6/7.II.296	2006	m2 sin.	Artiodactyla	Cervidae	Tribe Megacerini
19	CGG_1_016857	Dm.bXI.profile cleaning		long bone fragment of a herbivore	Indet.	Indet.	Indet.
20	CGG_1_016858	Dm.bXI.North.B1a.colleciton	2006	metapodium fragment	Artiodactyla	Cervidae	Tribe Megacerini
21	CGG_1_016859	D4.collection		fragments of pelvis and ribs of a large mammal	Indet.	Indet.	Indet.
22	CGG_1_016860	Dm.65/62.1.A1.collection	2011	P4 sin.	Artiodactyla	Cervidae	Tribe Megacerini
23	CGG_1_016861	Dm.64/63.1.B1z.collection	2010	fragment of an upper tooth	Perissodactyla	Equidae	<i>Equus stenoris</i>

723
724
725
726
727
728

Table 1. Fossil specimens selected for ancient protein and DNA extraction. 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 traditional comparative anatomy methods.

Specimen	Protein Name	Sequence length (aa)	Razor and unique peptides	Matched spectra*	Coverage after MaxQuant searches (%)	Final coverage after MaxQuant and PEAKS searches (%)	Final coverage (aa)
16628	Collagen alpha-1(I)	1158	5	8	3.2	3.2	37
16629	Amelogenin X	209	79	190	36.8	36.8	77
	Ameloblastin	440	51	84	25.0	25.0	110
	Enamelin	1129	58	133	6.2	6.5	73
	Collagen alpha-1(I)	1453	3	3	2.0	2.0	29
	Collagen alpha-1(III)	1464	2	3	1.4	1.4	20
	Amelotin	212	2	2	4.7	4.7	10
16630	Enamelin	1129	180 3	530 5	11.8 2.7	15.4	174
	Ameloblastin	440	105	231	30.9	31.4	138
	Amelogenin X	213	116	529	62.0	62.9	134
	Amelogenin Y	192	4	9	13.0	22.9	44
	Amelotin	212	5	6	8.0	8.0	17
16631	Enamelin	916	175	751	11.0	11.7	107
	Amelogenin X	213	156	598	48.8	61.5	131
	Amelogenin Y	90	5	18	15.6	25.6	23
	Ameloblastin	440	71	133	24.1	25.2	111
	MMP20	482	2	2	3.9	3.9	19
16632	Enamelin	1144	401	2160	17.9	19.1	219
	Amelogenin X	192	280	960	84.4	84.4	162
	MMP20	424	49	67	33.3	33.3	141
	Serum albumin	607	11	18	6.1	6.1	37
	Collagen alpha-1(I)	1513	4	4	2.6	2.6	40
16634	Amelogenin X	185	68	157	53.5	53.5	99
	Ameloblastin	440	47	58	23.4	23.4	103
	Enamelin	920	33	87	4.5	4.5	41
	MMP20	483	4	4	5.6	5.6	27
16635	Amelogenin X	206	394 3	2793 5	73.8 7.8	85.9	177
	Enamelin	1150	382 2	2966 2	18.3 1.6	25.1	289
	Ameloblastin	442	131	463	31.3	39.3	166
	Amelotin	267	26	148	9.9	9.9	20
	Serum albumin	607	34	64	18.5	24.5	149
	MMP20	483	15	25	11.8	15.3	74
16637	Collagen alpha-1(I)	1453	2	2	1.7	1.7	25
	Collagen alpha-1(II)	1421	2	2	1.9	1.9	27
	Collagen alpha-1(III)	1464	2	2	1.6	1.6	23
	Enamelin	1142	2 5	2 5	3.6 3.0	3.6	41
16638	Enamelin	1129	235 7	1155 13	11.8 4.7	12.9	146
	Amelogenin X	192	185 3	734 5	52.0 10.9	60.4	116
	Ameloblastin	440	64 2	120 4	30.0 5.7	36.4	160
	MMP20	481	6	7	8.1	9.1	44
16639	Enamelin	1129	202	726	12.0	12.6	142
	Amelogenin X	213	167	624	59.2	67.6	144
	Ameloblastin	440	88	155	26.8	30.5	134
	Amelogenin Y	192	13	13	18.8	18.8	36
16641	Amelogenin X	213	91	251	64.3	65.3	139
	Ameloblastin	440	69	122	28.9	28.9	127
	Enamelin	1129	24	75	7.8	7.8	88
	Amelotin	212	3	3	7.1	7.1	15
16642	Amelogenin X	185	89	245	42.7	42.7	79
	Enamelin	733	14	19	2.5	2.5	18
	Ameloblastin	421	3	3	7.1	7.1	30
	MMP20	483	2	2	3.5	3.5	17
16856	Amelogenin X	209	66 4	365 25	38.8	45.5	95
	Enamelin	916	58 13	153 70	8.2	10.2	93
	Ameloblastin	440	21	31	14.8	14.8	65
	Collagen alpha-1(I)	1047	8 10	9 11	14.5	16.9	177
	Collagen alpha-2(I)	1054	4 8	5 9	10.6	10.6	112
	Serum albumin	583	0 8	0 12	16.6	16.6	97
	Amelogenin Y	90	3	7	10.0	10.0	9
16857	Collagen alpha-1(I)	1047	18 14	24 18	21.7	23.4	245
	Collagen alpha-2(I)	1274	16 11	17 11	17.7	24.3	310
16860	Amelogenin X	192	46	98	30.7	32.3	62
	Ameloblastin	440	19	37	9.1	9.1	40
	Enamelin	900	15	25	3.8	3.8	34
16861	Amelogenin X	185	14	15	36.8	38.9	72
	Ameloblastin	343	2	2	4.4	4.4	15
	Enamelin	915	2	2	1.2	1.2	11
Neg. Contr. Gr. 1:							
235, 275, 706	ND						
Neg. Contr. Gr. 2:							
630, 875, 889	ND						
Neg. Contr. Gr. 3:							
1214, 1218	Amelogenin X	122	5	7	18.0	18.0	22

729
730
731
732
733
734
735

Table 2. Proteome composition and coverage. 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 MaxQuant (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.