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

Supplementary Materials, Methods and Results, with Figures and Tables

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

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

Table of contents

SUPPLEMENTARY MATERIALS

1. The archaeological site at Dmanisi and its geology

Reid Ferring, Gocha Kiladze, David Lordkipanidze

The archaeological and palaeontological site of Dmanisi is located about 65 km southwest of Tbilisi, the capital city of Georgia (South Caucasus), in the Kvemo Kartli region[1]. Today the site is situated at an elevation of 910 m MSL on a promontory that is isolated on two sides by the deeply entrenched Mashavera and Pinesauri rivers. Ca. 1.8 Ma ago, a flood of mafic lavas came down the Mashavera Valley, filling the valley. Those lavas also turned at the confluence and flowed up the Pinesauri Valley for about 3 km, forming a geological ramp, which created a lake on the southeastern edge of the Dmanisi promontory. The Mashavera basalt functions as bedrock for the overlying sediments that contain all the artifacts, faunas and hominin fossils at Dmanisi[1].

Shortly after the lavas cooled, ash falls began to accumulate on the promontory, thereby including all the artifacts and faunas as stratified deposits within Stratum A. Those deposits exhibit normal geomagnetic polarity, as does the Mashavera Basalt. Following a brief erosional phase, Stratum B sediments began accumulating, beginning with Stratum B1, which contains the vast majority (ca. 88%) of all recovered faunas, as well as all of the hominin fossils[2,3]. Stratum B continued to accumulate through Strata B2-B5, all of which contain artifacts and faunas. Stratum B deposits exhibit reversed geomagnetic polarity. Stratum B deposits are overlain by the Orozmani Basalt at two localities further upstream from Dmanisi (Extended Data Fig. 1a)[1,4].

Absolute ages and paleomagnetic data provide a chronology for all the deposits at Dmanisi. Stratum A is bound by the underlying Mashavera Basalt, dated to 1.8 Ma, and the A/B contact, which corresponds to the Olduvai-Upper Matuyama magnetic reversal, globally dated to 1.78 Ma. The Orozmani Basalt dates to 1.76 Ma, thereby bracketing the age of the Stratum B deposits date between 1.78 Ma and 1.76 Ma[2].

Well-preserved bones have been recovered from all nine of the strata (A1-B5), although the great majority are from Stratum B1. Stratum A deposits are all silt and very fine sand ashes, that accumulated rapidly, as indicated by weakly developed soils in each of the strata. Stratum B1 has the most complex, yet very brief depositional history. Stratum B1 ashes and some loess accumulated rapidly, based on weakly developed soil features, as well as exceptionally well-preserved faunas that occur as dense accumulations in all areas of B1. Significantly, pipes and gullies formed on and in the Stratum A deposits, and these are all filled with Stratum B1 sediments. Of additional significance is that thousands of stone artifacts have been recovered from Dmanisi. These have been found in each of the nine strata[2]. Extensive test excavation on the promontory has defined an area of at least 40,000 sq m containing in situ artifacts and faunas, indicating numerous, repeated occupations of the promontory.

The exceptional preservation of faunas and human fossils at Dmanisi can be attributed to several factors. First, all were buried by deposits of mafic ashes in Stratum A or ashes with calcareous loess in Stratum B. The sediments in each stratum were deposited rapidly, subjected to minor colluvial reworking. After deposition, each stratum exhibits weak weathering of the mafic minerals, and modest accumulation of pedogenic carbonates. All of the bones in and below Stratum B1 were further protected by the development of a laminated calcrete hardpan in the upper part of Stratum B1. This precluded disturbance as well as percolation of meteoric waters that would have been harmful to fossil and organic preservation.

1.1. Stratigraphy and formation context of the specimens

All of the faunal specimens analysed in this study were recovered in Stratum B1 deposits, which have been firmly dated between 1.77-1.76 Ma[1]. Stratum B1 was deposited very quickly, as can be inferred from the minimal evidence of soil development in those sediments[2], as well as the extremely low degree of weathering of the bones[3,5]. Despite its short depositional and weathering

history, Stratum B1 exhibits a number of geomorphic-sedimentary facies that create a complex microstratigraphy. Most of these are the result of piping and gullying, but also include large burrows, most of which were probably made by carnivores. *Histrix* is present in low numbers in the fauna, so some of their burrows may be included[5].

The stratigraphy of Dmanisi is represented by the M5 section, which is the deepest and most complete exposed so far (Fig. 1b, Extended Data Fig. 1a). The M5 section shows all nine strata, divided into groups A and B. All of the Stratum A deposits accumulated during the late Olduvai subchron, between 1.85 Ma and 1.78 Ma[1-3]. These accumulated as a series of ash falls with minor weathering and erosion between each of the strata A1 through A4. It was in these deposits that the gullies and pipes, filled with Stratum B1 sediments, form. In the M5 section, Stratum B1 overlies stratum A4 and is separated from that by a minor erosional disconformity. Also in the M5 section, Stratum B1 is a horizontal ash accumulation with no evidence of pipes or gullies. Everywhere on the promontory Stratum B1 is conformably overlain by Strata B2 and B3. Although Stratum B4 has been exposed in some of the exposures to the east of M5, Stratum B5 is thus far only preserved in the M5 excavation area.

The specimens investigated in this study were selected from five separate excavation areas on the Dmanisi promontory (Extended Data Tab. 1). Five specimens were recovered from a profile in the "Room 11" section. This was an area excavated in the middle 1980s and was partially backfilled. No excavations have taken place there since, although the profile was cleaned and photographed in 2006. These specimens were recovered during those activities and are ascribed to Stratum B1. In addition, one bone fragment was recovered from the profile of a filled pipe exposed in a cleaned Medieval cellar (aka "*darani*") along the west wall of Block 1 (Extended Data Fig. 1b).

Three specimens were collected from the northern part of Block 2 (Extended Data Fig. 1d). These also came from Stratum B1. The majority of the human fossil remains have been recovered from this part of the site. Here, Stratum A sediments that accumulated in a depression on the basalt surface were later subjected to piping, gullying and animal burrowing. Compared to the profile in the M5 section, the effects of these processes in Block 2 are clearly seen. The Stratum B1 deposits here include both horizontal beds of the B1 ashes (B1a) as well as pipe and gully facies (B1 x, y, z). It is the latter deposits that contain the highest densities of faunal materials as well as all of the human bones found in this part of the site.

Fifteen of the specimens, including the *Stephanorhinus* sp. molar (Dm.5/157-16635) were recovered from the newest excavation area, M17, located on the northern part of the promontory, near the modern edge of the Masavera Gorge. The M17 section (Extended Data Fig. 1c) reveals a geologic profile that is very different from the others in detail, but not in terms of the overall stratigraphic sequence. As can be seen in the profile, thick accumulations of black strata A1 ashes accumulated above the Masavera Basalt (the basalt is ca. 50 cm below the base of the shown profile). The A1 ashes were conformably overlain by A2. As in other parts of the site this is a continuous depositional sequence. Noticeably missing are Strata A3 and A4. The strata A deposits were quite heavily eroded, leaving a tall pillar of Strata A1-A2 in the northwest corner of the excavation block as shown in the profile. Following this erosion, Stratum B sediments began to accumulate, apparently quite thickly. It was in these deposits that large burrows were excavated, presumably by carnivores. All of the M17 Stratum B1 deposits contain dense faunal materials, including numerous articulated elements of large herbivores as well as numerous carnivore remains. Rapid deposition, minimal weathering and probably substantial carnivore accumulation of bones accounts for the large numbers and good preservation here. It is not yet understood what accounts for the extremely hard silica-rich coatings on these bones, which are diagenetic features not seen in any of the other excavation areas. It is important to note that the M17 test was located in an excavated Medieval house, with the goal of avoiding thick Medieval deposits. While the Medieval deposits were in fact thin, it also became clear that the Medieval house builders had themselves excavated down through the sediments removing all of Stratum B2 and higher part of the section. All of the pipe fills examined thus far at Dmanisi have yielded reversed paleomagnetic polarity and are facies of Stratum B1[1,3].

Rapid deposition coupled with minimal weathering accounts for the exceptional bone and tooth preservation here. Of note is the laminated calcrete in Stratum B1. This is a post-depositional feature related to probable fungal activity. This calcrete is extremely hard and appears to have acted as an aquiclude, protecting all of the fossils below from both meteoric waters and groundwater fluctuations which could have enhanced weathering. Nowhere on the promontory has this calcrete been breached by erosion, helping to account for the excellent preservation of bones and teeth below this horizon.

Excavations in the small block at M6 yielded quite high numbers of artifacts as well as faunal materials, especially in the thick Stratum B1 (Extended Data Fig. 1b). Stratum B1 is interpreted as gully fill deposits, which have been found in other parts of the site and are closely related to the development of pipes and gullies. The thickness of the B1 sediments here is greater than in section such as M5 or Block 2, because the gully had eroded entirely through the Stratum A deposits down to the Masavera Basalt at the base of this section. As in other parts of the site, however, the B1 ashes were deposited rapidly and appear to have quickly filled this gully.

2. Faunal remains and their taphonomy

Maia Bukhsianidze, Bienvenido Martínez-Navarro, Jordi Agusti, Ralf-Dietrich Kahlke, Martha Tappen, David Lordkipanidze

In addition to the unique paleoanthropological record, Dmanisi is one of the richest Early Pleistocene palaeontological sites of Eurasia's temperate zone. Long-term excavations of more than three decades revealed a highly diverse fossil vertebrate fauna and 49 taxa, inclusive of *Homo*, have been recorded so far[3,6-16] (Tab. S1). The large mammal record indicates an age position at the onset of the Late Villafranchian[3,17,18], while the occurrence of the water vole *Mimomys pliocaenicus* matches the Late Villanyan of the micromammal biochronology[19]. The small mammal association is dominated by steppic or dry-adapted rodents, such as the hamsters *Cricetulus* sp., and *Allocricetus bursae*, and the gerbil *Parameriones* aff. *obeidiensis*. Woodland (*Apodemus* aff. *atavus*) or fluviatile elements (*Mimomys pliocaenicus*) are very rare.

Whereas several of the herbivore faunal elements recorded from Dmanisi, such as *Palaeotragus* cf. *priasovicus* and *Gallogoral meneghinii sickenbergii*, are also known from Middle Villafranchian contexts of Western Asia and Europe, a significant number of genera, especially of artiodactyls, reflect the ongoing Late Villafranchian faunal renewal. Members of *Praemegaceros*, *Bison* (*Eobison*), *Capra*, *Soergelia* and *Pontoceros* from Dmanisi all represent the corresponding earliest evidence in the Western Palaearctic. *Cervalces* cf. *gallicus* and *Gallogoral meneghinii sickenbergii* were first recorded in the Southern Caucasus. Other herbivore species - *Arvernoceros insolitus*, *Pseudodama nestii*, *Palaeotragus* cf. *priasovicus*, *Equus stenonis*, *Stephanorhinus sp.* and *Mammuthus meridionalis taribanensis* - are to be considered as descendants of Middle Villafranchian predecessor populations of central Eurasia or the immediate Caucasus region. The remains of *Palaeotragus* and *Gallogoral* belong to the latest western Palaearctic records. Rare suid remains emphasize the fact that this family was probably present in the Southern Caucasus area during most of the Early Pleistocene[20-22]. In summary, the herbivore assemblage of Dmanisi is dominated by elements of Asian origin.

The composition of the species-rich guild of carnivores also points to a Late Villafranchian age. Among Canidae, wolfs are represented by a remarkably rich sample of *Canis etruscus*, while *Vulpes alopecoides* is elusive, possibly because of the territorial dominance of *Canis*. Ursidae are well represented by a series of typical *Ursus etruscus* specimens. Mustelids are only sparsely present: a few specimens represent the Galictine *Pannonictis nestii* and *Lutra simplicidens*. A limited number of fragmentary remains furthermore do not allow specific determinations for *Martes* sp. and *Meles* sp., but both genera are present. *Pachycrocuta brevirostris* is the only hyena species of the Dmanisi fauna. From the Villafranchian up to the early Middle Pleistocene, this hyaena, an African immigrant, played a permanent role as a food competitor of hominins within entire Eurasia[17,23-25]. Felids, the most diverse carnivore family of the site of Dmanisi, are

represented by unusually well and completely preserved remains of *Lynx issiodorensis*, the large *Acinonyx pardinensis*, the oldest Eurasian jaguar *Panthera onca georgica* and the dirk- and saber-toothed cats *Megantereon whitei* and *Homotherium crenatidens*. In summary, the Dmanisi fauna points to a mosaic-like landscape with remarkable differences in humidity and vegetation across its terrain. Whereas the immediate vicinity of the fossil site was situated in a more or less forested valley, the wider region was largely made up of dryer open tree savanna and grassland and by mountainous to semi-arid rocky terrains.

Preservation of the Dmanisi fauna and site formation is relevant for understanding the contemporaneity of the specimens and the situation in which the proteins were preserved. Well over half the specimens sampled for this work come from a new excavation block, M17 (Extended Data Tab. 1). This excavation block has produced hundreds of new fossils in the last few excavation seasons, but only a small portion have been analysed in detail. However, M17 appears to contain the same diverse taxa as other excavation units of the site and is notable for having numerous large complete skeletal elements and a few portions of articulated skeletons. A large number of carnivores come from this area, including an entire *Homotherium crenatidens* skull and *Megantereon megantereon* cranium, and specimens of *Panthera onca georgica*, *Pachycrocuta brevriostrus*, *Canis etruscus*, *Ursus etruscus* as well as a new *Homo* femur that is not yet described. Herbivores are also abundant and range from lagomorphs to rhinoceros in size. Coprolites and carnivore tooth marks were also identified on bones from this block, indicating contemporaneous carnivore activity. Other specimens analysed here come from the B1 stratum in Building 11, Block2 and M6. One specimen comes from Stratum B1 exposed in the walls of a cleaned medieval pit.

Bones in Stratum B1 occur on open surfaces in aggrading ashes, but more commonly in features associated with piping, collapsed pipes, and resulting shallow gullies that were quickly filled with B 1 deposits. Some of the pipes and gullies bear evidence of carnivore denning. For the site overall, new estimates of weathering concur with those reported previously by us[3,5], confirming that little subaerial weathering occurred. Hard sediment completely coats many of the recovered bone specimens, making accessing their surfaces very difficult. Of the 5395 specimens we could assign weathering stages[26] to, 73% are in Stage 0 (unweathered), and another 22% are in weathering stage 1. Higher weathering stages are present, but are much less frequent (Stage 2 = 3.35%; Stage 3= 1.3%; Stage 4 = 0.54 %). Since ~95% of the Dmanisi specimens are in the earliest two stages of weathering, we deduce that the bones were buried immediately within a few years after death. Although the bones often do not show subaerial weathering cracks, other forces have contributed to the destruction of the bone surfaces. The most common bone surface modifications are post-depositional, manganese and calcrete deposits, and various types of surface etching (28% of the non-tooth specimens contain fungal rhizomorph etching of their surfaces, 7.5% due to other chemical or diagenetic processes). Since these causes of etching were generally superficial, it did not affect preservation much.

Table S1. Dmanisi fauna from stratum B deposits.

Class	Order	Family	Genus	Species / Subspecies
Pisces	Salmoniformes	Salmonidae	*Salmo*	sp.
Amphibia	Anura	Bufonidae	*Bufo*	ex. gr. *Viridis*
Reptilia	*Testudinata*	Testudinidae	*Testudo*	*Graeca*
	Squamata	Lacertidae	*Lacerta*	ex. gr. *Viridis*
		Colubridae	*Elaphe*	ex. gr. *quatuorlineata*
			Natrix	sp.
		Colubridae indet.		
Aves	Struthioniformes	Struthionidae	*Struthio*	*dmanisensis*
	Galliformes	Gallidae	*Gallus*	*dmanisiensis*

	Strigiformes	Strigidae	*Strix*	*gigas*
Mammalia	Insectivora	Soricidae	*Beremendia*	*fissidens*
	Lagomorpha	Ochotonidae	cf. *Ochotona*	*lagreli*
		Leporidae	cf. *Hypolagus*	*brachygnathus*
	Rodentia	Muridae	*Apodemus*	aff. *atavus*
		Cricetidae	*Cricetulus*	sp.
			Allocricetus	*bursae*
		Arvicolidae	*Tcharinomys*	*tornensis*
			Mimomys	*pliocaenicus*
		Gerbillidae	*Parameriones*	aff. *obeidiensis*
		Hystricidae	*Hystrix*	*refossa*
	Primates	Hominidae	*Homo*	
	Carnivora	Canidae	*Canis*	*etruscus*
			Vulpes	*alopecoides*
		Ursidae	*Ursus*	*etruscus*
		Mustelidae	*Pannonictis*	*nestii*
			Martes	sp.
			Meles	sp.
			Lutra	*simplicidens*
		Hyaenidae	*Pachycrocuta*	*brevirostris*
		Felidae	*Lynx*	*issiodorensis*
			Acinonyx	*pardinensis*
			Panthera	*onca georgica*
			Megantereon	*whitei*
			Homotherium	*crenatidens*
	Proboscidea	Elephantidae	*Mammuthus*	*meridionalis taribanensis*
	Perissodactyla	Equidae	*Equus*	*stenonis*
		Rhinocerotidae	*Stephanorhinus*	*etruscus/hundsheimensis*
	Artiodactyla	Suidae	*Sus*	sp.
		Cervidae	*Pseudodama*	*nestii*
			Praemegaceors	*obscurus*
			Arvernoceros	*insolitus*
			Cervalces	cf. *gallicus*
		Giraffidae	*Palaeotragus*	cf. *priasovicus*
		Bovidae	*Bison* (*Eobison*)	*georgicus*
			Gallogoral	*meneghinii sickenbergii*
			Capra	*dalii*
			Soergelia	cf. *minor*
			Praeovibos	sp.
			Pontoceros	*surprine*
			Antilopini	indet.

3. Taxonomic attribution of the rhinoceros tooth specimen Dm.5/157–16635

Luca Pandolfi, Lorenzo Rook

The specimen Dm.5/157 (CGG GeoGenetics number 16635; Extended Data Tab. 1), was morphologically compared with several Pliocene and Pleistocene remains assigned to *Coelodonta antiquitatis* (Blumenbach, 1799), *Stephanorhinus kirchbergensis* (Jäger, 1839), *Stephanorhinus hemitoechus* (Falconer, 1859), *Stephanorhinus etruscus* (Falconer, 1868), and *Stephanorhinus hundsheimensis* (Toula, 1902). These specimens were collected from a number of European localities and are currently housed in several European Institutions[27,28]. The morphometric methodology follows that introduced by Guérin[29]. The anatomical descriptions and the dental terminology follow Guérin and Antoine[29,30]. The morphology of the lower teeth is quite conservative in Rhinocerotidae and hence it is difficult to determine the species attribution of isolated Pleistocene rhinoceros teeth. The studied specimen is an isolated left lower m1 or m2. The enamel is rough and the cement is observable on the vestibular side of the tooth (Fig. S1).

Figure S1. *Stephanorhinus* specimen Dm.5/157-16635. Lower left m1 or m2, after sampling for palaeoproteomic analysis. a) Occlusal view. b) Buccal view. c) Lingual view. Scale bar = 1 cm. Panel b replicates Fig. 1c.

In occlusal view, the talonid is rounded, the vestibular surface of the trigonid is rather flat. The vestibular groove is angular V-shaped, deep, and developed until the neck. The paralophid nearly reach the lingual rim. The metaconid and the entoconid are not constricted, the hypolophid is oblique. In lingual view, the anterior valley seems to have a V-shaped morphology and the posterior valley has a relatively broad V-shaped morphology. The ratio between the heights of the lingual valleys, 0.8, is relatively low. The lingual cingulum is absent and the lingual groove on the entoconid is absent. Mesial and distal cingula are present, and the mesial cingulum extends along the vestibular side of the tooth.

The tooth differs from those of *Coelodonta*, which displays more hypsodont teeth, an angular trigonid, almost sagittal hypolophid, and flat vestibular wall of the lophs. Specimen Dm.5/157-16635 also differs from *S. kirchbergensis* which has smooth enamel, usually displays U-shaped lingual valleys, and lacks mesial and distal cingula. The morphological differences between the studied specimen and *S. etruscus*, *S. hundsheimensis* and *S. hemitoechus* are relatively few. Angular V-shaped vestibular grooves are more frequent in the molars of *S. etruscus* (specimens from Upper Valdarno, Italy, and Senèze, France[27]) than in the other species. Deep vestibular grooves are frequently recorded in *S. hundsheimensis*[31] and this character has been reported in several specimens from the Middle Pleistocene of Germany[27]. In *S. etruscus* and *S. hemitoechus*, the vestibular groove on the molars is normally shallow[27,31]. The ratio between the height of the bottoms of the lingual valleys on m1 is normally high in *S. hundsheimensis* and *S. etruscus*[31] and rather small in *S. hemitoechus*. According to Guérin[29] the difference in height between the bottoms of the lingual valleys on m1 is medium to high in *D. etruscus brachycephalus* (= *S. hundsheimensis* in partim), variable in *S. hemitoechus*, and medium in *D. etruscus etruscus* (= *S. etruscus*). The presence of a vestibular cingulum is a variable character in *S. etruscus*[29], nevertheless, in this species the cingula are generally absent or poorly developed. Mesial and distal cingula are usually present on m1 of *S. hundsheimensis*. They are less frequent in *S. hemitoechus* and more often

absent in *S. etruscus*[31]. Concerning the M2, the differences in height between the bottoms of the valleys is normally important in *S. etruscus* and smaller in *S. hemitoechus* and *S. hundsheimensis*[31]. According to Guérin[29], the differences in height between the bottoms of the valleys is variable in *S. hemitoechus* and *S. hundsheimensis*, but in the latter species the vestibular and lingual cingula are rare. The vestibular cingulum is usually present on the m2 of *S. etruscus*[29].

The dimensions of the Dm.5/157-16635 specimen from Dmanisi are not useful to assign it to a well-defined species (Tab. S2). Consequently, due to the absence of diagnostic features and the presence of several characters shared in particular with *S. etruscus* and *S. hundsheimensis*, the studied specimen is assigned as *Stephanorhinus* ex gr. *etruscus-hundsheimensis*, pending a revision of the whole material collected from Dmanisi.

Table S2. Comparative lengths (in mm) of the specimen collected from Dmanisi (Georgia) and those of *S. etruscus*, *S. hundsheimensis* and *S. hemitoechus* collected from several Pleistocene European localities. m1: first lower molar; m2: second lower molar; Dm: Dmanisi collection, GNM: Georgian National Museum. Data from this study: [27-29,31-36]

Locality	Taxon	m1	m2	References/Museum
Dmanisi	Dm.5/157	51.8		GNM
Senèze	*S. etruscus*	40.8-45.3	41.6-43.18	Lacombat 2005, 2006
Val d'Arno	*S. etruscus*	35.13-43.73	40.35-45.98	Lacombat 2005, 2006
Monte delle Piche	*S. etruscus*	43.6	46.29	Pandolfi et al. 2015
Pietrafitta	*S. etruscus*	39.44-41.92	43.87-46.06	Pandolfi & Erten, 2017
Pietrafitta	*S. etruscus*	43.32	44.6	Pandolfi & Erten, 2017
Pietrafitta	*S. etruscus*	41.03-41.5	43.3-43.38	Pandolfi & Erten, 2017
Castellana	*S. cf. etruscus*	39.05	42.5	Pandolfi & Erten, 2017
Aivaliki	*S. cf. etruscus*	39.4	42.7	Symeonidis et al. 2003
Various localities	*S. etruscus*	36-50	40-54	Guérin 1980
Ceyssaguet	*S. hundsheimensis*	40.94		Lacombat 2005, 2006
Vallonnet	*S. hundsheimensis*	38.28-47.35	40.19-47.23	Lacombat 2005, 2006
Soleilhac	*S. hundsheimensis*	42.33	46.44	Lacombat 2005, 2006
Isernia	*S. hundsheimensis*	40.12-48	41.3-50.14	Lacombat 2005, 2006
Durfort	*S. hundsheimensis*		44.18	Lacombat 2005, 2006
Denizli	*S. hundsheimensis*	43	52	Pandolfi & Erten, 2017
Castellana	*S. hundsheimensis*		49.14	Pandolfi & Erten, 2017
Sussenborn	*S. hundsheimensis*	39-45	39.8-48.8	Kahlke 1969
Voigsted	*S. hundsheimensis*	40.5	45.8	Kahlke 1965
Mosbach	*S. hundsheimensis*	41.6-41.2	43.1-43.5	Koenigswald et al. 2007
Tegoleto	*S. cf. hundsheimensis*	47.61	54.26	Pandolfi & Erten, 2017
Various localities	*S. hundsheimensis*	36.5-57	41-57.5	Guérin 1980
Valle Radice	*S. hemitoechus*	53.39	55.88	Pandolfi & Tagliacozzo, 2015
Caune de l'Arago	*S. hemitoechus*	40.43-47.88	46.17-53.22	Lacombat 2005, 2006
Various localities	*S. hemitoechus*	36.5-56.5	41-60.5	Guérin 1980

SUPPLEMENTARY METHODS

4. Biomolecular preservation: amino acid racemization

Marc R. Dickinson, Kirsty Penkman

Chiral amino acid analysis was undertaken on six Pleistocene enamel samples from a range of species retrieved from the excavation block M17 and the stratum B1 from Dmanisi (Dm.5/154.3.A4.32-16629, Dm.5/154.2.A4.38-16632, Dm.7/151.2.B1/A4.1 -16634, Dm.5/157-16635, Dm.8/152.3.B1.2-16641 and Dm.M6/7.II.296-16856), in order to test the endogeneity of the enamel proteome through its degradation patterns. The current technique of amino acid analysis developed for geochronological purposes[37] combines a reverse-phase high-pressure liquid chromatography (RP-HPLC) method of analysis[38] with the isolation of an 'intra-crystalline' fraction of amino acids by bleach treatment[39]. This combination of techniques results in the analysis of D/L values of multiple amino acids from the chemically protected (closed system) protein fraction within the biomineral, thereby enabling both decreased sample sizes and increased reliability of the analysis.

The enamel chips were cleaned of dentine and other tooth tissues in Copenhagen, before they were transferred to York. Enamel chips were powdered with an agate pestle and mortar. All samples were prepared using modified procedures developed by Penkman *et al*.[37], but using a bleach time of 72 hours to isolate the intra-crystalline protein. Two subsamples were then taken from each tooth: one fraction was directly demineralised and the free amino acids analysed (referred to as the 'free' amino acids, FAA, F), and the second was treated to release the peptide-bound amino acids, thus yielding the 'total hydrolysable' amino acid fraction (THAA, H*). In addition to the procedure outlined previously[37], after demineralisation of the enamel, the pH of the solution was raised with KOH and then centrifuged for 5 min at 13,000 rpm, whereupon a biphasic solution formed. The supernatant was extracted and dried via centrifugal evaporation[40]. Samples were analysed in duplicate by RP-HPLC, with standards and blanks run alongside samples. During preparative hydrolysis, both asparagine (Asn) and glutamine (Gln) undergo rapid irreversible deamination to aspartic acid (Asp) and glutamic acid (Glu) respectively[41]. It is therefore not possible to distinguish between the acidic amino acids and their derivatives and they are reported together as Asx and Glx respectively.

The DL ratios of aspartic acid/asparagine, glutamic acid/glutamine, phenylalanine and alanine (D/L Asx, Glx, Phe and Ala) are assessed to provide an overall estimate of intra-crystalline protein decomposition (IcPD). In a closed system, the amino acid ratios of the FAA and the THAA subsamples should be highly correlated, enabling the recognition of compromised samples[42]. The D/L of all the amino acids will increase with time, but each amino acid racemises at a different rate, therefore providing different resolution over different timescales.

5. Proteomics

5.1. Sample collection and preparation

Anna K. Fotakis, Meaghan Mackie, Thomas W. Stafford Jr., Enrico Cappellini

All of the following procedures for sample preparation, up until nanoLC-MS/MS analysis, were conducted in laboratories dedicated to the analysis of ancient DNA, in clean rooms fitted with filtered ventilation and positive pressure, at the Natural History Museum of Denmark, University of Copenhagen. All work surfaces were UV irradiated overnight and repeatedly cleaned with bleach and ethanol, implementing all the measures necessary to avoid potential contamination from modern biomolecules[43]. Operators wore facemasks, nitrile gloves, hairnets and bodysuits continuously. A negative "extraction blank" control sample, not containing any mineralised material, was prepared, processed and analysed together with the ancient samples following the exact same procedure as with the samples.

Bone and tooth specimens selected for proteomic analysis were collected at the Georgian National Museum and moved to the University of Copenhagen after morphological assessment. The external surface of bone and dentine samples was gently removed using a rotating cutting disc connected to an electrically operated rotary tool. Approximately 1 g of the underlying compact bone and dentine material was then removed with a cutting disc and powdered either using a conventional hammer, after wrapping the samples in clean aluminium foil, or using a laboratory ball mill mechanical dismembrator (Braun Biotech). Fragments of enamel, occasionally mixed with dentine debris, were removed from teeth with a cutting disc. All bone, dentine and enamel specimens were stored in Protein Lo-Bind Tubes (Eppendorf) at -20˚C until further processing. Ancient protein residues were extracted from approximately 180-220 mg, if not otherwise specified, of bone and dentine powder, or dental enamel chips, using three different extraction protocols, hereinafter referred to as “A”, “B” and “C”, as described in detail below (Tab. S3).

Table S3. Specimens analysed. For each specimen, the Centre for GeoGenetics (CGG) reference number and the Georgian National Museum (GNM) specimen field number are reported, as well as the Stage-Tips number(s) identifying each protein extract. The type of mineral material extracted from each specimen, i.e. bone (BONE), dentine (DEN), or enamel+dentine (ENAM+DEN) is also indicated. NC: negative control, NA: not applicable.

CGG reference number	GNM specimen field number	Morphological identification	Stage-Tip number(s)_Extracted Material		
			Extr. method A	Extr. method B	Extr. method C
16486	Dm.bXI.sqA6.V._.	*Canis etruscus*			866_ENAM+DEN
16626	Dm.6/154.2/4.A4.17	Artiodactyla		620_BONE	
16628	Dm.7/154.2.A2.27	Cervidae		621_BONE	
16629	Dm.5/154.3.A4.32	Cervidae		622_BONE	867_ENAM+DEN
16630	Dm.6/151.4.A4.12	*Pseudodama nestii*		623_BONE	868_ENAM+DEN
16631	Dm.69/64.3.B1.53	Cervidae		624_BONE	869_ENAM+DEN
16632	Dm.5/154.2.A4.38	*Equus stenonis*			886_ENAM+DEN
16633	Dm.5/153.3.A2.33	*Equus stenonis*			888_BONE
16634	Dm.7/151.2.B1/A4.1	*Equus stenonis*			887_ENAM+DEN
16635	Dm.5/157.profile cleaning	*Stephanorhinus* sp.			870, 1213, 1217_ENAM+DEN
16636	Dm.6/153.1.A4.13	Rhinocerotidae		626_BONE	
16637	Dm.7/154.2.A4.8	Bovidae		627_BONE	
16638	Dm.5/154.1.B1.1	Bovidae		628_BONE	871_ENAM+DEN
16639	Dm.8/154.4.A4.22	Bovidae			872_ENAM+DEN
16640	Dm.6/151.2.A4.97	*Bison georgicus*		629_BONE	
16641	Dm.8/152.3.B1.2	*Bison georgicus*			873_ENAM+DEN
16642	Dm.8/153.4.A4.5	*Canis etruscus*			874_ENAM+DEN
16856	Dm.M6/7.II.296	Cervidae	230, 254, 272, 388, 390_DEN	700_ENAM+DEN	810, 816_ENAM+DEN
16857	Dm.bXI.profile cleaning	Indet.	231, 255, 256, 273, 274, 389, 391_BONE	701_BONE	811,817_BONE
16858	Dm.bXI.North.B1a.colleciton	Cervidae	232, 257, 280_BONE	702_BONE	812, 818_BONE
16859	D4.collection	Indet.	233_BONE	703_BONE	813, 819_BONE
16860	Dm.65/62.1.A1.collection	Cervidae		704_ENAM+DEN	814, 820_ENAM+DEN
16861	Dm.64/63.1.B1z.collection	*Equus stenonis*		705_ENAM+DEN	815, 821_ENAM+DEN
Neg. contr. (blank)	NA	NA	235, 251, 275, 392_NC	630, 706_NC	875, 889, 1214, 1218_NC

Extraction protocol A - FASP

Tryptic peptides were generated using a filter-aided sample preparation (FASP) approach[44,45]. Briefly, bone (_BONE) or dentine (_DEN) powder was demineralized by suspending it in 1 mL 0.5 M EDTA pH 8.0. After 24 h incubation with gentle shaking at 4˚C, the solution was centrifuged at 14000 ×g for 15 min in a benchtop centrifuge. The supernatant was discarded and the demineralization step was repeated one more time. The pellet was re-suspended in 300 µL of 4% SDS, 0.1 M dithiothreitol (DTT), 0.1 M Tris-HCl pH 8.0, heated at 80˚C for 2 h, then heated at 95˚C for 10 min with mechanical agitation. Once the suspension reached room temperature, it was centrifuged at 14000 ×g for 5 min in a benchtop centrifuge and the supernatant was then collected for further processing. The supernatant (SDS-fraction) was carefully mixed with 2 mL 8 M urea in 0.1 M Tris-HCl pH 8.0 and “washed”, i.e. ultrafiltered using Amicon Ultra-4 (Millipore, Billerica,

USA) centrifugal filter units with 10 kDa NMWL, at 4500 ×g in a benchtop centrifuge for the time necessary to remove all the solvent, except for what remained in the dead volume. FASP protocol continued by washing the fraction retained above ultrafilter with 2 mL 8 M urea, 0.1 M Tris-HCl pH 8.00. Alkylation was achieved by re-suspending in 500 µL of 50 mM 2-Chloroacetamide (CAA), 8 M urea, 0.1 M Tris-HCl pH 8.0 and incubating 20 min in the dark at room temperature. CAA was removed by washing with 1 mL of 8 M urea, 0.1 M Tris-HCl pH 8.0 for two times, then urea was removed by washing with 1 mL 50 mM ammonium bicarbonate, pH 7.5-8.0, for two times. The fraction retained above ultrafilter was re-suspended in 300 µL of 50 mM ammonium bicarbonate pH 7.5-8.0.

Protein digestion was started by adding 4 µL of 0.5 µg/µL sequencing-grade trypsin solution (Promega). After mixing, pH was checked using pH strips and, if necessary, adjusted to 7.5-8.0. The ultrafiltration units were incubated at 37˚C overnight. The following morning one supplemental microgram of fresh trypsin was added and the digestion was extended for 6 additional hours. Ultrafiltration units were then transferred into new sterile 15 mL tubes and centrifuged for 10 minutes at 4500 ×g. An additional aliquot of 500 µL 0.05 M ammonium bicarbonate pH 7.5-8.0 was added to the filter, vortexed, and centrifuged 4500 ×g to elute remaining peptides. The passed-through filtrate was then transferred to a sterile 1.5 mL Protein Lo-Bind Tube (Eppendorf) and acidified with 10% trifluoroacetic acid (TFA) to a final concentration of 0.2-0.8%, as necessary to reach a pH < 2.00.

Stage-Tips[46,47] were prepared in-house by pushing two disks, previously pierced with a 14 gauge blunt-end needle, out of a Empore C18 membrane (3M), at the end of each P200 micropipette tip. Stage-Tips were sequentially conditioned with 150 µL methanol, 150 µL 80% acetonitrile solution (80% acetonitrile, 0.1% trifluoroacetic acid (TFA), in ultrapure molecular biology-grade H_2O -v/v/v-), and 150 µL 0.1% TFA, in H_2O (v/v). The acidified peptides were then loaded onto the Stage-Tips and immobilized on the C18 filter by centrifugation at 1300 ×g for as long as necessary to pass all the solvent through the Stage-Tips. The filter was then washed with 150 µL 0.1% TFA, in H_2O (v/v), centrifuged 1300 ×g until dry, and stored at -20˚C. Just before LC-MS/MS analysis, Stage-Tips were transferred to the physically-separated MS/MS analysis facility at the Novo Nordisk Foundation Center for Protein Research (CPR), University of Copenhagen, where the tryptic peptides were eluted from the Stage-Tips membrane using 20 µL of 40% acetonitrile (ACN), 0.1% TFA in H_2O (v/v/v) and then 10 µL of 60% ACN, 0.1% TFA in H_2O (v/v/v), directly into a 96-well plate. The samples in the plate were concentrated for about 17 mins, in a centrifugal evaporator at 40˚C to a volume of approx. 2-5 µL, carefully avoiding to completely dry them, and re-suspended in 0.1% TFA, 5% acetonitrile to reach a final volume of 10 µL.

Extraction protocol B – GuHCl lysis solution and digestion

Bone(_BONE) or enamel+dentine (_ENAM+DEN) powder aliquots were placed in 2 mL Protein Lo-Bind Tubes (Eppendorf). Specimen samples, and accompanying negative control, were then suspended in 1 mL 0.5 M EDTA, pH 8.0 (Ambion P/N AM8262), mechanically shaken for approximately one minute and incubated overnight at room temperature under mechanical agitation. The following day the samples were centrifuged at 17000 ×g for 10 mins and the supernatant was removed and discarded. One additional millilitre of EDTA was added and the step was repeated one more time. The third day, after removal of EDTA supernatant, all demineralised pellets were re-suspended in 100 µL of 0.1 M Tris pH 8.0, mechanically shaken for approximately one minute, and precipitated by centrifugation at 17000 ×g for 10 mins. The supernatant was then removed and discarded. The wash step with 100 µL of 0.1 M Tris pH 8.0 was repeated two more times. The pelleted samples were then re-suspended in a 300 µL solution containing 2 M guanidine hydrochloride (GuHCl, Thermo Scientific), 100 mM Tris pH 8.0, 20mM 2-Chloroacetamide (CAA), 10 mM Tris (2-carboxyethyl)phosphine (TCEP) in ultrapure H_2O[48,49]. The pH was measured using pH strips and, if necessary, adjusted to a value between 7.5 and 8.5. Samples were then incubated in a heat block for 2 hours at 80˚C, then heated at 95˚C for 10 min with mechanical shaking at 800 rpm. After samples reached room temperature, 0.2 µg of mass spectrometry-grade rLysC (Promega P/N V1671) enzyme was added to each sample. Samples

were then incubated for 3-4 hours at 37˚C under agitation. Samples and negative controls were subsquently diluted to 0.6 M GuHCl final concentration with a solution containing 10% acetonitrile (ACN), 25 mM Tris pH 8.0, in ultrapure water. Afterwards, the pH was checked using pH strips, if necessary adjusted to fall between 7.5 and 8.5, and 0.8 µg of mass spectrometry-grade Trypsin (Promega P/N V5111) were added. Samples and negative controls were then incubated overnight under mechanical agitation at 37˚C. On the following day, samples were acidified, using 10% TFA in ultrapure water, to reach pH < 2.0, and then centrifuged at 17000 ×g for 1 hour to eventually precipitate remaining debris. The resulting peptide mixtures in the supernatant fractions were collected, cleaned, concentrated and immobilised using in-house prepared C18 solid-phase extraction Stage-Tips, as previously described in the paragraph: “Extraction protocol A”.

Extraction protocol C - digestion-free HCl/TFA demineralisation

Approximately 250 mg of dental enamel powder, with possible trace amounts of dentine, (_ENAM+DEN) or bone powder (_BONE) were placed in a 15 mL sterile tube, suspended in 1 mL of 1.2 M HCl, and demineralised overnight at room temperature under mechanical shaking. The following day, after centrifugation at 4500 ×g for 15 min, the HCl supernatant was collected and stored at -18˚C. The demineralisation step was repeated one more time following the same procedure. The third day the samples were again centrifuged at 4500 ×g for 15min and, for each sample, the HCl supernatant was collected and merged with the HCl supernatant fraction stored after the first demineralisation round. The pH of merged supernatant fractions for each sample and accompanying negative control samples were measured, using pH strips, to confirm it was lower than 2. Peptides suspended in the HCl demineralisation fraction were then directly cleaned, concentrated and immobilised, skipping reduction, alkylation and digestion steps, using in house assembled C18 solid-phase extraction Stage-Tips, as described in the paragraph “Extraction protocol A”. Any remaining pellets were stored at -18˚C. Unlike the other extracts processed using this method, the extract identified with Stage-Tip number “1217” was demineralised using 10% TFA, instead of 1.2 M HCl, overnight at 4˚C, and not at room temperature. All the other experimental parameters and procedures were identical to those used for all the other samples extracted using protocol C.

5.2. LC-MS/MS analysis

Rosa Rakownikow Jersie-Christensen, Meaghan Mackie, Jesper V. Olsen

Q Exactive Classic

Samples (Stage-Tips n. 230-235, 272-275 and 280, Table S3) were analyzed by nanoflow liquid chromatography tandem mass spectrometry (nanoLC-MS/MS) using a 50 cm fused silica column (75 µm inner diameter) in-house pulled and packed with 1.9 µm C18 beads (Reprosil-AQ Pur, Dr. Maisch) on an EASY-nLC™ 1000 system connected to a Q-Exactive (Thermo Scientific, Bremen, Germany) mass spectrometer. The peptide mixtures were separated with a linear gradient from 5% to 25% buffer B in 110 minutes followed by a 15 minutes step to 40% buffer B. At the end of the gradient the column was washed by increasing buffer B content to 80% in 5 minutes and keeping it there for 5 min before re-equilibrating the column back to 5% buffer B for 5 minutes, resulting in a total acquisition time of 145 min. Buffer B consisted of 80% acetonitrile and buffer A of milliQ water, both containing 0.1% TFA. The flow rate was maintained at 250 nL/min throughout the gradient, wash, and equilibration.

The Q-Exactive was operated in data-dependent acquisition mode using a ‘top 10’ method. Full scan MS were recorded at resolution of 70,000 at m/z 200 in a mass range of 300–1750 m/z with a target value of 1e6 and a maximum injection (IT) time of 20 ms. Higher-energy Collisional Dissociation (HCD) fragment MS/MS spectra were recorded with a maximum ion injection time set to 108 ms with a target value of 2e5 and recorded at a resolution of 35,000 with a fixed first mass set to 100 m/z. Normalized collision energy (NCE) was 25%. The isolation window was set at 2.5 m/z and the dynamic exclusion of former target ions for MS/MS to 30 seconds.

Q Exactive Plus

Samples (Stage-Tips n. 251 and 257, Table S3) were analyzed by nanoflow LC-MS/MS using a 50 cm PicoFrit® column (75 µm inner diameter) in-house packed with 1.9 µm C18 beads (Reprosil-AQ Pur, Dr. Maisch) on an EASY-nLC™ 1000 system connected to a Q-Exactive Plus (Thermo Scientific, Bremen, Germany) mass spectrometer. The peptides were separated with a gradient going from 2% to 25% buffer B in 110 minutes followed by a 25 minute step to 40%. After the gradient, the column was washed by increasing buffer B content to 60% in 5 minutes and remaining there for 5 min and re-equilibrated back to 2% for 15 minutes resulting in a final acquisition time of 165 min. Buffer B consisted of 80% acetonitrile and buffer A of deionized water, both containing 0.1% TFA. The flow rate was set to 200 nL/min throughout the gradient, wash and equilibration.

The Q-Exactive Plus was operated in data-dependent 'top 12' mode. Full scan MS were recorded at resolution of 70,000 at m/z 200 in a mass range of 300–1750 m/z with a target value of 3e6 and a maximum injection time of 20 ms. Fragment HCD-MS/MS were recorded with a maximum ion injection time set to 108 ms with a target value set to 1e5 and recorded at a resolution of 35,000 with a fixed first mass set to 100 m/z. Normalized collision energy was 28%. The isolation window was set at 1.3 m/z and the dynamic exclusion to 30 seconds.

Q Exactive HF

All remaining samples were analyzed by nanoflow LC-MS/MS using a 50 cm PicoFrit® column (75 µm inner diameter) in-house packed with 1.9 µm C18 beads (Reprosil-AQ Pur, Dr. Maisch) on an EASY-nLC™ 1000/EASY-nLC™ 1200 system connected to a Q-Exactive HF (Thermo Scientific, Bremen, Germany) mass spectrometer. The peptides were separated with a linear gradient from 2% to 25% buffer B in 110 minutes followed by a 25 minute step to 40%. After the gradient, the column was washed by increasing buffer B to 60% in 5 minutes and remaining a this concentration for 5 min before re-equilibration back to 2% for 15 minutes resulting in a final total acquisition time of 165 min. Buffer B consisted of 80% acetonitrile and buffer A of deionized water, both containing 0.1% TFA. The gradient flow rate was 200 nL/min throughout the gradient and wash.

The Q-Exactive HF was operated in data-dependent acquisition mode using a 'top 10' method. Briefly, full scan MS were recorded at resolution of 120,000 at m/z 200 in a mass range of 300–1750 m/z with a target value of 3e6 and a maximum injection time of 20 ms. Fragment HCD-MS/MS were acquired with a maximum ion injection time set to 108 ms with a target value of 2e5 and recorded at a resolution of 60,000 with a fixed first mass set to 100 m/z. Normalized collision energy was 28%. The isolation window was set at 1.3 m/z and the dynamic exclusion to 30 seconds.

MS/MS Wash and Blank Runs

Before and after each MS/MS run acquiring data for ancient or extraction blank samples, an MS/MS run with no sample intake ("MS/MS wash" run), followed by another MS/MS run with an intake exclusively of the buffer (0.1% TFA, 5% ACN) used to re-suspend the samples ("MS/MS blank" run) were added. These runs were included in each data acquisition session in order to monitor and prevent potential carryover contamination of downstream samples in case of incomplete purge of the nanoLC column after each sample MS/MS run. The MS and nanoLC operation parameters for these runs can be found in the tables below (Tabs. S4, S5).

Table S4. Mass spectrometer settings for the wash and blank samples on the Q-Exactive (QE) Classic, QE Plus, and QE HF. IT: injection time and NCE: normalized collision energy.

		Wash			Blank		
		QE Classic	**QE Plus**	**QE HF**	**QE Classic**	**QE Plus**	**QE HF**
Method		Top 10	Top 3	Top 5	Top 10	Top 3	Top 10

Full Scan	Resolution	70,000 at m/z 200	70,000 at m/z 200	120, 000 at m/z 200	70,000 at m/z 200	70,000 at m/z 200	120, 000 at m/z 200
	Mass Range	m/z 300-1750	m/z 350-1400	m/z 200-2000	m/z 300-1750	m/z 350-1400	m/z 200-2000
	Target Value	1e6	3e6	3e6	1e6	3e6	3e6
	Max IT	60 ms	20 ms	100 ms	60 ms	20 ms	100 ms
Fragment Scan	Resolution	35,000	17,500	30,000	35,000	17,500	60,000
	Fixed first mass	100 m/z	100 m/z	—	100 m/z	100 m/z	100 m/z
	Target Value	2e5	1e5	2e5	2e5	1e5	1e5
	Max IT	108 ms	45 ms	50 ms	108 ms	45 ms	60 ms
	NCE	25%	28%	25%	25%	28%	28%
	Isolation Window	4.0 m/z	4.0 m/z	2.0 m/z	4.0 m/z	4.0 m/z	1.3 m/z
	Dynamic exclusion	40 s	10 s	10 s	40 s	10 s	20 s

Table S5. nLC settings for the wash and blank samples on the Q-Exactive Classic, Q-Exactive Plus, and Q-Exactive HF.

Run Type	Instrument	Injection Volume (µL)	Flow rate (nL/min)	Time (mm:ss)	Duration (min, from Time)	% Buffer B
Wash	QE Classic	0	250	00:00	0	30
				05:00	5	95
				10:00	5	5
				15:00	5	95
				20:00	5	5
				25:00	5	5
	QE Plus	0	200	00:00	0	5
				05:00	5	80
				07:00	2	80
				09:00	2	30
				11:00	2	30
				13:00	2	80
				15:00	2	80
				20:00	5	5
				30:00	10	5
	QE HF	0	200	00:00	0	5
				05:00	5	90
				07:00	2	90
				09:00	2	90
				11:00	2	90
				16:00	5	5
				26:00	10	5
Blank	QE Classic	2	250	00:00	0	5
				15:00	15	30
				25:00	10	80

				30:00	5	80
				35:00	5	5
				40:00	5	5
	QE Plus	2	200	00:00	0	5
				20:00	20	30
				30:00	10	80
				32:00	2	80
				37:00	5	5
				42:00	5	5
	QE HF	2	200	00:00	0	5
				10:00	10	30
				15:00	5	80
				17:00	2	80
				19:00	2	5
				24:00	5	5

5.3. Peptide-spectrum matching

Enrico Cappellini, Frido Welker

Raw files generated during MS/MS spectral acquisition were searched on a workstation using MaxQuant[50], version 1.5.3.30, and the commercial tool PEAKS[51], version 7.5. A two-stage peptide-spectrum matching approach was adopted (Fig. S2). At the end of the first, “preliminary”, stage, enamel proteins and collagen were identified and their sequences were partially reconstructed, enabling, in the most favourable cases, an initial taxonomic placement of the ancient reconstructed sequences. These results were used to create a smaller protein database, which was used in a second, “advanced”, stage of peptide-spectrum matching, to extend the sequence coverage of the ancient proteins previously identified.

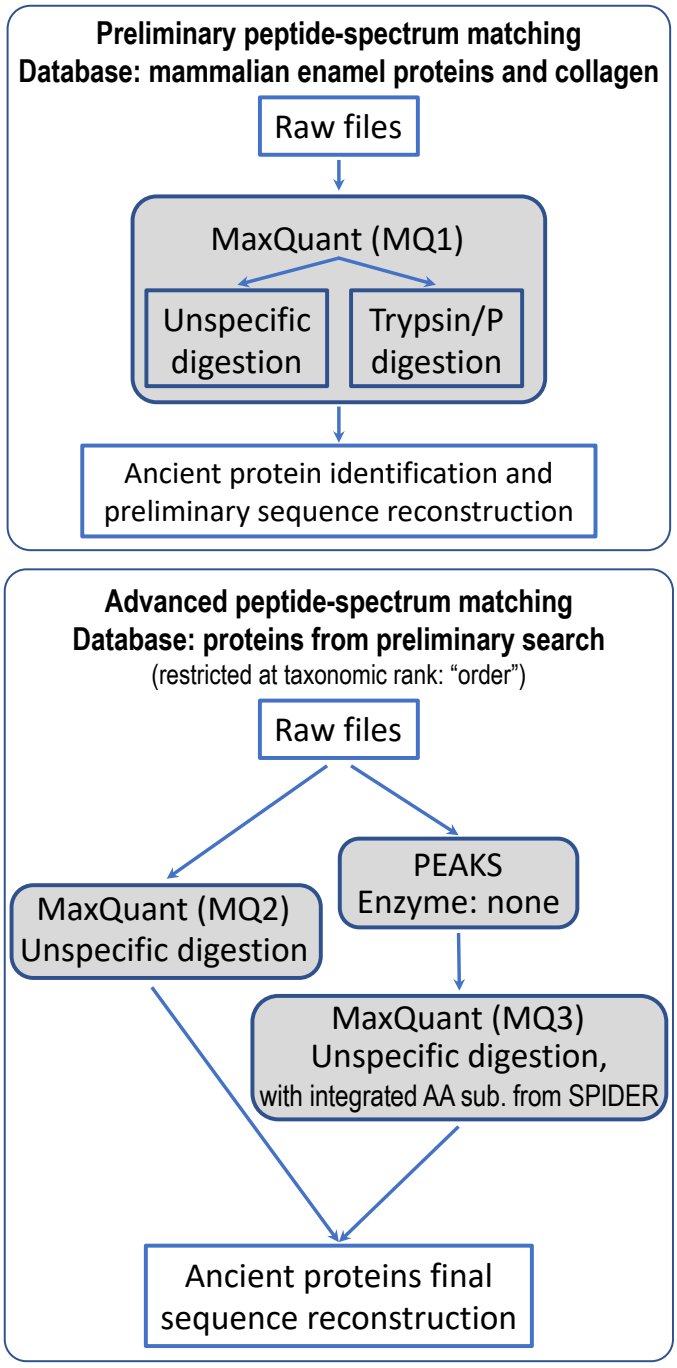

Figure S2. Schematic representation of the workflow used for peptide-spectrum matching and ancient protein sequence reconstruction.

MaxQuant preliminary (MQ1) peptide-spectrum matching

Raw files were initially searched against a target/reverse database of collagen and enamel proteins retrieved between 15th February and 1st March 2016 from the UniProt and NCBI Reference Sequence Database (RefSeq) archives, taxonomically restricted to mammalian species. Enamel proteins were extracted from the UniProt archive using the keywords: "AMTN", "AMELX", "AMELY", "AMBN", "ENAM", "DSPP", "MMP20", and "KLK4", in the "gene name" field, as well as "tuftelin" and "serum albumin" in the "protein name" field. Collagen sequences were retrieved from the RefSeq filtering for the keywords: "COL1A1", "COL1A2", "COL2A1", and "COL3A1", in the "gene name" field. A database of partial "COL1A1" and "COL1A2" sequences from cervid species

was also included[52]. The “contaminants” file distributed with the MaxQuant version used was included in each search.

The minimum Andromeda score for unmodified and modified peptides was set to 40. The false discovery rate (FDR) for peptide-spectrum matches (PSMs) and for proteins was set to 1%. In the main search, the peptide tolerance was set to 5 ppm, and the MS/MS match tolerance was set to 20 ppm. All the data generated from each animal specimen, independently from the number of fractions generated and the extraction protocols applied, were processed in a single search. For all the animal specimens analysed, the parameter “digestion mode” was set to “unspecific”, and the following variable modifications were included: oxidation (M), deamidation (NQ), N-term Pyro-Glu (Q), N-term Pyro-Glu (E), hydroxylation (P), phosphorylation (S). Carbamidomethylation (C) was set: (i) as a fixed modification, for searches of data generated from sets of sample fractions exclusively digested with trypsin, or (ii) as a variable modification, for searches of data generated from sets of sample fractions partially digested with trypsin. For searches of data generated from sample fractions that were not digested, carbamidomethylation (C) was not included as a modification, neither fixed, nor variable. For searches of data generated from sample fractions partially or exclusively digested with trypsin, i.e. using extraction protocols A and B, another MaxQuant and PEAKS search were conducted using the “enzyme” parameter set to “Trypsin/P”, allowing a maximum of 2 missed cleavages.

MaxQuant advanced (MQ2) peptide-spectrum matching

The results from the preliminary analysis were used for a first, provisional reconstruction of protein sequences. For specimens whose dataset showed a narrower, though not fully resolved, initial taxonomic placement, a second MaxQuant search (MQ2) was performed, in unspecific digestion mode, using the same parameters as previously described apart from searching against a different, smaller protein database. The new protein database was generated by extracting from the RefSeq archive the sequences matching the “gene name” and the “protein name” fields of all the enamel and collagen proteins identified in the previous search. Duplicates within the same species and partial, shorter, sequences were manually removed.

Finally, for the advanced MaxQuant matching of the MS/MS spectra generated from the animal specimen with reference number Dm.5/157-16635, attributed by morphological analysis to *Stephanorhinus* sp. (Supplementary Information §3), partial sequences of serum albumin and enamel proteins: “ALB”, “AMELX”, “AMBN”, “AMTN”, “ENAM”, and “MMP20” from Sumatran rhinoceros (*Dicerorhinus sumatrensis*), Javan rhinoceros (*Rhinoceros sondaicus*), Indian rhinoceros (*Rhinoceros unicornis*), woolly rhinoceros (*Coelodonta antiquitatis*), and Black rhinoceros (*Diceros bicornis*), were also added to the protein database searched. All the protein sequences from these species were reconstructed from unpublished DNA reads generated by high-throughput full-genome sequencing (Supplementary Information §7).

PEAKS peptide-spectrum matching

The datasets re-analysed with MQ2 “advanced” search, were also processed with the PEAKS software using the entire workflow (PEAKS de novo to PEAKS SPIDER) to detect novel single amino acid polymorphisms (SAPs) not present in the reference databases used. The error tolerance was set to 5 ppm for the precursor and to 0.05 Da for the fragment ion, respectively. Independent of the extraction protocol, the mass spectrometry data of each animal specimen was processed in a single search. The parameter “Enzyme” was set to “None”, and the following variable modifications were included: oxidation (M), deamidation (NQ), N-term Pyro-Glu (Q), N-term Pyro-Glu (E), hydroxylation (P), phosphorylation (S), and carbamidomethylation (C). No fixed modifications were included, and up to 5 PTMs per peptide match were allowed. Data processing with the “Peaks PTM” and “SPIDER” algorithms was included in the workflow. Filtering was conducted using a false discovery rate (FDR) for peptides of 1% and a minimum score for proteins of -10 lgP ≥ 20. De novo ALC% was set to ≥ 95. For all the animal specimens processed, partially or exclusively, with extraction protocols A and B, i.e. with trypsin digestion, the PEAKS search was refined by searching the previously unassigned spectra once more. This time the “Enzyme”

parameter was set to “Trypsin”, modified to allow cleavage even with P as the amino acid trailing the cleavage site, and allowing a maximum of 2 missed cleavages. The following variable modifications were included: oxidation (M), deamidation (NQ), N-term Pyro-Glu (Q), N-term Pyro-Glu (E), hydroxylation (P), and phosphorylation (S). Carbamidomethylation (C) was included as a fixed modification.

The raw files were matched against the same database used in the MQ2 “advanced” search. Any amino acid substitution detected by the “SPIDER” homology search algorithm was validated by repeating the MaxQuant “advanced” search (MQ3), while taking into account the PEAKS guidelines for SAP acceptance proposed elsewhere[53]. The amino acid substitutions detected by the “SPIDER” algorithm were included in the taxonomically restricted protein database used in MQ2. All the other search parameters were not modified.

5.4. Ancient protein sequence reconstruction

Enrico Cappellini, Diana Samodova, Frido Welker, Christian Kelstrup

The peptide sequences confidently identified by the MaxQuant and PEAKS searches were aligned and a consensus sequence for each protein from each specimen was generated in FASTA format. Rare inconsistencies between ancient sequences reconstructed with MaxQuant and PEAKS were resolved by manually inspecting the data, in particular when dealing with novel protein sequences suggested by PEAKS and confirmed by MaxQuant[53].

Sequence reconstruction from MaxQuant output files

For each animal specimen, after removing contaminant and reverse sequences, all the peptide sequences confidently identified and manually inspected were exported from the “peptides.txt” file and converted to FASTA format. For each protein, peptides were then aligned to the sequence of the mammalian species that presented the highest coverage. The alignment was obtained by using the program Geneious[54] version 5.4.4, selecting substitution matrix BLOSUM62, gap open penalty 12 and gap extension penalty 3. No filtering based on depth of coverage was applied. The ancient sequences reconstructed were manually inspected and exported to FASTA format. Amino acid positions that were not confidently reconstructed were indicated with “X”. The output of the “advanced” MaxQuant peptide-spectrum matching (MQ2 and 3) was used to extend the coverage of the ancient protein sequences initially identified in the “preliminary” MaxQuant iteration (MQ1).

Sequence reconstruction from PEAKS output files

After removing contaminants, all the peptide sequences confidently identified by PEAKS for each animal specimen were exported as “Supporting peptides” text files (protein-peptides.csv). For each protein, peptides were aligned to the sequence of the mammalian species that presented the highest coverage, using an in-house developed script. No filtering based on depth of coverage was applied. The ancient sequences reconstructed were manually inspected and exported in FASTA format. Amino acid positions that were not confidently reconstructed were indicated with “X”. The output of the “advanced” PEAKS peptide-spectrum matching was used to extend the coverage of the ancient protein sequences initially identified in the “preliminary” MaxQuant iteration. Finally, the amino acidic substitutions detected by the “SPIDER” algorithm, in the “advanced” PEAKS peptide-spectrum matching, were included, after validation through the final MaxQuant search (MQ3), in the final version of the ancient protein sequence reconstruction.

Spectral and peptide validation

To validate the identification of the phylogenetically relevant protein sequences, we used a combination of three different approaches applied to the *Stephanorhinus* specimen Dm.5/157-16635.

First, we show that our phylogenetically relevant positions are covered by multiple PSMs and peptides (Extended Data Fig. 2). The same is true for the phosphorylated peptides (Extended Data Fig. 2).

Second, we show that we have near-complete fragment ion series spanning around these phylogenetically informative positions (see document “Key MS-MS Spectra.pdf”). This also includes validation by manual sequence reconstruction for a subset of MS2 spectra, showing that manual and automatic approaches result in the same phylogenetic information.

Third, we used MS^{2}PIP[55] to predict the relative intensity of b and y fragment ions in the spectra supporting the identification of a relevant set of peptides. As the currently available MS^{2}PIP models are trained using tryptic peptides only, we had to restrict the application of MS^{2}PIP to the 14 peptides, with C-terminal lysine (K) or arginine (R), covering phylogenetically informative SAPs. Similarly, we could apply MS^{2}PIP only to the single peptide with C-terminal arginine (R) covering one phosphorylation but no additional PTMs. We then compared the MS^{2}PIP-predicted relative fragment ion intensities for this peptide subset with the corresponding ones from the measured MS2 spectra attributed by MaxQuant to the same set (Figure S3a). This shows a high, positive Pearson correlation score (0.848) indication good agreement between the predicted and observed intensities. This positive correlation between predicted and observed intensities is also present for all peptides individually (Figure S3b), with a distribution similar to the one observed during MS^{2}PIP benchmarking for HCD datasets[55].

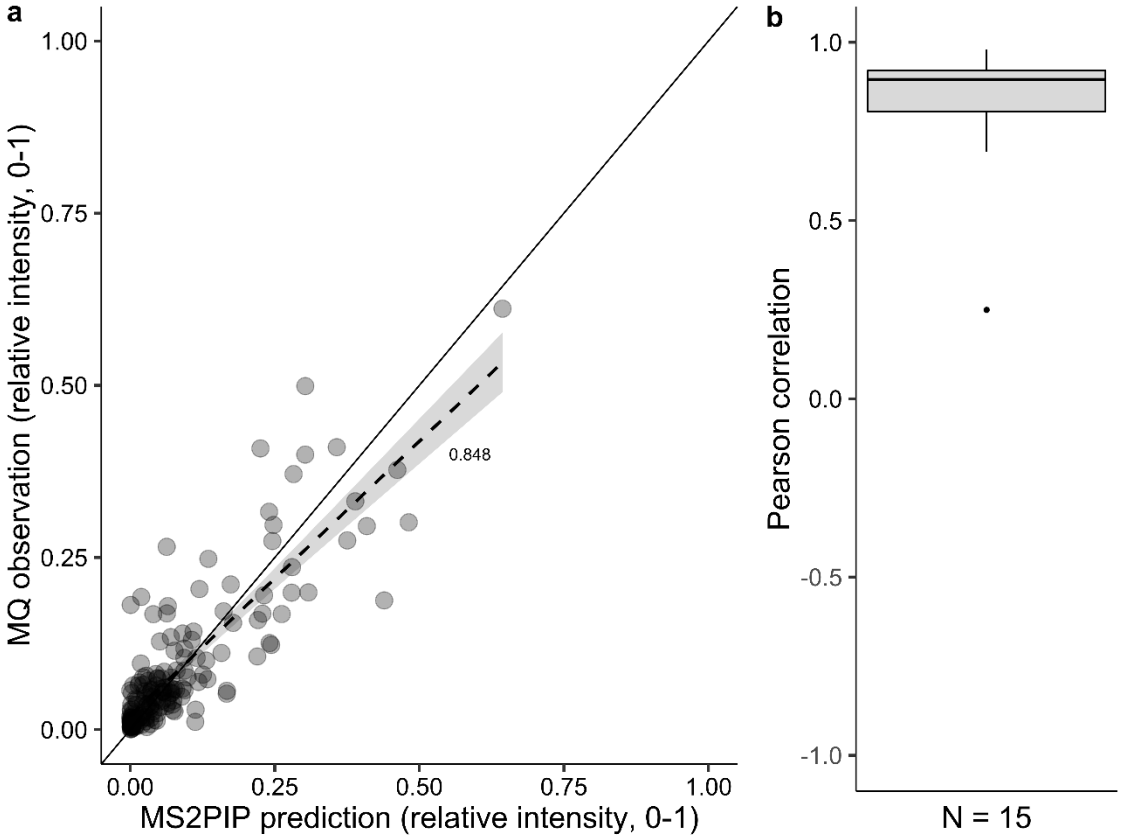

Figure S3. MS^{2}PIP fragment ion intensity prediction. a) Correlation between the predicted intensity by MS^{2}PIP and observed intensity by MaxQuant for 250 fragment ions deriving from 15 “pseudo-tryptic” peptides. Dashed line: linear relationship between predicted and observed intensity. Solid line: x = y. **b)** Boxplot of Pearson correlation observed for each of the 15 peptides showing a positive, highly significant correlation.

Fourth, we analysed 19 synthetic peptides covering phylogenetically-relevant SAPs, and three synthetic peptides containing a single phosphorylation. The synthetic peptides were ordered from JPT Peptide Technologies (Berlin, Germany), solubilized in 5% acetonitrile (ACN)/0.1% trifluoroacetic acid (TFA), pooled together, purified by loading on C18 material (Stage-Tips), and analysed on a Q-Exactive HF-X.

The peptide mixture was eluted from each Stage-Tip, using 50 µL of 40% ACN/0.1% FA into a LoBind Eppendorf tube. The sample was placed in a vacuum centrifuge at 45°C for 18 min to evaporate ACN (30 µL final volume). Peptide concentration was thereafter measured by NanoDrop spectrcophotometer (ThermoFisher Scientific, Waltham, USA) at a wavelength of 280 nm. 12 ng of the peptides were then separated on a 15 cm column (75 µm inner diameter) in-house laser pulled and packed with 1.9 µm C18 beads (Dr. Maisch, Germany) on an EASY-nLC 1200 (Proxeon, Odense, Denmark) connected to a Q-Exactive HF-X (Thermo Scientific, Bremen, Germany) on a 30.5 min gradient. Buffer A was 0.1% FA in milliQ water. The peptides were separated, by increasing buffer B concentration (80% ACN and 0.1% FA) from 10% to 30% in 25 min, 30% to 45% in 5 min, 45% to 80% in 0.5 min, and held at 80% for 3.5 min respectively. The mobile phase flow rate was 350 nL/min. The column temperature was maintained at 40°C, using an integrated column oven.

The Q-Exactive HF-X mass spectrometer was operated in parallel reaction monitoring (PRM) Top11 mode. The inclusion list, containing mass-over-charge (m/z) ratios, charge and normalized collision energy (NCE), was utilised for identification and fragmentation of the relevant peptide precursors. Spray voltage was 2 kV, S-lens RF level at 40, and heated capillary at 275°C. Full scan mass spectra were recorded at a resolution of 60,000 at m/z 200 over the m/z range 350–1000 with a target value of 5e5 and a maximum injection time of 118 ms. HCD-generated product ions were recorded with a maximum ion injection time set to 118 ms and a target value set to 5e5 at a resolution of 60,000. Normalized collision energy was set to 28% and the isolation window was 0.7 m/z.

The raw file containing synthetic peptide MS data, was manually inspected in order to, firstly, validate the purity of prepared peptide mixture, and, secondly, to select the high-quality MS2 scans of the corresponding synthetic precursors. The mass-over-charge values of synthetic peptide precursor ions were extracted from the total ion chromatogram (TIC), except for the peptide HAPPDNPP (m/z = 385.6827, z = 2), the synthesis of which was not successful. This extraction procedure allowed to assess the yield of the synthesis and to evaluate purity of the peptides. Finally, high-quality MS2 scans were selected from the apex point of precursor peaks for further alignment with the experimental dataset.

The high-quality MS2 spectra, selected during manual validation, were aligned to the corresponding experimented peptide spectra, using the Interactive Peptide Spectral Annotator (IPSA)[56] freeware. The synthetic peptides-containing raw file was first converted to mgf format, using the MSConvert tool, available within the ProteoWizard cross-platform toolkit[57], and uploaded to the tool server, as a bulk. The experimental and synthetic peptide MS2 spectra were thereby annotated and overlaid, in order to prove the authenticity of the ancient peptide sequences. These are available via the PRIDE repository with the data set identifier PXD011008.

The obtained results indicate almost full correspondence between the automated, manual, and synthetic peptide sequence reconstructions, except for the peptide HAPPDNPP (m/z = 385.6827, z = 2), the synthesis of which was not successful. The obtained results show good overall correspondence between the synthetic and the ancient peptide MS2 spectra (See “Key MS2 Spectra.pdf” SI file). This is true for both the synthetic peptides overlapping SAPs of relevance, and for the phosphopeptides. It is worth stressing that both the synthetic and ancient peptide MS2 spectra contain some high-intensity fragment ions, not identified as b- or y- ions in the corresponding peptide sequences. This could probably be explained by the co-fragmentation of impurities not present in the synthetic peptides. Based on these four independent lines of evidence, we conclude that our MS spectral interpretation is reliable and conforms to the highest standards in the field.

5.5. Post Translational modifications

Phosphorylation

David Lyon, Rosa Rakownikow Jersie-Christensen, Christian D. Kelstrup, Jesper V. Olsen

All raw LC-MS/MS files were analysed and processed together in a single search as separate experiments using MaxQuant[50] version 1.5.3.36 without matching between runs and using unspecific enzymatic digestion mode. Matched variable modifications were oxidation (M), deamidation (NQ), Gln->pyro-Glu, Glu->pyro-Glu, carbamidomethylation (C), oxidation (P), and phosphorylation (ST). Peptide-spectrum matches (PSMs) were identified using the Andromeda search engine by searching de-isotoped MS/MS spectra against a forward and reversed version of a combined FASTA protein database consisting of mammalian protein sequences of enamel proteins, as previously described. The resulting peptide dataset was filtered based on the forward and reversed peptide identifications using the posterior error probability (PEP) values applying a 1% false discovery rate (FDR) at both peptide and protein level and requiring a minimum Andromeda score of 40 for all peptide identifications. After removing contaminants, reversed hits and phosphorylation site entries with site probabilities lower than 0.98 were filtered out from the Phospho(ST)Sites output file. The remaining unique phosphorylation site sequence windows (±6aa) were used to perform linear kinase sequence motif enrichment analysis with the IceLogo software tool[58] using all serine/threonine residues in the detected non-phosphorylated peptides as background.

Deamidation

David Lyon

Deamidation was calculated per proteomic dataset as detailed in Mackie et al.[59]. In short, MaxQuant's "proteinGroups" file was filtered to remove contaminants and identifications matching reversed protein sequences. The "Evidence IDs" of the remaining protein entries were mapped to the "evidence" file in order to select the relevant entries. This was done for results of both MaxQuant searches, performed with unspecific and tryptic cleavage. These filtered "evidence" files were merged, and duplicates removed, thereby combining the results of different search strategies per sample. Peptide level deamidation rates were calculated separately for asparagine and glutamine, taking relative abundance values of all the PSMs (peptide-spectrum-matches) into account, resulting in a deamidation rate between 0 and 1 for each unique peptide sequence and charge state. The various charge states were collapsed to the peptide level by calculating the mean, resulting in a separate deamidation rate for asparagine as well as glutamine, respectively. Specimens for which either glutamine or asparagine deamidation values were calculated from less than 20 peptides were filtered out. In order to gain an estimate of the error of the calculation, the peptide level deamidation rates were sampled with replacement, i.e. bootstrapped, 1000 times and therefrom the mean and the standard deviation calculated. Generally, the higher the number of peptides detected with the potential to be deamidated, the more accurate the measurement and thus the lower the error will be. Nevertheless, the method is intended for comparative purposes and not as an absolute measure of deamidation. The method is freely available to the scientific community at: https://github.com/dblyon/deamidation, and further details are available in Mackie et al.[59].

Other forms of ancient protein damage

Diana Samodova, Patrick L. Rüther

Determination of PTMs with PEAKS software

The initial analysis of damage-related chemical modifications other than deamidation in ancient proteins was conducted using the PEAKS PTM tool[60], which performs a "blind" search for all the PTMs listed in Unimod eventually present in the proteins previously identified. The PEAKS PTM search parameters were set as follows: parent mass error tolerance was set to 5.0 ppm and fragment mass error to 0.05 Da; De novo ALC score threshold was set to 15% and peptide hit threshold (-10logP) to 30; enzyme was set to none and FDR estimation was enabled. The number of max variable PTMs per peptide was set to 2. After the search was completed, the obtained results were filtered using the following criteria: 5% false discovery rate (FDR), De novo ALC score ≥ 50%, protein hit threshold (-10logP) ≥ 20 and unique peptides ≥ 0. In order to quantify different

modifications for each sample, i.e. to determine the number of peptide-spectrum matches for each modification, the following extraction procedure was applied to each raw file:

1. Export “protein-peptides.csv” output file from PEAKS PTM directory;
2. Sort by “-10logP” value in decreasing order;
3. For each amino acid and PTM, filter and retain occurrences with “AScore” values above threshold;
4. Count peptide-spectrum matches (PSMs), related to each modified amino acid, by summing up values from “#Spec” column;
5. Remove all modifications per amino acid with a total of less than 10 PSMs;
6. Update the list of obtained modifications by excluding those that are either not related to aging processes (e.g. esterification, acetylation, alkylation, carbamylation, guanidination), or those that have been used as variable or fixed modifications in the PEAKS DB search.

Determination of PTMs with MaxQuant

The analysis of degradation-related PTMs, excluding deamidation was divided in two parts. The goal of the first search was unbiased discovery, i.e. assessing the presence of as many different PTMs as possible. In order to do this, the “dependent peptides” algorithm of MaxQuant[61], which searches for similarities between identified “base peptides” (unmodified or carrying a specified variable modification) and non-identified fragmentation spectra, was used. For similar fragmentation patterns, the mass shift to the base peptide is calculated and located in the base peptide sequence by matching equal and shifted masses to the predicted fragment masses. Relevant modifications were identified by ranking the mass-shifts by occurrence (after reverse hits and contaminants were removed) and then counting how often the dependent peptides contained each mass shift at each amino acid. Secondary searches were then carried out using variable modifications in order to map selected PTMs on all identified peptides for better coverage and comparability between samples. To identify PTMs caused by long-term protein degradation in the range of one million years, we analysed the raw-files together with enamel from a medieval goat/sheep, which had also been prepared by acidic peptide extraction and without trypsin digest, as a reference sample.

Highly redundant overlapping peptide sequences make intensity-based analysis of PTMs very complex. Therefore, the extent of modification was estimated by “relative PSM counts”. This is the number of peptide spectrum matches containing the PTM of interest divided by the total number of PSMs per raw-file. This approach effectively compensated for the large differences in sequencing depth limited by peptide concentration and degradation. Furthermore, with the relatively small number of modified sites, MaxQuant’s own FDR control at the modification-site level removes many sites, despite the solid coverage and matching spectra. Therefore, the database searches were carried out without a cutoff, which was then manually adjusted to 5% using the results of the decoy search.

Workflow:

1. Extraction of data from “evidence.txt”
2. Sorting by score (decreasing order)
3. Calculation of FDR (cumulative number of reverse hits divided by rank)
4. Removal of PSMs above FDR threshold (5%), reverse hits, and contaminants
5. Counting PSMs by summing “MS/MS count” column per modification

Calculating relative PSMs by dividing with the total number of PSMs from a raw file

The results represent the percentage of PSMs with a certain PTM, in a sample. Since this value takes the total number of PSMs into account, it is comparable between samples with different peptide abundance and FDR cutoff. Comparison of relative PSM counts can only be made if the same amino acid is affected. To test this method, relative PSM counts for deamidation, Pro-, and Met-oxidation were included in all MaxQuant searches and showed consistency across samples of similar age.

6. Dental enamel proteome from Medieval remains

Enrico Cappellini, Anna K. Fotakis, Meaghan Mackie

Ancient proteins from a medieval ovicaprine (*Capra hircus*/*Ovis aries*) tooth, originally recovered from the archaeological site called "Hotel Skandinavia" in Århus, Denmark, were analysed. The specimen was provided by the Natural History Museum of Denmark, and is approximately 500-1000 years old, based on stratigraphy and archaeological context. Enamel proteins were solubilised using extraction protocol C, and MS data were generated using the "Q Exactive HF" setup previously described. The Stage-Tip numbers assigned to the medieval ovicaprine control specimen and to the relative extraction blank are: 1036 and 1038, respectively. Spectra from the two raw files were matched, using MaxQuant, against the complete goat proteome, downloaded on May 31st 2017 from the RefSeq archive, taxonomically restricted to "*Capra hircus*" (42,702 entries). The same settings previously described when using "unspecific" digestion mode were used. Carbamidomethylation (C) was not included as a modification.

Kristian Murphy Gregersen is thanked for providing the medieval control specimens.

7. Reconstruction of unpublished extinct and extant rhinoceros dental enamel protein sequences

Jazmín Ramos-Madrigal, Eleftheria Palkopoulou, J. Víctor Moreno-Mayar, Mikkel-Holger S. Sinding, Shanlin Liu, Irina Kirillova, Peter D. Heintzman, Yoshan Moodley, Senthilvel KSS Nathan, Marcela Sandoval Velasco, Yvonne L. Chan, Joshua D. Kapp, Anders Götherstrom, Beth Shapiro, Love Dalén, M. Thomas P. Gilbert

7.1. Generation of high-coverage genomic data

The following gene sequences: "*ALB*", "*AMBN*", "*AMELX*", "*AMTN*", "*ENAM*", "*MMP20*" were extracted from unpublished high-coverage genomic data from the following extinct and extant rhinoceros species (Supplementary Data): woolly rhinoceros (†*Coelodonta antiquitatis*), Merck's rhinoceros (†*Stephanorhinus kirchbergensis*), Sumatran rhinoceros (*Dicerorhinus sumatrensis*), Javan rhinoceros (*Rhinoceros sondaicus*), Indian rhinoceros (*Rhinoceros unicornis*), and Black rhinoceros (*Diceros bicornis*). Our dataset thereby covers all extant species of rhinoceros as well as two Late Pleistocene extinct species. Summary details can be found in table S6.

Woolly rhinoceros (†*Coelodonta antiquitatis*). A femur from a woolly rhinoceros was collected during fieldwork along the Yana river in Russia. The specimen was found in the Muus-Khaya bone accumulation, approximately 500 m upstream from the Yana RHS site[62], and was subsampled for DNA analysis. Once having been brought back to the ancient DNA laboratory at the Swedish Museum of Natural History, 50 mg of bone powder was collected using a Dremel tool. DNA was extracted from the woolly rhinoceros (*Coelodonta antiquitatis*) specimen using a modified version of protocol C in Yang et al.[63] as described in Brace et al.[64] An indexed double-stranded library was generated from the genomic DNA following the protocol from Meyer & Kircher[65] including uracil excision with the USER (New England Biolabs) enzyme during the blunt-end repair step. Paired-end sequencing was performed at the SciLifeLab (Stockholm, Sweden) on 8 lanes of the Illumina HiSeq 2500 flowcell for 2×100bp cycles in HighOutput mode. Sequencing reads were processed with HiSeq Control Software 2.0.12.0/RTA 1.17.21.3 The white rhinoceros (*Ceratotherium simum*)

genome sequence cerSim1 available from UCSC was used as the reference genome for read alignment, implementing parameters recommended for ancient DNA in BWA[66] v.0.7.12 'aln' as described by Kircher[67]. SAMtools[68] v. 0.1.19 was used to sort and convert SAM format files to BAM format, and to remove duplicate reads using the 'rmdup –s' option.

Merck's rhinoceros (†*Stephanorhinus kirchbergensis*). A first molar (M1) tooth root from a complete Merck's rhinoceros cranium (F-4160) was subsampled, as previously reported by Kirillova *et al*.[69]. The University of California Santa Cruz Paleogenomics laboratory supplemented the DNA extract from Kirillova *et al*. with a new DNA extract, following previously reported methods[70]. We then converted each extract into four single-indexed, double-stranded DNA libraries following Meyer & Kircher (2010)[65], for eight libraries in total. These libraries were pooled and sent to SciLifeLab (Stockholm, Sweden) for sequencing on two lanes of the Illumina HiSeq-X platform, using 2× 150 cycles of paired-end chemistry. The resulting 880 million reads were merged and adapter-trimmed using SeqPrep (https://github.com/jstjohn/SeqPrep) and sequence complexity filtered using a DUST cut-off of 7 in PRINSEQ-lite[71] (v0.20.4). We discarded merged reads shorter than 30 bp. We then mapped the remaining merged and unmerged data to the Sumatran rhinoceros draft genome (Dalen and Gilbert, unpublished data) assembly in BWA aln (v0.7.7; 57), with the seed disabled[72]. We collapsed duplicate reads using rmdup in SAMtools (v.0.1.19; 59) and used BEDTools[73] (v2.25) to determine a modal genome-wide sequence coverage estimate of 7×. Using SAMtools flagstat, we determined that the sample has an endogenous DNA content of 42%, based on alignment to the Sumatran rhinoceros' genome. The mean DNA fragment length was estimated to be 73 bp. We next performed an assessment of ancient DNA damage patterns, using mapDamage2[74] (v2.0.5). The results clearly show the deamination and depurination-induced strand break patterns typical of ancient DNA (Fig. S4).

Fedor Shidlovskey and the Ice Age Museum, Moscow (Russia), are thanked for providing access to the studied *S. kirchbergensis* sample.

Sumatran rhinoceros (*Dicerorhinus sumatrensis*). A fresh blood specimen from a Sumatran rhinoceros (Kertam) was collected by staff at the Sabah Wildlife Department in Borneo, Malaysia. Genomic DNA was extracted from blood and two libraries with 180bp and 650bp insert size were build using Illumina's TruSeq DNA PCR-Free kit. Paired-end reads were processed as described above for the woolly rhino data, except for that the 180bp-insert library reads were merged only (not the 650bp-insert library reads) and that default parameters were implemented in BWA v.0.7.12. The SAMtools v.0.1.19 default 'rmdup' option was used for the 650bp-insert library since the reads were aligned as paired-end reads.

Javan rhinoceros (*Rhinoceros sondaicus*). Three specimens of dried soft tissue and one of bone were sub-sampled from a Javan rhinoceros skull collected in 1839 in Java, kept in the collections of the Natural History Museum at the University of Oslo (Museum id: 734). Samples were extracted and prepared for sequencing in the ancient DNA laboratories at the Centre for GeoGenetics – Natural History Museum of Denmark, following standard clean lab procedures[46,75]. Bone and dried soft tissue were extracted using, respectively, a EDTA - proteinase k buffer, as described in Gilbert *et al.* (2007)[76], and a urea - proteinase k buffer, as described in Ersmark *et al.* (2015)[77]. Next Generation Sequencing libraries were prepared following the single-tube protocol described by Carøe *et al.* (2017)[78], with the modifications as in Mak *et al.* (2017)[79], using Illumina-specific adapters[65], or BGISEQ500 specific adapters[79] for single read sequencing on Illumina HiSeq 2500 or BGISEQ500 respectively. Raw reads were mapped to the white rhinoceros reference genome (cerSim1; accession ID: GCA 0002831551.1) using bwa[66] and PCR duplicates were removed using picard-tools (http://broadinstitute.github.io/picard).

Indian rhinoceros (*Rhinoceros unicornis*). The specimen was provided by the San Diego Zoo (ID= KB14498, SB137). The DNA was extracted from a cell culture, and the DNA was sent to Stockholm where 180bp and 670bp insert libraries were constructed and sequenced on one lane each on the Illumina HiSeq X platform at SciLifeLab (Stockholm, Sweden), following the procedure described above. Raw reads were mapped to the white rhinoceros reference genome (cerSim1;

accession ID: GCA 0002831551.1) using bwa[66] and PCR duplicates were removed using picard-tools (http://broadinstitute.github.io/picard).

Reads mapping to the coordinates of the genes of interest were retrieved from all alignments using SAMtools[68]. The annotation of the reference genome was obtained from the GBshape genome browser database[80]. In table S6 we show the average depth of coverage for each protein, for each rhinoceros species except the Black and Merck's rhinoceros.

Black rhinoceros (*Diceros bicornis*). The specimen was provided by the Zululand Rhino Reserve, South Africa (ID=46373). The DNA was extracted from an ear biopsy using a Kingfisher Duo robot. 180bp and 670bp insert libraries were constructed and sequenced on one lane each on the Illumina HiSeq X platform at SciLifeLab (Stockholm, Sweden), following the procedure described above. For the Black rhinoceros, the genome assembly was searched for the genes of interest using the protein2genome model embedded in *Exonerate* (version 2.2)[81] and the protein sequences obtained from the white rhinoceros reference genome.

Table S6. Depth of coverage (high-throughput sequencing) overview for unpublished rhinoceros genomic data. For each rhinoceros species, for each gene, we show its genomic coordinates with respect to white rhinoceros (*Ceratotherium sinum simum*) genome sequence cerSim1, the proportion of bases with depth greater than 2X (%bases > 2X) and the average depth of coverage (AverageDepth). The latter were computed for sites with depth of at least 1X. See Supplementary Information §7.1, for a detailed description of the provenance of these data.

Rhinoceros	Protein	Contig	StartPosition	EndPosition	%bases>2X	AverageDepth
Indian	ALB	JH767724	74890191	74907074	99.73	77.23
Indian	AMBN	JH767724	72670086	72682952	99.76	75.49
Indian	AMELX	JH767774	1198717	1206021	99.60	63.93
Indian	AMTN	JH767724	72622131	72635475	99.88	76.00
Indian	ENAM	JH767724	72699344	72716716	99.54	71.03
Indian	MMP20	JH767744	19500006	19547511	99.65	71.72
Javan	ALB	JH767724	74890191	74907074	97.91	46.84
Javan	AMBN	JH767724	72670086	72682952	97.81	47.97
Javan	AMELX	JH767774	1198717	1206021	99.26	58.49
Javan	AMTN	JH767724	72622131	72635475	98.85	47.63
Javan	ENAM	JH767724	72699344	72716716	98.30	55.96
Javan	MMP20	JH767744	19500006	19547511	97.78	56.19
Sumatran	ALB	JH767724	74890191	74907074	97.90	25.65
Sumatran	AMBN	JH767724	72670086	72682952	96.91	27.03
Sumatran	AMELX	JH767774	1198717	1206021	97.37	14.20
Sumatran	AMTN	JH767724	72622131	72635475	97.37	25.09
Sumatran	ENAM	JH767724	72699344	72716716	97.99	27.18
Sumatran	MMP20	JH767744	19500006	19547511	98.58	26.20
Woolly	ALB	JH767724	74890191	74907074	48.52	2.67
Woolly	AMBN	JH767724	72670086	72682952	53.90	3.06
Woolly	AMELX	JH767774	1198717	1206021	24.50	2.01
Woolly	AMTN	JH767724	72622131	72635475	51.77	2.91
Woolly	ENAM	JH767724	72699344	72716716	54.22	2.98
Woolly	MMP20	JH767744	19500006	19547511	49.63	2.90

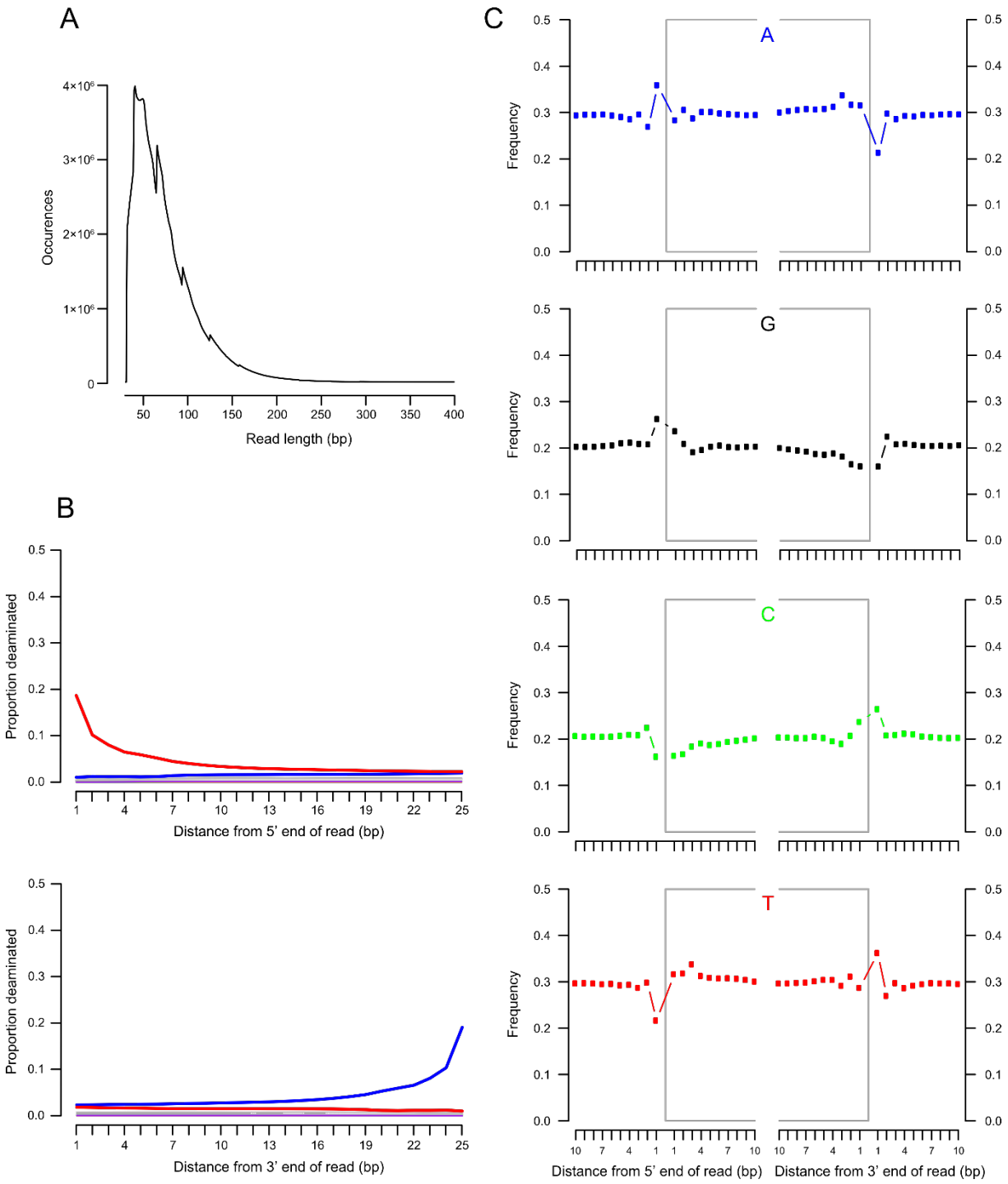

Figure S4. Ancient DNA damage patterns for the Merck's rhinoceros (*Stephanorhinus kirchbergensis*) specimen. (A) DNA fragment length distribution; (B) proportion of cytosines that are deaminated at fragment ends (red: cytosine > thymine; blue: guanine > adenine); and (C) average base frequencies immediately upstream and downstream of the 5' and 3' ends of mapped reads.

7.2. Protein sequence translation

For each protein in each species, we built a majority rule consensus sequence using ANGSD[82], after filtering reads with mapping quality lower than 30 and nucleotides with base quality lower than 20. Additionally, for the woolly rhinoceros and *S. kirchbergensis*, we excluded the first and last five

nucleotides of each read in order to minimize the effect of post-mortem ancient DNA damage. We formatted each consensus sequence as a separate blast nucleotide database and performed a tblastn alignment using the corresponding white rhinoceros sequence as a query. Since this step is aimed at recovering the entire translated and spliced protein sequence from the consensus, we did not filter low-complexity regions (-F F), we favoured alignments without gaps (-E 32767 -G 32767) and used the PAM70 substitution matrix to score the alignment (-M PAM70). Finally, we recovered the translated and spliced protein sequence from the local blast alignments using the ProSplign algorithm from the NCBI Eukaryotic Genome Annotation Pipeline[83]. Genomic BAM files used for Rhinocerotidae protein sequence translation are included as Supplementary Data 1.

8. Phylogenetic analysis of *Stephanorhinus* specimen Dm.5/157-16635

Jazmín Ramos-Madrigal, J. Víctor Moreno-Mayar

8.1. Protein sequence alignment

We assembled a comparative reference dataset including publicly available protein sequences from species that belong to the following families: Equidae, Rhinocerotidae, Suidae and Bovidae. The proteomic sequences from all the Dmanisi specimens assigned by morphology to the Cervidae family (Extended Data Tab. 1, 2) could not be processed for phylogenetic inference, due to the insufficient availability of protein sequences from this taxonomic group in public reference databases. For each of the nine ancient proteins that were reconstructed from the entire set of faunal mineralised specimens retrieved in Dmanisi (Supplementary Information §5.4), we first obtained the human reference sequence as annotated in UniProt. We used the *human* sequence as a query in a blastp search against the Refseq[84] database and restricted the results to sequences derived from ungulates (taxid 91561 and 9787). Note that for each protein, we only kept full-length sequences from species for which all nine proteins were available in Refseq. In table S7 we show the NCBI identifiers for each of the sequences that we included in the final dataset. Protein sequence alignments used for phylogenetic reconstruction are included as Supplementary Data 1.

Table S7. Comparative reference dataset provenance and ancient protein sequence overview. We assembled a comparative reference dataset as detailed in Supplementary Information §7.2. For each publicly available protein sequence, we show the corresponding NCBI identifier. For rhinoceros species, we use “FROM_BAM” to denote sequences that were translated from high-throughput sequencing data (Supplementary Information §7.1) or “FROM_GA” for those obtained from a genome assembly. For each ancient specimen, we use “SEQ” to denote protein sequences that were partially reconstructed and “NA” to indicate missing sequences.

Sample	ALB	AMBN	AMELX	AMTN	COL1A2	COL2A1	COL3A1	ENAM	MMP20
HUMAN	P02768	Q9NP70.1	Q99217.1	Q6UX39	P08123.7	P02458	P02461.4	Q9NRM1.3	O60882
Bison bison bison	XP_010836494.1	XP_010837412.1	XP_010828933.1	XP_010837411.1	XP_010838069.1	XP_010835850.1	XP_010861218.1	XP_010837410.1	XP_010832385.1
Bos taurus	NP_851335.1	NP_776413.1	NP_001014984.1	NP_001289882.1	NP_776945.1	NP_001001135.2	NP_001070299.1	XP_605463.4	NP_776816.1
Bubalus bubalis	XP_006047011.1	XP_006069380.1	XP_006042101.1	XP_006071182.1	XP_006054012.1	XP_006059500.1	XP_006073767.1	XP_006069381.1	XP_006053124.1
Capra hircus	XP_005681801.1	XP_005681784.1	NP_001272545.1	XP_005681989.1	XP_005678993.1	XP_013819296.1	XP_005675926.1	XP_005681785.1	XP_005689422.1
Ceratotherium simum simum	XP_004419220.1	XP_004419200.1	XP_004435161.1	XP_004419286.1	XP_004431414.1	XP_004428994.1	XP_004426778.1	XP_004419201.1	XP_004427457.1
Equus asinus	NP_001310707.1	XP_014702755.1	XP_014710898.1	XP_014702740.1	XP_014708844.1	XP_014705536.1	XP_014708400.1	XP_014702756.1	XP_014715496.1
Equus caballus	NP_001075972.1	XP_001489030.3	NP_001075296.2	XP_014594023.1	XP_001492989.1	XP_005611139.1	AAD13114.1	XP_001487944.1	XP_001500102.1
Equus przewalskii	XP_008524663.1	XP_008528890.1	XP_008523219.1	XP_008528888.1	XP_008522480.1	XP_008511884.1	XP_008507124.1	XP_008528891.1	XP_008520978.1
Ovis aries	NP_001009376.1	XP_004009935.1	XP_011961603.1	XP_012035497.1	XP_004007775.1	XP_012029774.1	XP_004004563.1	XP_004009936.1	XP_004016022.1
Pantholops hodgsonii	XP_005983856.1	XP_005983834.1	XP_005977934.1	XP_005983833.1	XP_004007775.1	XP_005981079.1	XP_005971121.1	XP_005983835.1	XP_005970113.1
Sus scrofa	NP_001005208.1	NP_999202.1	NP_998965.1	XP_013844773.1	NP_001230584.1	NA	NP_001230226.1	NP_999406.1	NP_999070.1
Coelodonta antiquitatis	FROM_BAM	FROM_BAM	FROM_BAM	FROM_BAM	NA	NA	NA	FROM_BAM	FROM_BAM
Dicerorhinus sumatrensis	FROM_BAM	FROM_BAM	FROM_BAM	FROM_BAM	NA	NA	NA	FROM_BAM	FROM_BAM
Rhinoceros sondaicus	FROM_BAM	FROM_BAM	FROM_BAM	FROM_BAM	NA	NA	NA	FROM_BAM	FROM_BAM
Rhinoceros unicornis	FROM_BAM	FROM_BAM	FROM_BAM	FROM_BAM	NA	NA	NA	FROM_BAM	FROM_BAM
Diceros bicornis	FROM_GA	FROM_ GA	FROM_ GA	FROM_ GA	NA	NA	NA	FROM_ GA	FROM_ GA
Stephanorhinus kirchbergensis	FROM_BAM	FROM_BAM	FROM_BAM	FROM_BAM	NA	NA	NA	FROM_BAM	FROM_BAM
16632	SEQ	NA	SEQ	NA	NA	NA	NA	SEQ	SEQ
16635	SEQ	SEQ	SEQ	SEQ	NA	NA	NA	SEQ	SEQ
16638	NA	SEQ	SEQ	NA	NA	NA	NA	SEQ	SEQ
16639	NA	SEQ	SEQ	NA	NA	NA	NA	SEQ	SEQ
16641	NA	SEQ	SEQ	NA	NA	NA	NA	SEQ	NA
16857	NA	NA	NA	NA	SEQ	SEQ	SEQ	NA	NA

For each protein, for all publicly available and unpublished rhinoceros reference sequences, we built a multiple sequence alignment using mafft[85]. Resulting alignments were inspected visually

and overhanging terminal sequences present in only one species were pruned. For each protein of each ancient individual, we aligned the ancient protein sequence to its corresponding reference alignment (Tab. S7) using mafft --add. The properties of these alignments are summarized in table S8. For each ancient individual, we created a concatenated alignment, which only included the proteins that were sequenced for that particular individual. Finally, to account for isobaric amino acids which are indistinguishable using the tandem MS approach previously described (Supplementary Information §5.2), we substituted all isoleucines in the alignments with leucines at sites where the ancient sample carried either of those amino acids. We use these concatenated alignments to infer the phylogenetic relationship between the ancient samples and the species included in the reference dataset, using three approaches: distance-based neighbor-joining, maximum likelihood and Bayesian inference.

Table S8. Retrieval of informative SAPs in ancient Dmanisi proteomes. Data derives from final *mafft* alignments.

GNM specimen ID	Specimen ID	Gene	Total sites	Coverage in ancient sample	Total segregating sites	Segregating sites in ancient	Non-singletons in ancient
Dm.5/154.2.A4.38	16632	ALB	609	37	268	22	19
Dm.5/154.2.A4.38	16632	AMELX	230	162	59	45	28
Dm.5/154.2.A4.38	16632	ENAM	1183	219	561	91	68
Dm.5/154.2.A4.38	16632	MMP20	487	141	119	43	18
Dm.5/157	16635	ALB	609	149	268	75	60
Dm.5/157	16635	AMBN	449	166	175	45	2
Dm.5/157	16635	AMELX	230	177	60	49	30
Dm.5/157	16635	AMTN	213	20	99	7	6
Dm.5/157	16635	ENAM	1183	289	561	128	97
Dm.5/157	16635	MMP20	487	74	120	15	12
Dm.5/154.1.B1.1	16638	AMBN	450	160	175	45	28
Dm.5/154.1.B1.1	16638	AMELX	245	116	59	30	18
Dm.5/154.1.B1.1	16638	ENAM	1183	146	561	41	34
Dm.5/154.1.B1.1	16638	MMP20	487	44	122	6	1
Dm.8/154.4.A4.22	16639	AMBN	449	134	175	33	18
Dm.8/154.4.A4.22	16639	AMELX	230	144	59	03	25
Dm.8/154.4.A4.22	16639	ENAM	1183	142	561	44	37
Dm.8/152.3.B1.2	16641	AMBN	449	127	175	35	19
Dm.8/152.3.B1.2	16641	AMELX	230	139	59	36	22
Dm.8/152.3.B1.2	16641	ENAM	1183	88	561	24	20
Dm.bXI.profile cleaning	16857	COL1A1	1472	245	118	15	5
Dm.bXI.profile cleaning	16857	COL1A2	1366	310	159	38	25

8.2. Phylogenetic inference

Neighbor-joining trees

Since ancient proteins cover only around 25% of the non-singleton segregating sites across alignments, we expect the genetic distances required by the neighbor-joining (NJ) algorithm to be overestimated. We circumvented this issue by using a modified version of the concatenated alignments that only includes sites in which the ancient samples have non-missing data. We computed the genetic distance between taxa and built the corresponding neighbor-joining trees using the phangorn[86] R package. For each alignment, we considered pairwise deletions and used the JTT model to estimate the genetic distances. We estimated the support for each bipartition through a non-parametric bootstrap approach based on 500 replicates.

Maximum likelihood phylogenetic inference

In contrast to the NJ algorithm, maximum likelihood (ML) inference is based on discrete characters. Thus, we used the complete concatenated alignments for this approach. We used PhyML[87] 3.1 9 to infer the topologies, branch lengths and substitution rates (-o tlr) for each alignment using the JTT model, based on which frequencies were estimated (-m JTT -f m). For each dataset, we used the maximum likelihood estimates of these two parameters: 1) the gamma distribution shape parameter to model rate heterogeneity (-a e), and 2) the proportion of invariable sites (-v e). We started from three random trees and for each optimization step, we kept the best tree between those generated through the 'nearest neighbor interchange' and 'subtree prune and regraft' routines (-s BEST --rand_start --n_rand_starts 3). We estimated the support for each bipartition based on 500 non-parametric bootstrap replicates.

Bayesian phylogenetic inference

In addition to the NJ and ML approaches, we carried out Bayesian phylogenetic inference using MrBayes[88] 3.2.6 10. We partitioned the concatenated alignments by gene, which allowed us to independently infer parameters for each partition. For each gene, we allowed the Markov chain to sample the following parameters: 1. the substitution rates from a set of predetermined matrices, 2. the shape parameter of a gamma distribution for modelling across-site rate variation (exponentially distributed), and 3. the proportion of invariable sites (uniformly distributed) (unlink Statefreq=(all) Ratemultiplier=(all) Aamodel=(all) Shape=(all) Pinvar=(all)). Note that the overall tree topology and branch lengths were inferred for the whole alignment. We ran the MCMCMC algorithm (two parallel runs) with 4 chains and a temperature parameter of 0.2 for 5,000,000 cycles, and sampled a set of parameters every 500 steps. For parameter inference, we discarded the first 25% of the iterations as burn-in. We assessed the convergence of the algorithm using Tracer v.1.6.0 (Fig. S5). For each ancient individual, we examined the effective sample sizes for all parameters (all greater than 5,500), as well as the overall log-likelihood trace on both replicates (Fig. S5). Based on these results and the large concordance between methods, we consider that we reached a reasonable local maximum for all runs.

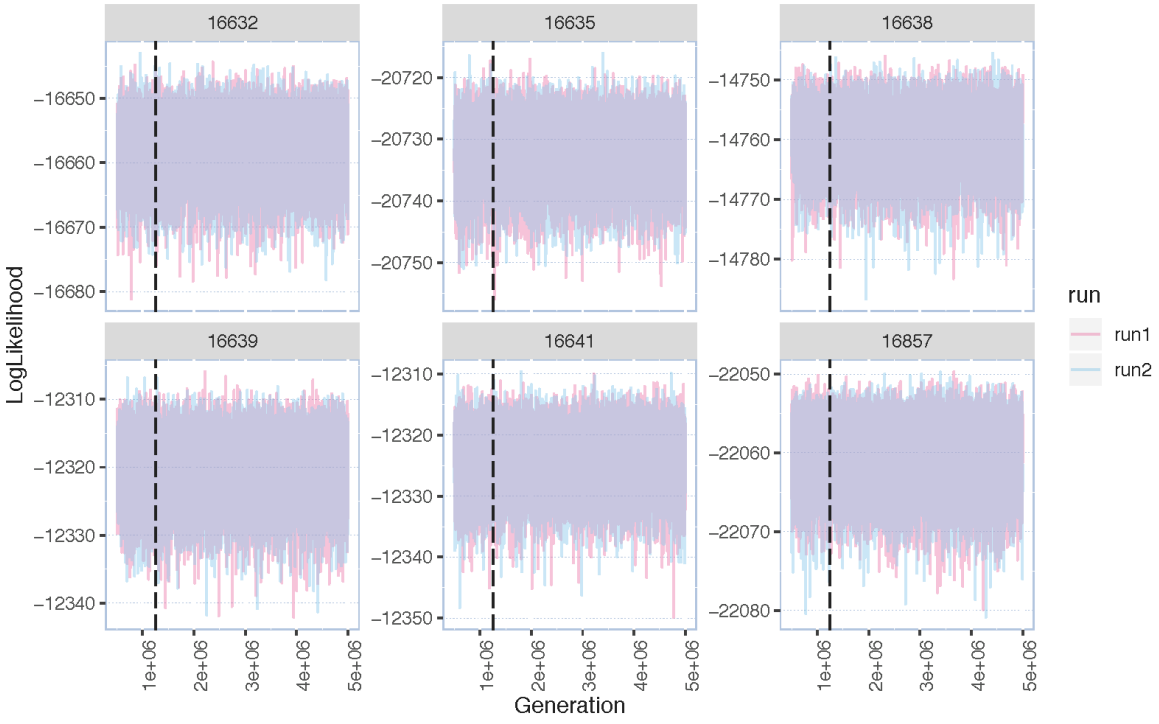

Figure S5. Bayesian inference log-likelihood traces for each ancient sample. We carried out Bayesian phylogenetic inference as detailed in SI Section 8.2. For each sample, we show the log-

likelihood of the model parameters sampled in each generation, for both independent runs. Samples to the left of the dotted line (25%) were discarded as burn-in.

9. Ancient DNA analysis

Morten E. Allentoft, Aurelien Ginolhac, Ludovic Orlando, Eske Willerslev

The bone and dentine specimens were processed using strict aDNA guidelines[89,90] in a clean lab facility at the Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen. DNA extraction was attempted on five of the ancient animal specimens (Tab. S9). Parts of the samples were drilled into powder and extracted using a silica-in-solution method[91,92]. To prepare the samples for NGS sequencing 20 µL of DNA extract was built into a blunt-end library using the NEBNext DNA Sample Prep Master Mix Set 2 (E6070) with Illumina-specific adapters. The libraries were prepared as described previously[91] and PCR-amplified with inPE1.0 forward primers and custom-designed reverse primers with a 6-nucleotide index[65]. Because of low DNA concentrations (Tab. S9), several attempts of library amplification proved unsuccessful for MA397, MA398 and MA400. Only after a total of 30 PCR cycles in a 2-round setup (12 cycles with total library, followed by 18 cycles with 5 µL library from the first amplification) did the reaction result in a sequenceable library product for extract MA399 and in a library product for extract MA2481, the latter originating from the *Stephanorhinus* ex gr. *etruscus-hundsheimensis* specimen Dm.5/157-16635 studied in-depth proteomically. The library for extract MA2481 amplified somewhat better in the PCR reaction, requiring 15 cylces (1 round) to work. We note that these were the only two samples with a detectable DNA concentration in the extract (Tab. S9). Two different libraries (MA399_L1 and MA399_L2) were constructed from extract MA399 and each library was pooled with others for sequencing on an Illumina 2000 platform using 100bp single read chemistry. Library MA2481 was constructed from extract MA2481 and was pooled with other libraries for sequencing on an Illumina 2500 platform using 81bp single read chemistry.

The data were base-called using the Illumina software CASAVA 1.8.2 and the sequences were demultiplexed with a requirement of full match of the 6 nucleotide indexes that were used. This resulted in a total of 11,001,436 reads for MA399_L1, 21,885,830 reads for MA399_L2 and 5,767,560 reads for MA2481. Raw reads were processed using the PALEOMIX pipeline following published guidelines[93], on a high performance computer[94]. Briefly, raw reads were trimmed for adapters using AdapterRemoval[95] v1.5, for the MA399 libraries, and v2.2.2, for the MA2481 library, and trimmed reads shorter than 25 bp were discarded. After trimming, a total of 8,629,253 reads were retained for library MA399_L1 (average length 45.0 bp), 21,617,987 reads were retained for library MA399_L2 (average length 70.9 bp) and 3,879,288 reads were retained for library MA2481 (average length 37.0 bp). To assess the authenticity of the DNA obtained from the animal remains, several genomes were used as reference for mapping. Given the extract MA399 was generated from a pelvis bone of a large mammalian herbivore (Cervidae or Bovidae), and the extract MA2481 was generated from dentine of *Stephanorhinus* ex gr. *etruscus-hundsheimensis*, we used as reference:

White rhino nuclear genome (*Ceratotherium_simum simum*, CerSimSim1)

Cow nuclear genome (*Bos taurus* 4.6.1, accession GCA_000003205.4)

Cow mitochondrial genome (*Bos taurus*)

Red deer mitochondrial genome (*Cervus elaphus*, accession AB245427.2)

Human nuclear genome (GRCh37/hg19)

The trimmed reads were mapped using BWA backtrack[66] v0.7.15 with the seed disabled. All other parameters were set as default. PCR duplicates from mapped reads were discarded using the

picard tool MarkDuplicate [http://picard.sourceforge.net/)]. Reads were not filtered for mapping quality, as a hit on a genome was more informative than its actual real genomic location.

Table S9. Ancient DNA extractions and library concentrations for Dmanisi fauna.

CGG reference number	GNM specimen number	Extraction	Mass	DNA conc. (ng/mL)
16635	Dm.5/157.profile cleaning	MA2481	≈500 mg	183
16856	Dm.M6/7.II.296	MA397	915 mg	<100
16857	Dm.bXI.profile cleaning	MA398	1015 mg	<100
16859	D4.collection	MA399	1281 mg	234
16860	Dm.65/62.1.A1.collection	MA400	504 mg	<100

SUPPLEMENTARY RESULTS

10. Biomolecular preservation: amino acid racemization results

Marc R. Dickinson, Kirsty Penkman

To test the extent of protein degradation, and therefore the likelihood of endogeneity, in the enamel of the *Stephanorhinus* (Dm.5/157-16635) tooth, six specimens, i.e. Dm.5/154.3.A4.32-16629, Dm.5/153.3.A2.33-16632, Dm.7/151.2.B1/A4.116634, Dm.5/157-16635, Dm.8/152.3.B1.2-16641 and Dm.M6/7.II.296-16856, from the deposits at Dmanisi were analysed for their intra-crystalline amino acid content. The sample with the most notable difference in amino acid composition is Dm.8/152.3.B1.2-16641 (*Bison*) with a reduced Gly composition and increased contribution from Leu (Fig. S6). Where comparable teeth of the same/similar species are available (Dm.5/153.3.A2.33-16632 and Dm.7/151.2.B1/A4.1-16634, from *Equus*), similar amino acid compositions are observed. The *Stephanorhinus* (Dm.5/157-16635) enamel proteins are consistent with the compositions of the other species of the same antiquity, indicating a similar fraction of proteins has been analysed.

The total concentration of amino acids in the *Bison* enamel sample is higher than the other taxa from Dmanisi, for both the free and bound fractions (Fig. S7). The lowest concentrations were observed in the *Equus* enamel samples. FAA are thought to be more labile than peptide bound amino acids and thus are more readily lost from the inter-crystalline fraction; retention of FAA indicates a closed system is present[42]. The extent of peptide bond hydrolysis ranges from 53-65% FAA, which is therefore consistent with closed system behaviour in samples of this age.

The high correlation between the THAA and FAA D/L values is consistent with closed system degradation of the intra-crystalline fraction. Increases in THAA racemization also correspond with higher FAA D/L values (Extended Data Fig. 4). The relationship between the extent of FAA and THAA racemisation is not expected to be linear due to the intrinsic nature of IcPD, and may not be synonymous between different taxa[96,97]. However, despite the potential differences between taxa, the two variables exhibit strong correlation ($R^2 > 0.93$) in all of the amino acids studied (calculated from polynomial best fit trendlines; Extended Data Fig. 4). This indicates the enamel proteins analysed from the taxa from Dmanisi are endogenous[40].

The extent of racemization in the Dmanisi samples varies between taxa. The *Bison* enamel sample has the highest levels of racemization, consistent with its amino acid composition, which shows lower levels of Ser (a relatively unstable amino acid), a higher contribution of Ala (a common amino acid breakdown product; Fig. S6) and high percentage of FAAs (Fig. S7), consistent with greater peptide bond hydrolysis. Interestingly the *Bison* also had the highest concentrations of intra-crystalline protein. For cases where the same or similar species have been analysed (*Equus* and Cervidae), extents of racemization in both the FAA and THAA are similar. This indicates that the overall spread of the data from Dmanisi is due to the different protein diagenesis patterns between taxa, as is also observed for mollusc shells[98]. In summary, the protein composition, level of peptide bond hydrolysis and extent of racemization in the *Stephanorhinus* enamel (Dm.5/157-16635), and all the other samples analysed in this study, is consistent with the isolation of a closed system of endogenous proteins.

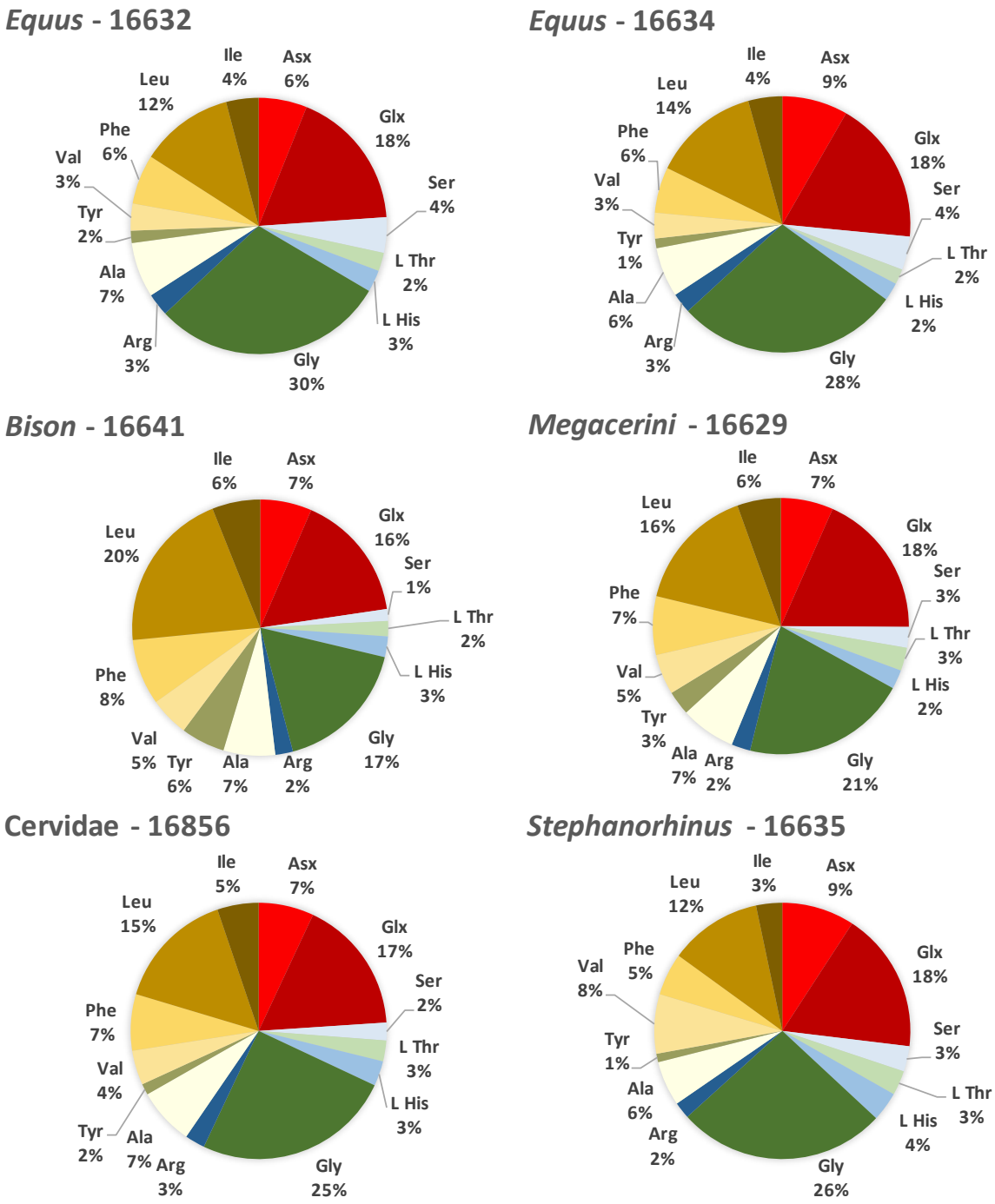

Figure S6. The intra-crystalline THAA composition of tooth enamel from various taxa from the deposits in Dmanisi. Amino acids are coloured according to their side chain properties: ionisable side chains with a negative charge (red), ionisable side chains with a positive charge (blue), non-ionisable side chains that are polar (yellow) and non-ionisable side chains that are non-polar (green). See Extended Data Tab. 1 for cross-reference of CGG and GNM specimen numbers.

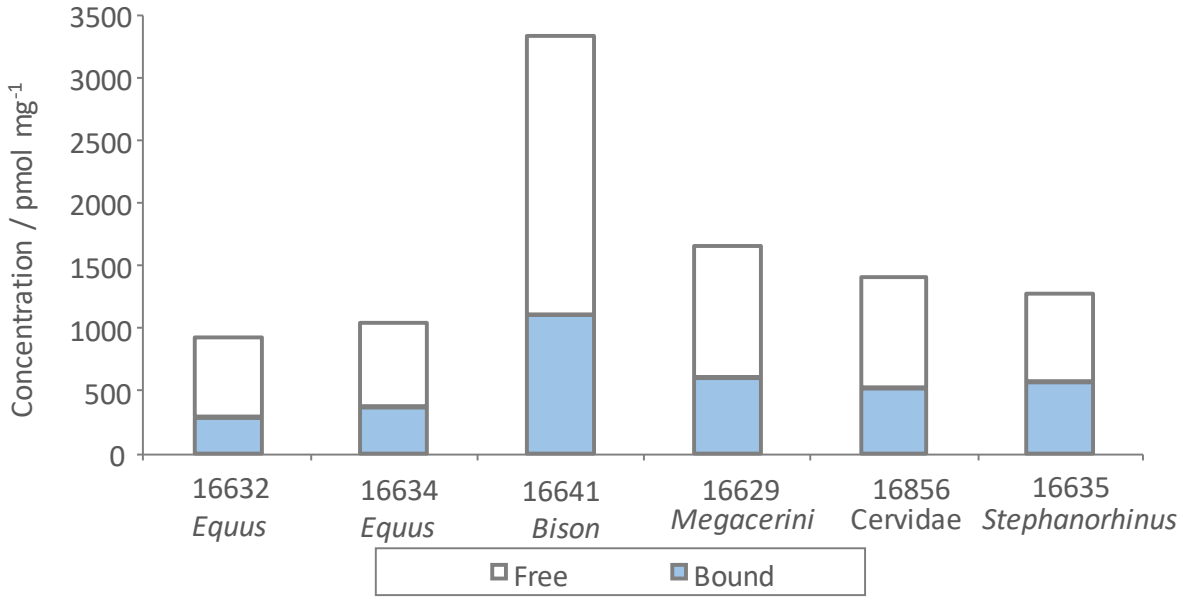

Figure S7. Total concentration of amino acids in the enamel fossil samples from Dmanisi. Solid colour indicates the concentration of the bound amino acids and the upper outlined section represents the free amino acid concentration. The total amino acid concentrations are indicated by the combined height of the bars.

11. Palaeoproteomic results

Enrico Cappellini

11.1. Peptide/protein/isoform recoveries

Several specimens (Extended Data Table 1) returned no or very few confidently identified MS/MS spectra, mostly assigned to common contaminants. The total number of identified MS/MS spectra, including those assigned to contaminant proteins, exceeded 1500 in a few cases and only for one specimen, Dm.5/157-16635, more than six thousand MS/MS spectra were confidently identified. Similarly, only for a few specimens the fraction of identified spectra was above 5% (Tab. S10).

Table S10. MS/MS spectra acquired and identified (including contaminants) for each specimen. In those cells reporting two values separated by the "|" symbol, the first value refers to MaxQuant (MQ) searches performed selecting unspecific digestion, while the second value refers to MQ searches performed selecting trypsin digestion. For those cells including one value only, it refers to MQ searches performed selecting unspecific digestion.

N.	GNM specimen ID	CGG specimen ID	MS/MS spectra in the raw files (n.)	MS/MS spectra identified (n.)	MS/MS spectra identified (%)
1	Dm.bXI.sqA6.V._.	16486	10525	0	0
2	Dm.6/154.2/4.A4.17	16626	11553	41 | 10	0.35 | 0.09
3	Dm.7/154.2.A2.27	16628	10949	38 | 24	0.35 | 0.22
4	Dm.5/154.3.A4.32	16629	31461	410 | 17	1.30 | 0.05
5	Dm.6/151.4.A4.12	16630	38205	1324 | 31	3.47 | 0.08
6	Dm.69/64.3.B1.53	16631	41879	1578 | 26	3.77 | 0.06
7	Dm.5/154.2.A4.38	16632	32707	1771	5.41

8	Dm.5/153.3.A2.33	16633	16899	4	0.02
9	Dm.7/151.2.B1/A4.1	16634	18564	281	1.51
10	Dm.5/157	16635	90982	6393	7.03
11	Dm.6/153.1.A4.13	16636	12731	10 | 18	0.08 | 0.14
12	Dm.7/154.2.A4.8	16637	28200	64 | 88	0.23 | 0.31
13	Dm.5/154.1.B1.1	16638	42796	1926 | 137	4.50 | 0.32
14	Dm.8/154.4.A4.22	16639	24871	1515	6.09
15	Dm.6/151.2.A4.97	16640	7613	41 | 7	0.54 | 0.09
16	Dm.8/152.3.B1.2	16641	25419	409	1.61
17	Dm.8/153.4.A4.5	16642	15368	276	1.80
18	Dm.M6/7.II.296	16656	63827	689 | 402	1.08 | 0.63
19	Dm.bXI	16857	125980	1634 | 2121	1.30 | 1.68
20	Dm.bXI.North.B1a	16858	64541	33 | 116	0.05 | 0.18
21	D4.collection	16859	40747	3 | 19	0.01 | 0.05
22	Dm.65/62.1.A1	16860	36131	179 | 32	0.50 | 0.09
23	Dm.64/63.1.B1z	16861	28786	55 | 33	0.19 | 0.11
		Neg. contr. 1	17275	78 | 122	0.45 | 0.71
		Neg. contr. 2	20284	70 | 67	0.35 | 0.71
		Neg. contr. 3	13416	8	0.06

The total number of identified MS/MS spectra in specimen Dm.5/157-16635, i.e. 6393, is approximately half of the number of spectra, i.e. 12,823, identified in another well characterised ancient specimen: a mammoth ~43,000 years old from Siberia[46]. For the Dmanisi specimens, the final number of ancient peptides and proteins confidently identified after removing contaminants and reverse sequences, as well as filtering for proteins identified with a minimum of two unique non-overlapping peptides, can be found in Extended Data Table 2.

Ancient protein sequences were confidently retrieved from 15 out of the 23 ancient specimens investigated (Extended Data Tab. 1, 2). Bone specimens, e.g. Dm.7/154.2.A2.27-16628 and Dm.bXI.profile cleaning-16857, sporadically returned collagen. Occasionally, collagen was also retrieved in enamel samples not entirely separated from dentine debris, e.g. samples Dm.5/154.2.A4.38-16632 and Dm.M6/7.II.296-16856. The proteomes identified in the other samples include instead, almost exclusively, dental enamel specific proteins, highlighting how this material appears to be particularly favourable to ancient protein preservation. Erupted dental enamel is a highly mineralized bioceramic and the hardest material in the vertebrate body[99]. The high mineral content in enamel, ~95% of its mass (wt%), turns it into a matrix that is highly inert to chemical alterations and particularly resistant to diagenesis occurring over thousands or millions of years. The organic extracellular matrix is quite limited, ~3 wt%, mostly represented by proteinaceous residues.

Enamel extracellular matrix proteins are tissue-specific and play a critical role in enamel formation[100]. Their mutations can affect the formation of enamel leading to several genetic disorders globally referred to as "*amelogenesis imperfecta*", a condition featuring soft, porous enamel[101]. During enamel maturation, most proteins are removed from the matrix by enamel proteases: matrix metalloproteinase-20 (MMP-20) and kallikrein 4 (KLK4), before complete mineralization is achieved[102]. Some residual protein fragments, however, persist in the mineralized matrix of erupted enamel. Enamel-specific proteins were previously identified in permanent enamel of human erupted teeth using LC-MS/MS[103]. Although the recovery of dental enamel proteins from

ancient samples was previously reported[104,105], this is the first time, to the best of our knowledge, that dental enamel proteins are retrieved from samples older than 1 Ma.

Amelogenins isolated *in vivo* from developing enamel appear to be the translation products of alternatively spliced RNA messages eventually processed by proteinases. The alternative splicing of amelogenins was first proposed to explain the origin of the leucine-rich amelogenin polypeptide (LRAP) which was identical to the major amelogenin protein at its amino and carboxyl-termini, but lacked a large segment from the centre of the protein[99]. Our results demonstrate the approach we used allows to recover the LRAP isoform at least from some of the Dmanisi specimens (Extended Data Fig. 3).

The unprecedented implementation of a digestion-free sample preparation approach in palaeoproteomics significantly increased the number of identified peptides. Relying on spontaneous protein backbone hydrolysis occurring over very wide time-ranges to generate tandem MS-compatible peptides, instead of using tryptic digestion, brings other significant advantages. First, it allows to identify exactly which fragments were generated by spontaneous hydrolysis with no bias introduced by enzymatic digestion (Fig. 2c). For example, diagenesis-induced PTMs might hinder enzymatic cleavage, or result in random cleavage patterns potentially unrecognized during data analysis. In the future, modelling the fragmentation patterns in ancient proteins could improve the reconstruction of spontaneous protein fragmentation, potentially leading to the identification of “fragile” sites in the amino acid sequence. Second, digestion-free sample preparation allows to sequence extended sets of randomly fragmented, partially overlapping peptides that can conceptually be treated similarly to DNA reads generated during high-throughput genomic sequencing. They can be aligned or mapped to a reference to generate consensus sequences in which each amino acid position can be covered multiple times by different overlapping peptides, significantly increasing the confidence of each amino acid call. This possibility is particularly useful to confidently discover amino acid substitutions private to extinct organisms for which no genomic information is, and arguably will not be, available.

11.2. Sex determination

Molecular sex determination of ancient specimens is commonly conducted by screening for the presence of DNA fragments mapping to the Y-chromosome, or PCR-based amplifications of X or Y chromosome-specific sequences[106,107]. Recently, sex identification based on peptide sequences specific for amelogenin X (AMELX) and Y (AMELY) was introduced as an alternative[108], with the oldest AMELY peptides deriving from the mid-Holocene[105]. Our data demonstrates that four enamel specimens from Dmanisi contain AMELY-specific peptides, indicating they can be assigned to male individuals (Extended Data Fig. 7) and that protein-based sexing of mammals is possible across most of the Pleistocene for temperate sites.

Crucially, the absence of AMELY-specific peptides might be either due to the specimen originating from a female individual, or to allele drop causing false negative detection of AMELY-specific peptides. The latter option can’t be excluded as the amount of expressed AMELY is known to represent approximately 10% of the expression of AMELX[109]. Future methodological developments in the mass spectrometry screening for AMELX- and AMELY-specific peptides could circumvent this issue.

11.3. Post Translational Modifications

Phosphorylation
David Lyon, Rosa Rakownikow Jersie-Christensen, Christian D. Kelstrup, Jesper V. Olsen

The search for serine/threonine phosphorylated peptides identifies a number of phosphopeptides covering several sites mainly on three proteins; Enamelin, Ameloblastin, Amelogenin (Fig. 3, Extended Data Figs. 2, 5). We identified 40 unique phosphorylated sequences defined by +/- 6 amino acids surrounding the phosphorylated residues, including several spectra with deamidated glutamine and asparagine sites (Extended Data Fig. 5). IceLogo-based sequence motif analysis (Fig. 3) reveals a strong overrepresentation of phosphorylated serine with a glutamate in +2 position. This S-x-E sequence is a well-established substrate motif for the Fam20C kinase[110], which is the major protein kinase modifying extracellular proteins and is known to phosphorylate enamelin, amelotin and amelogenin involved in biomineralization[111], further confirming the endogenous origin of our enamel proteomes.

Deamidation

David Lyon

Overall deamidation rates are high, ranging from over 50 to 100% deamidation (Fig. 2a). With a few exceptions, for example sample Dm.M6/7.II.296-16856 and Dm.bXI.profile cleaning-16857 which mainly contain collagen, glutamine deamidation rates are between 90-100%, suggesting near-complete diagenetic modification through this pathway. Biological pathways, such as conversion of asparagine to aspartic acid during deglycosylation via N-glycanase are unlikely to influence our data. For example, N-glycanase deamidation occurs on the substrates N-x-S ad N-x-T. However, out of the 33 asparagines confidently detected in our *Stephanorhinus* sp. sample (Dm.5/157-16635), only four match the N-x-S motif and only one represents the N-x-T motif. Given the high deamidation values observed, we are confident that these are not the result of N-glycanase deglycosylation.

Unexpectedly, for the enamel+dentine proteomes, asparagine deamidation is not always higher than glutamine. On the contrary, for ancient bone proteomes the anticipated pattern, with asparagine deamidation higher the glutamine one, is fully respected (Fig. 2a). Given: (i) the small difference between the asparagine and glutamine deamidation in relation to the high degree of deamidation in general, (ii) the consistent observation of more glutamines than asparagines, and (iii) the margin of error, this observation might not hold given more data. For example, CI intervals for glutamine and asparagine overlap for all samples, although sometimes only to a small extent. Still, we note that we observe the same pattern in the enamel proteome of our Medieval control, which has lower deamidation values for both asparagines and glutamines in comparison to the Dmanisi samples, but higher deamidation for the glutamines than for the asparagines. Although detailed modelling of the deamidation processes occurring in Early Pleistocene dental enamel proteome is beyond the scope of this manuscript, we suggest that the higher deamidation rate of Q compared to N in all the Dmanisi enamel specimens can be due to the cumulative effect of deamidation and peptide bond cleavage[112] followed by terminal hydrolysis.

Other forms of protein damage

Diana Samodova, Patrick L. Rüther

We profiled the samples from Dmanisi for other age-related post-translational modifications (PTMs) using two complimentary algorithms, PEAKS PTM[60] and MaxQuant[113] dependent peptide search.

The “PEAKS PTM” search uses high-quality *de novo* sequences (filtered by ALC score) for the identification of all modifications in the Unimod database[60]. MaxQuant, on the other hand was operated in “Dependent Peptides” mode[61,113], which performs a regular database search followed by a screening for similarities between the identified (termed base peptides) and non-identified spectra. This unbiased search yielded mass-shifts and localizations within the peptides, which were used to find relevant PTMs and re-search them specified as “variable modifications” in

regular search mode. Both searches were set to a false discovery rate of 5% that was controlled manually in the MaxQuant results, and automatically in PEAKS PTM with a modified target-decoy approach[114].

Ancient proteomes represent a challenge for conventional intensity-based quantification approaches. The high diversity of peptides, which are obtained by unspecific protein degradation and modified over time, only leaves very few pairs of modified peptides and unmodified counterparts for the calculation of relative intensities. Therefore, we decided to use the simpler approach of spectral counting, which in our hands resulted in the best comparability between the two search algorithms[115]. In order to accommodate for the high variability of peptide abundances and the score cut-offs defined by FDR between the different samples, relative peptide-spectrum matches (% PSMs) were calculated. These were obtained from the MaxQuant variable modification searches with specified PTMs of interest by dividing the number of PSMs with a certain modification by the total number of PSMs in the respective search. This “relative PSM” percentage allowed for the comparison between individual samples and most importantly between the ancient samples from Dmanisi and a younger enamel control sample from a medieval goat. Unfortunately, this approach is not applicable to PEAKS PTM search or MaxQuant dependent peptide search results because these algorithms have a much larger theoretical search space and have less robust FDR control varying with the number of unidentified spectra. In order to compare between the two search algorithms, absolute PSM values were used. For quantitative comparison between ancient samples, we believe that relative PSM values offer a good approach.

The molecular damage patterns determined in samples Dm.5/154.2.A4.38-16632, Dm.5/157-16635 and Dm.M6/7.II.296-16856 can be divided into two main groups – oxidation and glycation reaction products. Damage patterns were also evaluated in the bone sample Dm.bXI.profile cleaning-16857 that had the highest collagen sequence coverage. However, no novel potentially aging-related PTMs were detected, in this specimen. In the following results, we report both the PSMs obtained with PEAKS PTM search and the conventional MaxQuant search with specified variable modifications. Although in both cases the FDR was controlled at 5%, the numbers of matched spectra differ in rare cases. This is likely due to the different approaches of FDR calculation. Nevertheless, both search engines produced comparable results.

Oxidation

The results show that some amino acids underwent mono- and di-oxidation, during aging processes (Tab. S11). Sample Dm.5/157-16635 showed the highest number of PSMs corresponding to Trp and Tyr mono- and di-oxidative modifications based on both PEAKS PTM and MaxQuant search results. Searched with PEAKS, samples Dm.5/154.2.A4.38-16632 and Dm.M6/7.II.296-16856 revealed only the presence of Trp oxidation products, whereas Tyr mono- and di-oxidation were detected by MaxQuant search. Besides, Asp and Lys oxidative PTMs were also detected by both algorithms with low PSM counts. However, these values were not significantly higher than the negative control, data not shown.

Table S11. Peptide-spectrum matches (PSMs) corresponding to mono- and di-oxidative modifications of Trp and Tyr in the studied samples.

PEAKS PTM search					MQ Variable modification search			
Modification (Da)	Amino acid	16635	16632	16856	16635	16632	16856	Neg. control
Mono-oxidation	Trp	482	43	18	441 (6.28%)	29 (0.96%)	17 (1.43%)	7 (0.08%)
ΔM = +15.994915	Tyr	54	n.d.	n.d.	367 (5.23%)	28 (0.93%)	8 (0.67%)	43 (0.48%)
Di-oxidation	Trp	570	76	18	446 (6.36%)	133 (4.40%)	25 (2.10%)	5 (0.06%)

ΔM = +31.989828	Tyr	16	n.d.	n.d.	318 (4.53%)	93 (3.08%)	15 (1.26%)	20 (0.21%)

The relative PSM counts showed an up to 20-fold increase of tryptophan and tyrosine oxidation comparing the ancient samples with enamel from a medieval ovicaprine specimen used as a negative control. The balance between mono- and di-oxidations of tryptophan and tyrosine side chains was shifted towards di-oxidative modifications in sample Dm.5/154.2.A4.38-16632, while it was close to 1:1 in samples Dm.5/157-16635 and Dm.M6/7.II.296-16856. However, the interpretation of these results was not trivial, as these oxidative PTMs are often the intermediates of further amino acid oxidation (Fig. S8). Hence, a low relative PSM count could be indicative for both, a low and a very high stage of degradation. Besides, we observed similar absolute PSM values with two different algorithms (PEAKS PTM and MaxQuant) for Trp oxidations, but significantly higher numbers for Tyr with MaxQuant.

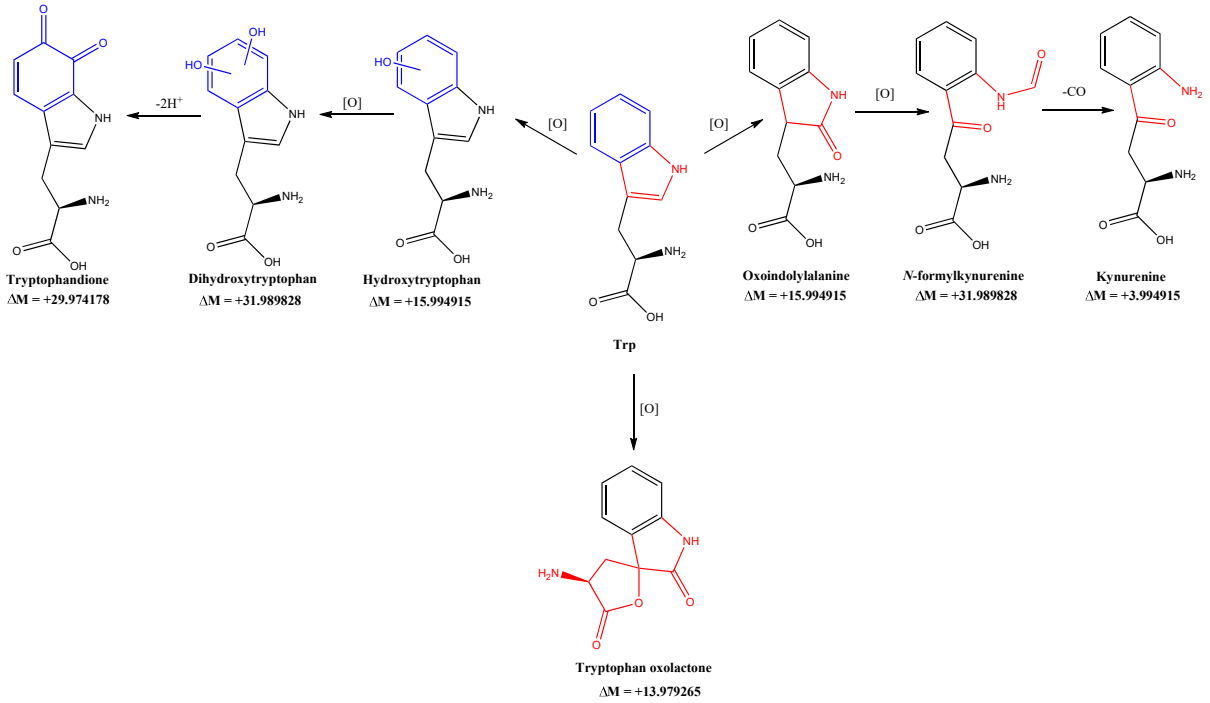

Figure S8. Oxidative degradation of tryptophan (Trp) via different pathways. ΔM: mass shift (Da).

In the case of tryptophan, we were able to identify various advanced oxidation products (Fig. S8). This phenomenon can be explained by the high reactivity of the easily oxidized indole group in the amino acid's chemical structure. The indole aromatic heterocyclic structure consists of electron rich benzene and pyrrole rings that makes it favourable for redox reactions[116,117]. There are multiple pathways of Trp oxidation resulting in formation of various oxidation products[118-121]. Oxidation of the pyrrole cycle yields either kynurenine or tryptophan oxolactone, while oxidation of the benzene ring leads to tryptophandione (Fig. S8). In practice, some Trp oxidation products cannot be distinguished by mass shift (e.g. dihydroxytryptophan and *N*-formylkynurenine). However, detection of the final reaction products, such as tryptophandione and kynurenine, suggests existence of the mentioned intermediates in the degradation pathways. In the current study, all intermediates with a mass shift of +15.9949 or +31.9898 were classified as mono- and di-oxidative modifications.

In the samples from Dmanisi, mass shifts indicative for the advanced oxidation products kynurenine, tryptophandione, and oxolactone were found. PEAKS PTM search only yielded identifications of kynurenine and Trp-dione in the sample Dm.5/157-16635, while oxolactone was

also found in samples Dm.5/157-16635 and Dm.5/154.2.A4.38-16632. MaxQuant database search revealed identification of the three modifications in all samples (Tab. S12). Kynurenine and oxolactone were found in samples Dm.5/157-16635 and Dm.5/154.2.A4.38-16632 in 1 to 3.5% of all PSMs, while they were almost absent in the medieval goat control sample. Because of the low general ID-rate in sample Dm.M6/7.II.296-16856, only few peptides with Trp modifications were identified. Tryptophandione was only identified with very low spectral counts in all samples, making the PSM percentages rather unreliable, but they were still higher than in the control. Noteworthy, in over 65 % of the peptides with a Trp→oxolactone conversion, the Trp residue was located at the C-terminus, while this percentage was below 30%, for the four other Trp modifications. Possibly, the modification is part of the mechanism of peptide bond breakage after Trp residues or C-terminal Trp residues are more susceptible to oxolactone conversion[122]. Overall, the results of the study show that Trp can be modified into various oxidation products, over time, thus confirming the high redox activity of the amino acid. Based on this, we conclude that the modifications oxolactone, kynurenine, and tryptophandione can potentially be used as markers of age-related oxidation.

Table S12. Peptide-spectrum matches (PSMs) corresponding to Trp oxidative modifications in the studied samples. Trp mono- and di-oxidation values are reported in table S11.

PEAKS PTM search				MQ Variable modification search			
Modification (Da)	16635	16632	16856	16635	16632	16856	Neg. control
Kynurenine ΔM = +3.994915	156	n.d.	n.d.	248 (3.53%)	55 (1.82%)	5 (0.42%)	1 (0.01%)
Oxolactone ΔM = +13.979265	116	21	n.d.	67 (0.95%)	69 (2.28%)	3 (0.25%)	4 (0.04%)
Tryptophandione ΔM = +29.974178	14	n.d.	n.d.	10 (0.14%)	4 (0.13%)	2 (0.17%)	3 (0.03%)

Histidine oxidation-linked degradation

Oxidative degradation products of the His side chain have been previously described for O_2 and UV-exposed ancient samples[59,123,124]. While the degradation of processed proteins in cultural heritage material with exposure to air and sunlight happens relatively fast over time periods of hundreds of years[59], we hypothesize that the reaction is slowed down significantly for proteins embedded in enamel or bone.

We found oxidized and di-oxidized His, as well as His to Asp conversion, with both search algorithms (Tab. S13). The potential conversion of His to hydroxyglutamic acid[59] was only identified with MaxQuant and not with PEAKS PTM search, as it was not included in the UniMod database at the time of writing. Relative PSM counting showed comparable levels of His oxidation and di-oxidation in the three specimens from Dmanisi and in the medieval control. Conversion of His to aspartic and hydroxyglutamic acid (Fig. S9), however, was increased by a factor of 10 to 30 in the older samples from the Dmanisi site. The relative counts were slightly lower for sample Dm.5/157-16635 indicating that it was less degraded or His was converted into another unknown degradation product.

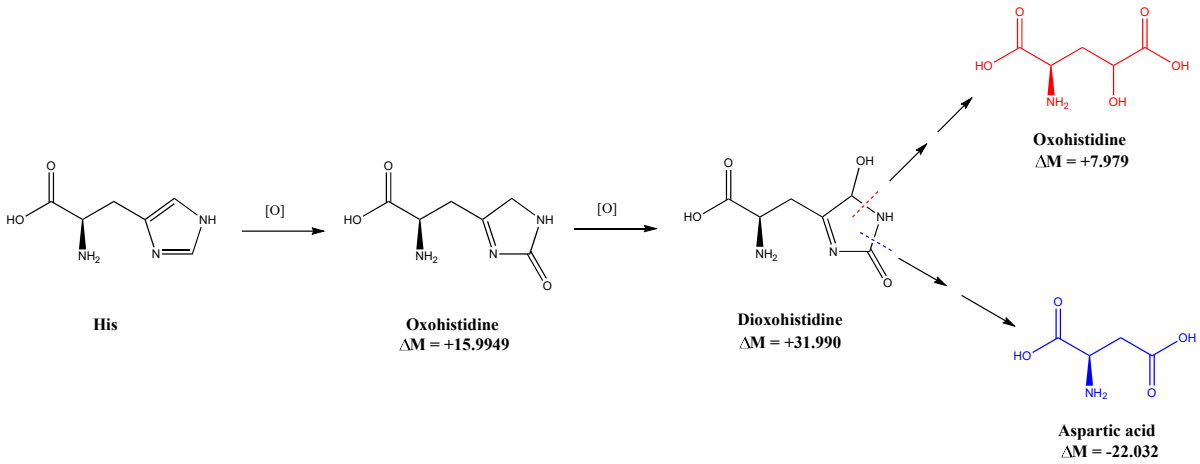

Figure S9. Oxidative degradation of histidine (His). ΔM: mass shift (Da).

Table S13. Peptide-spectrum matches (PSMs) corresponding to His degradation–linked modifications in the studied samples.

MQ Variable modification search				
Modification (Da)	16635	16632	16856	Neg. control
Oxohistidine ΔM = +15.9949	18 (0.26%)	8 (0.26%)	4 (0.34%)	66 (0.68%)
Dioxohistidine ΔM = +31.990	18 (0.26%)	36 (1.19%)	3 (0.25%)	18 (0.19%)
His -> Hydroxyglutamate ΔM = +7.979	159 (2.27%)	116 (3.84%)	38 (3.19%)	12 (0.14%)
His -> Aspartic Acid ΔM = -22.032	89 (1.27%)	61 (2.02%)	59 (4.95%)	12 (0.13%)

Ornithine formation

Relatively high numbers of PSMs corresponding to peptides with Arg to ornithine conversion (ΔM = -42.0218 Da) were observed with both search algorithms (Tab. S14). Arg can be glycosylated via Maillard reaction producing AGEs[125] that may serve as precursors for *in vivo* conversion of arginine to ornithine[126]. In order to find these potential intermediates, we included carboxymethylation and –ethylation in the MaxQuant database search.

The numbers of PSMs with Arg to ornithine conversion matched between the PEAKS PTM and MaxQuant searches for samples Dm.5/157-16635 and Dm.M6/7.II.296-16856, while PEAKS PTM search tool did not identify ornithine in the sample Dm.5/154.2.A4.38-16632 (Tab. S14). Since the relative number of PSMs with Arg to ornithine conversion was much higher in the Dmanisi samples compared to the medieval control, we believe that it could serve as a marker for high sample age. Our data showed very low abundance of Arg-linked AGEs, proposed as intermediates in glycation reaction, compared to the negative control. Hence, we hypothesize that the slow conversion of Arg into ornithine is based either on a different mechanism, or on different intermediates not included in the current search.

Table S14. Peptide-spectrum matches (PSMs) corresponding to Arg–linked modifications.

PEAKS PTM search				MQ Variable modification search			
Modification (Da)	16635	16632	16856	16635	16632	16856	Neg. control
Carboxymethylation ΔM = +39.9949	n.d.	n.d.	n.d.	0 (0%)	5 (0.17%)	1 (0.08%)	49 (0.53%)
Carboxyethylation ΔM = +54.0106	n.d.	n.d.	n.d.	2 (0.03%)	7 (0.23%)	2 (0.16%)	78 (0.86%)
Ornithine ΔM = -42.0218	228	n.d.	41	248 (3.53%)	202 (6.68%)	82 (6.88%)	75 (0.83%)

Terminal losses during peptide degradation

Traditionally, tryptic digestion is an essential step in proteomics workflow for any modern or ancient sample because it produces peptides, which are in a good size range and are easily ionized[127]. In the samples from Dmanisi, protein degradation had progressed so much that the peptide length had reached the lower limit of seven amino acids with our LC-MS/MS and data analysis setup (Fig. 2c). Hence, most of the samples were extracted and analysed without digestion.

Since it is not well understood how peptides degrade over aging periods of more than one million years, we decided to have a closer look at the peptide backbone. Besides the predominant hydrolysis of the amide bond, some peptides lost the N-terminal amino group (1-5% of all PSMs) and others lost the C-terminal carboxyl group (1-2% of all PSMs). This was potentially due to asymmetric backbone cleavage or further degradation of the terminal groups. In the case of N-terminal loss, the relative amounts were higher in the Dmanisi samples (2-5%) compared to the medieval control (1%; Tab. S15). There are probably more modifications of the termini to be discovered in the future.

Table S15. Peptide-spectrum matches (PSMs) corresponding to terminal group losses.

MQ Variable modification search				
Modification (Da)	16635	16632	16856	Neg. control
C-terminal COO-loss ΔM = +39.9949	77 (1.10%)	42 (1.39%)	22 (1.85%)	99 (1.08%)
N-terminal NH-loss ΔM = +54.0106	123 (1.75%)	161 (5.32%)	39 (3.27%)	91 (0.98%)

12. Phylogenetic reconstruction

Jazmín Ramos-Madrigal, J. Víctor Moreno-Mayar

We found largely consistent results across the three different phylogenetic inference approaches (Figs. S10-S15). While we were not able to resolve certain clades using the NJ approach, discrete character-based methods yielded results that are more robust. We attribute this difference in resolution to the conservative ascertainment scheme implemented for NJ, where we only included sites where the ancient samples have non-missing data (Supplementary Information §8.2). In what follows, we base our inference on the ML and Bayesian approaches only. Specimen Dm.bXI.profile cleaning-16857 was confidently placed in Bovinae, improving its morphology-based taxonomic assignment, i.e. “herbivore” (Extended Data Table 1; Fig. S15, Extended Data Fig. 6). In agreement with the taxonomic assignment based on morphology, specimens Dm.5/154.2.A4.38-16632 and Dm.8/152.3.B1.2-16641 were placed, respectively, in the Equidae, and in Bovidae. In the latter case only through Bayesian inference, though. Finally, specimens Dm.5/154.1.B1.1-16638 and Dm.8/154.4.A4.22-16639 were tentatively placed basal to or within the the Caprinae and Antilopinae subfamilies. This instability is either due to the recovery of an insufficient number of informative positions or the absence of reference sequences for particular subfamilies within the Bovidae.

Specimen Dm.5/157-16635, which had been previously assigned to the *Stephanorhinus* genus based on morphology (Supplementary Information §3), was placed within the Rhinocerotidae family. Furthermore, results from Bayesian and ML phylogenetic inference suggest that this specimen forms a clade with the Sumatran rhinoceros, the *Stephanorhinus kirchbergensis* and the woolly rhinoceros. More specifically, it falls basal to both the *S. kirchbergensis* and the woolly rhinoceros (Fig. S11).

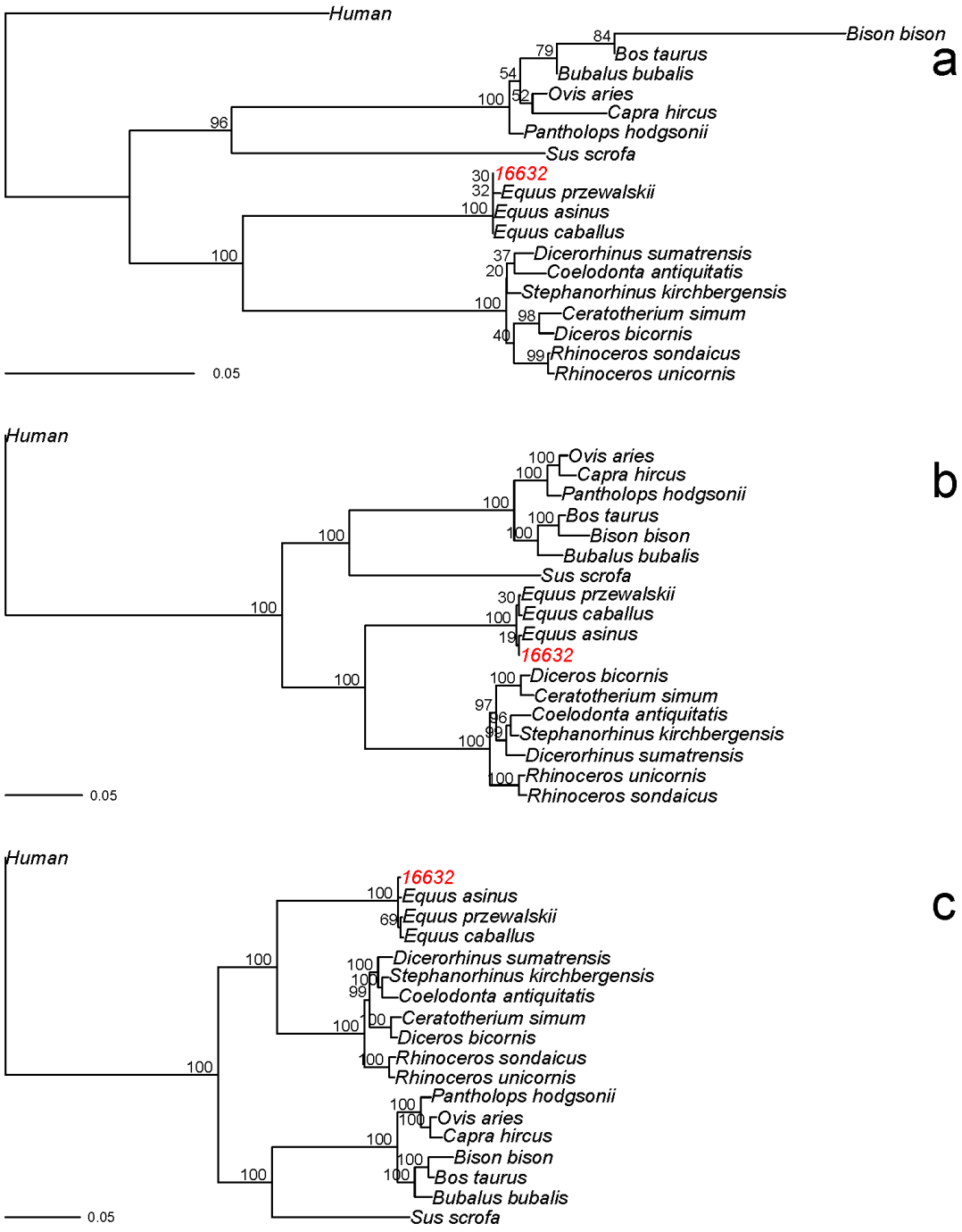

Figure S10. Phylogenetic relationships between the comparative reference dataset and specimen Dm.5/154.2.A4.38-16632. We carried out phylogenetic inference for each ancient specimen based on the concatenated alignments described in Sections 7 and 8. **a**. Neighbor-joining tree based on the modified concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through a non-parametric bootstrap procedure based on 500 replicates. **b**. Maximum likelihood tree based on the concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through 500 non-parametric bootstrap replicates. **c**. 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.

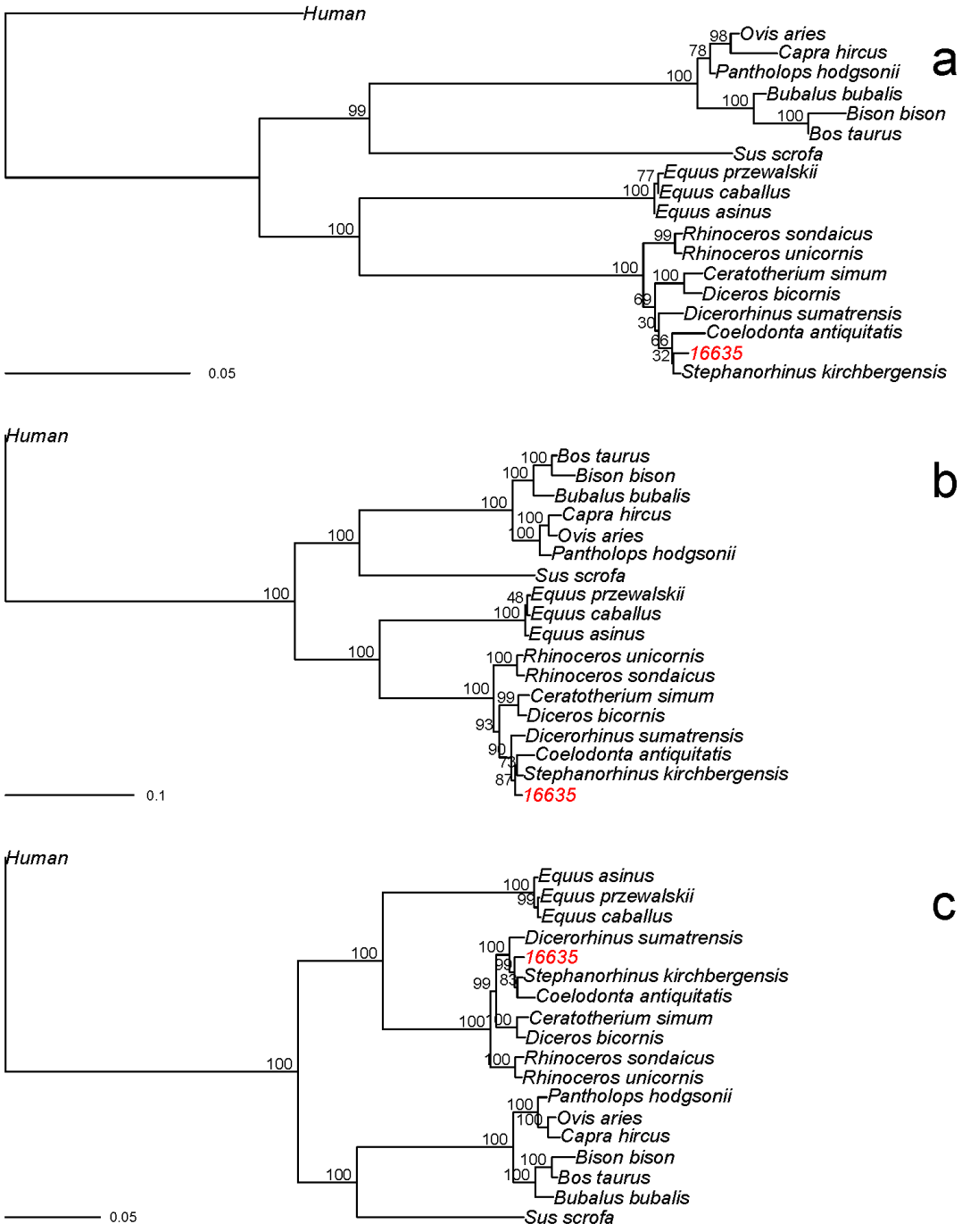

Figure S11. Phylogenetic relationships between the comparative reference dataset and specimen Dm.5/157-16635. We carried out phylogenetic inference for each ancient specimen based on the concatenated alignments described in Sections 7 and 8. **a**. Neighbor-joining tree based on the modified concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through a non-parametric bootstrap procedure based on 500 replicates. **b**. Maximum likelihood tree based on the concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through 500 non-parametric bootstrap replicates. **c**. 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. The Bayesian consensus tree in panel c is also reported, modified, in Fig. 4.

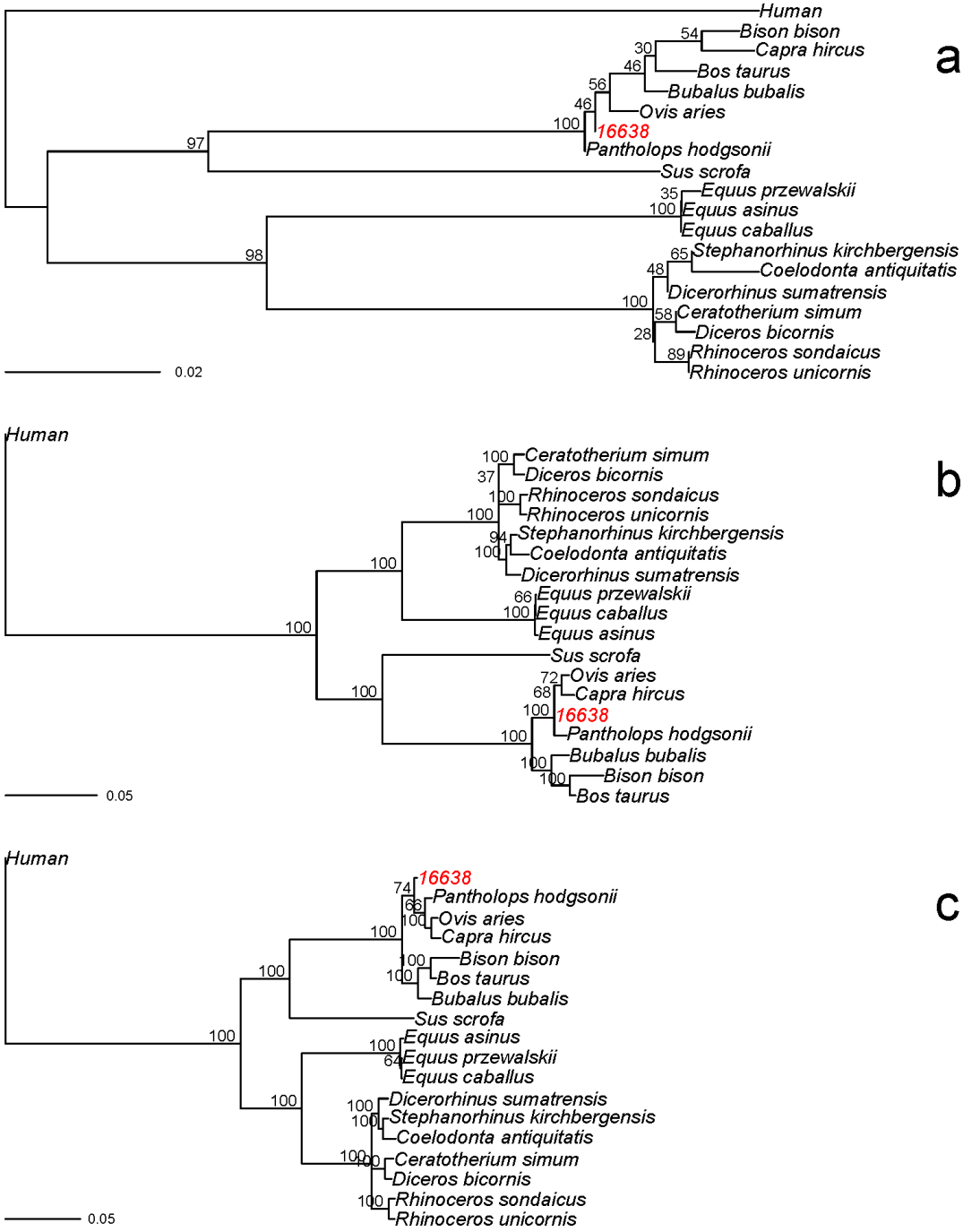

Figure S12. Phylogenetic relationships between the comparative reference dataset and specimen Dm.5/154.1.B1.1-16638. We carried out phylogenetic inference for each ancient specimen based on the concatenated alignments described in Sections 7 and 8. **a**. Neighbor-joining tree based on the modified concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through a non-parametric bootstrap procedure based on 500 replicates. **b**. Maximum likelihood tree based on the concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through 500 non-parametric bootstrap replicates. **c**. 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.

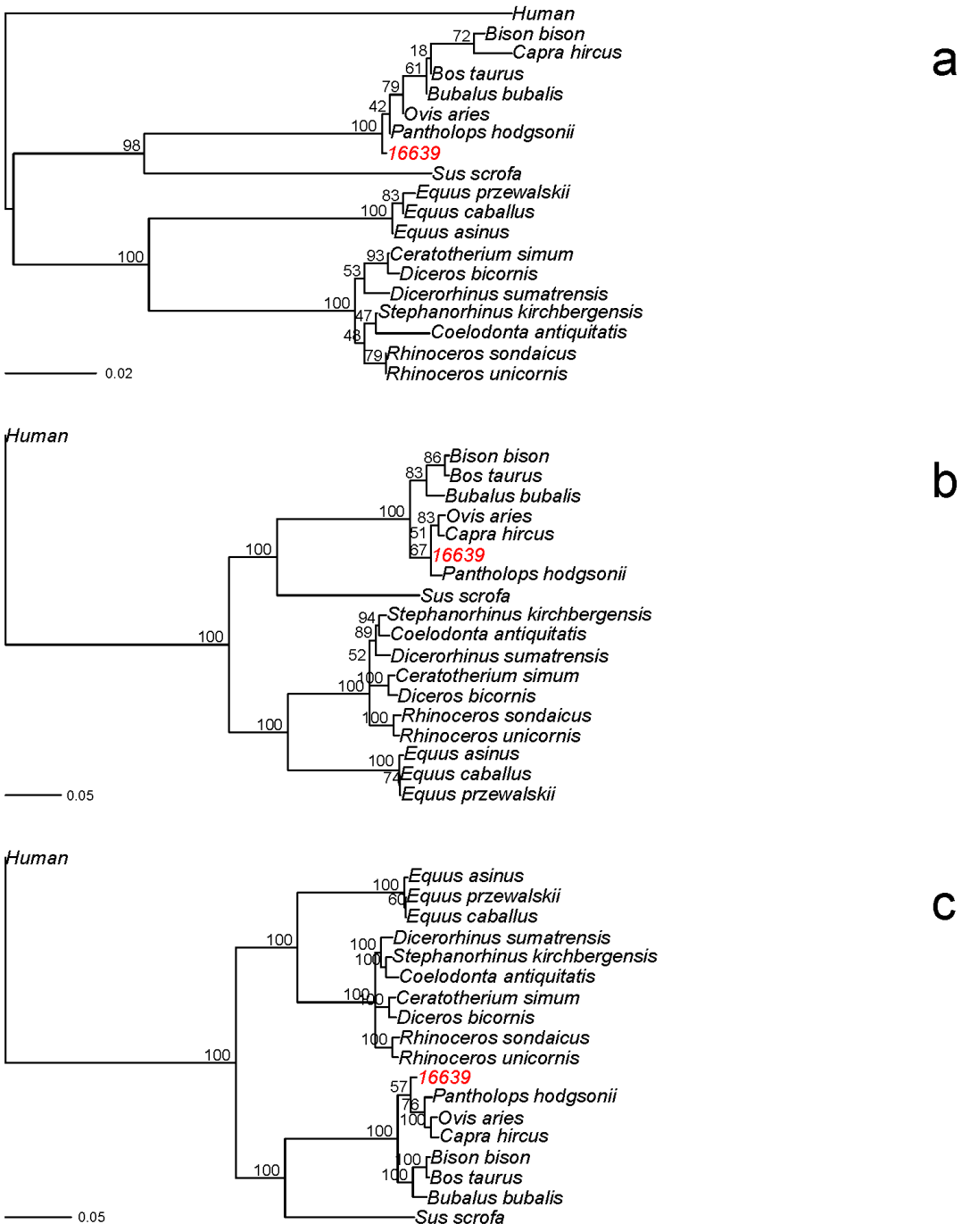

Figure S13. Phylogenetic relationships between the comparative reference dataset and specimen Dm.8/154.4.A4.22-16639. We carried out phylogenetic inference for each ancient specimen based on the concatenated alignments described in Sections 7 and 8. **a**. Neighbor-joining tree based on the modified concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through a non-parametric bootstrap procedure based on 500 replicates. **b**. Maximum likelihood tree based on the concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through 500 non-parametric bootstrap replicates. **c**. 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.

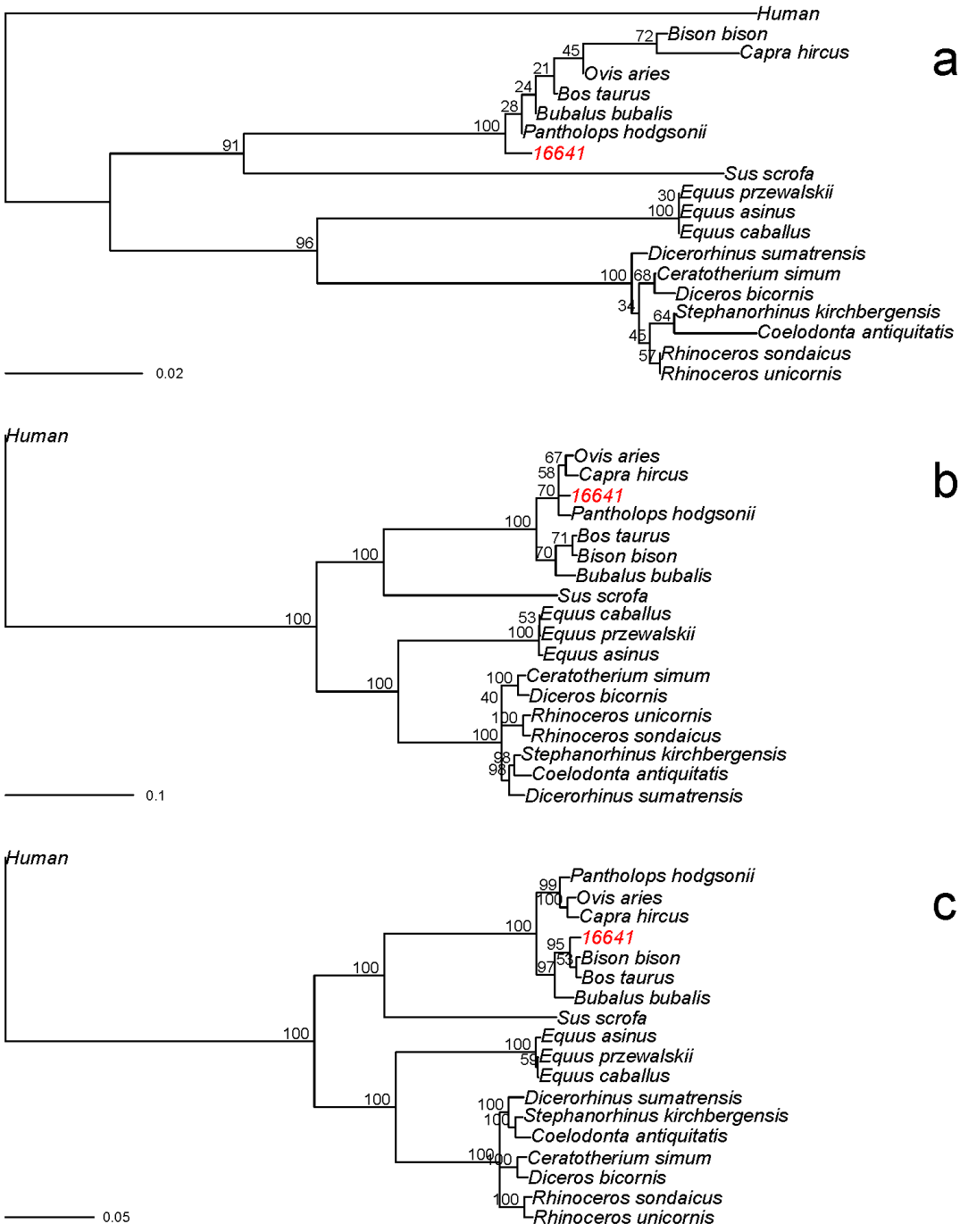

Figure S14. Phylogenetic relationships between the comparative reference dataset and specimen Dm.8/152.3.B1.2-16641. We carried out phylogenetic inference for each ancient specimen based on the concatenated alignments described in Sections 7 and 8. **a**. Neighbor-joining tree based on the modified concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through a non-parametric bootstrap procedure based on 500 replicates. **b**. Maximum likelihood tree based on the concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through 500 non-parametric bootstrap replicates. **c**. 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.

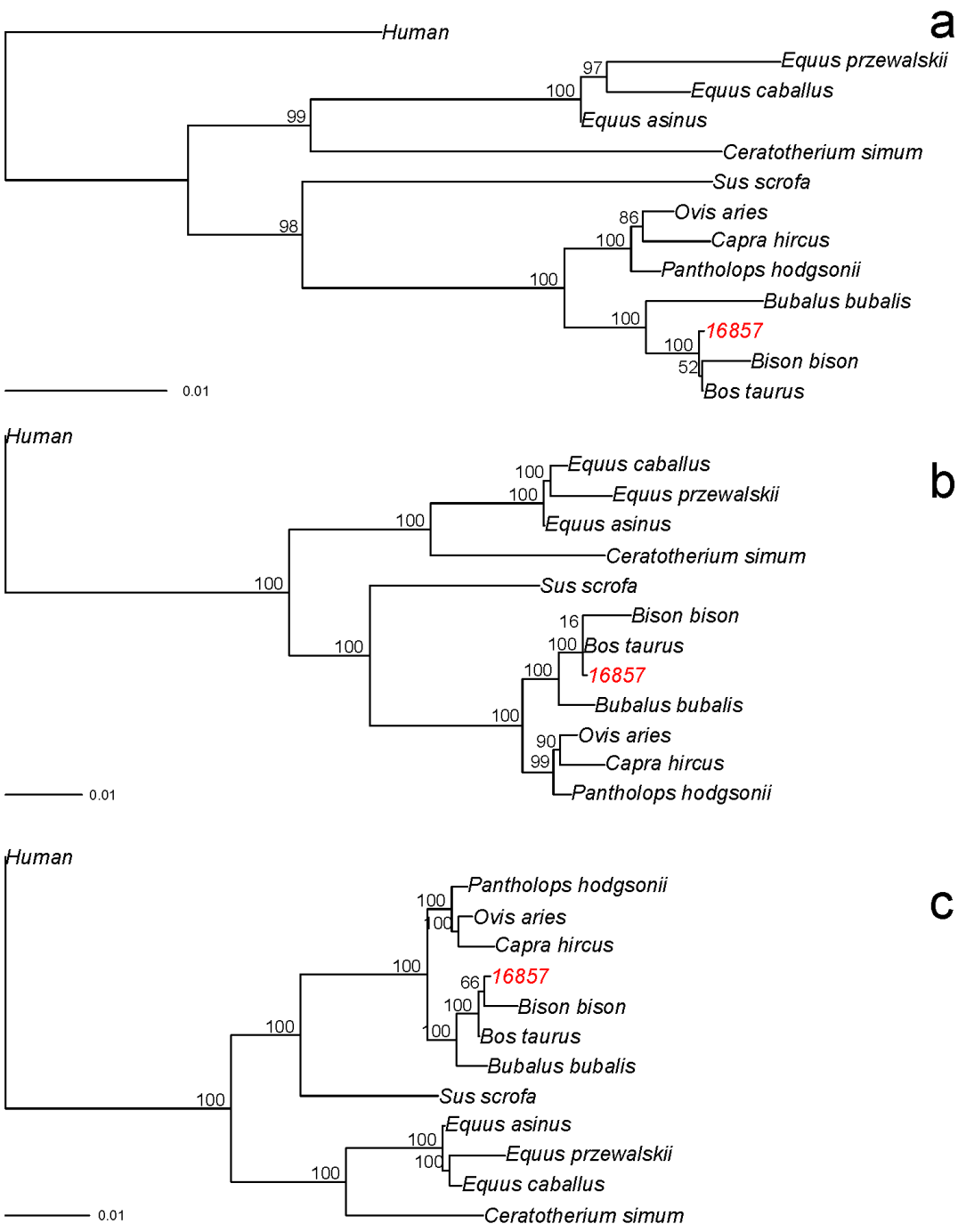

Figure S15. Phylogenetic relationships between the comparative reference dataset and specimen Dm.bXI.profile cleaning-16857. We carried out phylogenetic inference for each ancient specimen based on the concatenated alignments described in Sections 7 and 8. **a**. Neighbor-joining tree based on the modified concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through a non-parametric bootstrap procedure based on 500 replicates. **b**. Maximum likelihood tree based on the concatenated alignment. Bipartition support (percentage to the left of each node) was estimated through 500 non-parametric bootstrap replicates. **c**. 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. The Bayesian consensus tree in panel c is also reported in Extended Data Fig. 6.

13. Ancient DNA analysis

Morten E. Allentoft, Aurélien Ginolhac, Anders Krogh, Ludovic Orlando, Eske Willerslev

Very few reads were observed on the expected red deer and cow mitogenomes, and when mapping against nuclear genomes, more hits were obtained on the human reference compared to the cow and rhinoceros genomes (Tab. S16, Fig. S16).

DNA patterns, such as purine frequencies increased at read positions -1 and C>T at read extremities are proxies for ancient template authenticity[128]. The mapDamage tool[74] was run and the results provided no evidence for ancient DNA signatures for any set of retrieved sequences. Here are depicted the 3,241 hits unique to the White rhino genome after filtering for mapping quality 30: 1,782 hits (Fig. S16). Moreover, the read nucleotide composition does not fit the rhino genome. Additionally, we observe a high rate of insertions compared to the reference (in purple in Fig. S17) and an extremely short length, 99% of the 1,782 reads have a length < 29 bp, suggesting that some sequence alignments are spurious and likely incorrect. Based on our analyses, we conclude that no ancient DNA could be confidently retrieved using the techniques described in this study. Given our current knowledge on the rate of DNA decay[129], it is not a surprise that we cannot identify authentic >25 bp DNA fragments in ~1.77 Ma old skeletal remains from a site in a temperate climate zone. The DNA we have identified is likely to primarily represent small amounts of modern human DNA contamination, and DNA from the living microbes found in the soil at the Dmanisi site.

Table S16. Hits mapping after removing for PCR duplicates.

Library	Trimmed reads	Rhinoceros nuclear	Cow nuclear	Hits hg19	Cow mt	Deer mt
MA2481	3,879,288	4,486	4,540	4,955	0	0
MA_399_L1	8,629,253	NA	3,735	5,027	2	4
MA_399_L2	21,617,987	NA	2,503	2,890	18	13

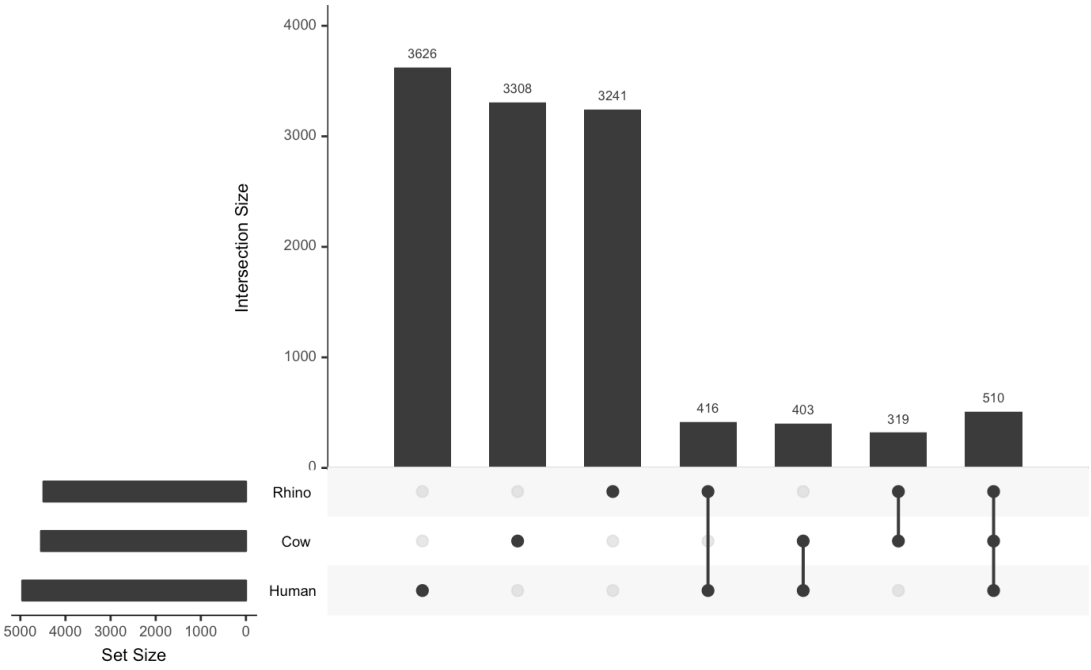

Figure S16. Number of reads from library MA2481, generated from dentine of specimen Dm.5/157-16635 (*Stephanorhinus* sp.), shared between the three nuclear genomes. Using the R package UpSetR[130].

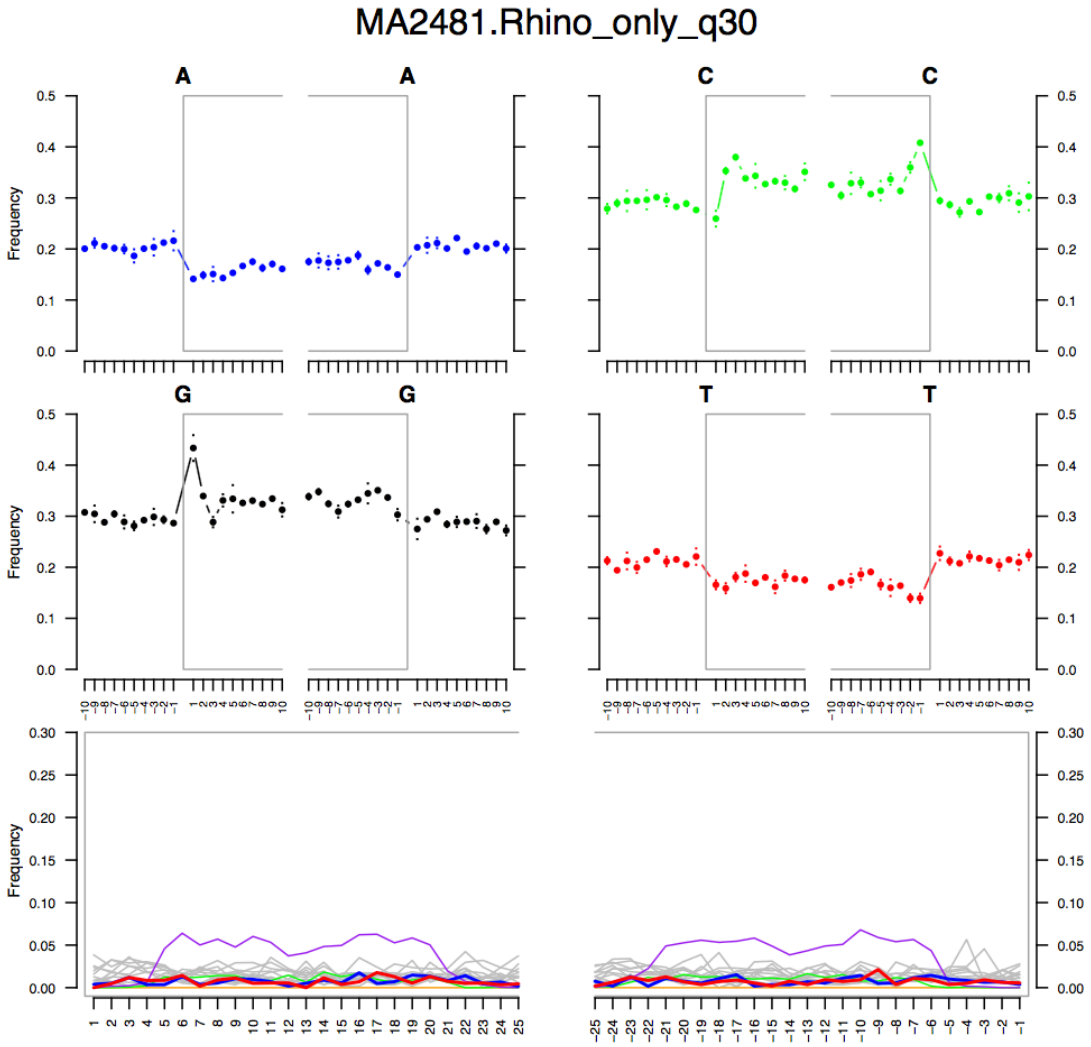

Figure S17. Lack of DNA damage signature. The analysis shows no increase of C-T deamination damage (red line) towards position 1 in the reads, as otherwise expected for authentic ancient DNA. Plot based on the 1,782 reads from MA2481 library that mapped to rhinoceros and not the Cow and Human genomes and with a mapping quality ≥ 30.

APPENDIX

14. Systematics of the genus *Stephanorhinus*

Luca Pandolfi, Lorenzo Rook

Stephanorhinus is an extinct genus of Rhinocerotidae, which characterized the Eurasian faunal assemblages during the Plio-Pleistocene time period. The debate on the taxonomic status this genus and its phylogenetic relationships is well documented in the literature[27-29,35,131-142]. During the 19th century and the first half of the 20th century, the European Pleistocene rhinoceroses, not included within the genus *Coelodonta*, were referred to the genus *Rhinoceros* (nowadays represented by the one-horned rhino *R. unicornis* and by the Java's rhino *R. sondaicus*) and later to the genus *Dicerorhinus* (nowadays represented by the Sumatran rhino *D. sumatrensis*), but also to *Ceratorhinus* and *Atelodus*[143-151].

The genus *Stephanorhinus* was established by Kretzoi[131,132] and included the species *Rhinoceros etruscus*, *R. megarhinus-leptorhinus*, *R. kirchbergensis* and *R. pikermiensis*. According to Kahlke (in Tong and Wu 2010[152]), the name *Stephanorhinus* derives from the name of King Stephen I of Hungary. Kretzoi indicated the species *R. etruscus* as holotype of the new genus but he did not report an exhaustive diagnosis for the new established taxon[131,132]. In addition, Kretzoi erected the new genus *Procerorhinus* for the Middle Pleistocene species *R./P. hemitoechus*[132]. *Stephanorhinus* and *Procerorhinus* were not immediately adopted by other authors.

The genus *Stephanorhinus* was adopted into the Rhinocerotidae systematic literature only thirty years after the Kretzoi's papers[133,136,153-155]. Guérin, in his monumental work on Rhinocerotidae, ascribed the European Pleistocene fossil rhinoceroses (except the woolly rhino) to the genus *Dicerorhinus* and the species *jeanvireti*, *etruscus*, *hemitoechus* and *mercki* to the new subgenus *Brandtorhinus*[29]. *Brandtorhinus* was later elevated to the genus rank[135]. Contrary to other authors, Guérin clearly rejected the names *Stephanorhinus* and *Procerorhinus*[29] because they were proposed "à la va-vite"[156] in a short paper of about five pages which did not report any interesting news and discussed "enrafale" sixteen different taxa. The genus *Stephanorhinus*, according to Guérin[29,134,156], has been defined by Kretzoi only on the basis of the reduction of the anterior teeth and the presence of a small rugosity for the frontal horn while the subgenus (and later genus) *Brandtorhinus* was established on a well-defined and detailed series of characters. In addition, according to Guérin[29,156], the morphological features listed by Kretzoi[131,132] as typical of the genus *Stephanorhinus* are not exclusive to this taxon and the development of the rugosity for the insertion of the frontal horn shows an high variability within Rhinocerotidae[29,156]. Consequently, Guérin suggested the combination *Dicerorhinus* (*Brandtorhinus*) for the fossil European Pleistocene rhinoceroses not included within *Coelodonta* or *Elasmotherium*[29]. However, the Pleistocene European rhinoceroses show little similarity with the genus *Dicerorhinus* and they are closer to *Rhinoceros*; therefore it is incorrect, from a phylogenetic point of view, to continue to refer these species to *Dicerorhinus*[136]. Geraads[157] suggested that *Stephanorhinus* is a paraphyletic group and *Coelodonta* represents the name for the monophyletic group which includes the species characterised by loss of anterior teeth, firmly closed external auditory pseudomeatus and backward-movement of the nasal notch. This hypothesis, however, is in need of a detailed discussion and requires further investigations. Despite these contributions, some authors continued to refer the fossil European Pleistocene rhinoceroses to the living genus *Dicerorhinus*[138,156].

In 1993, Fortelius *et al*.[137] suggested that *Brandtorhinus* is identical to *Stephanorhinus* (as previously proposed by Geraads[157]), and the latter name ‘‘is preferable to Guérin’s because it has priority and has been used quite extensively in literatures’’. Consequently, Fortelius referred the European Pleistocene rhinoceros species *jeanvireti*, *etruscus*, *hundsheimensis* (= *etruscus brachycephalus* in Guérin[29]), *hemitoechus* and *kirchbergensis* (= *mercki* in Guérin[29]) to the genus *Stephanorhinus*. This viewpoint was later followed by several other authors[35,158-160].

Deng et al.[161] recently referred the species *kirchbergensis* to *'Dihoplus'*, together with the Miocene species *ringstroemi* and *pikermiensis* and the Pliocene megarhinus. According to these results, morphologic characteristics distinctive of the genus *Stephanorhinus*, such as a partially ossified nasal septum and loss of functional incisors, evolved twice in the Pleistocene. The inclusion of *kirchbergensis* within the genus *'Dihoplus'*, however, seems to be unfounded and not supported by strong morphological evidence.

According to Pandolfi &Tagliacozzo[28], the comments reported by Guérin[29,156] on the invalidity of the genus *Stephanorhinus* are correct and well-founded; nevertheless, *Stephanorhinus* seems to be a valid name according to the ICZN rules, because the works of Kretzoi[131,132] satisfy the ICZN Articles 8, 11.8, 13.1, and 13.3.

14.1. Phylogeny of the genus *Stephanorhinus*

Phylogenetic relationships of the fossil Eurasian rhinoceroses within the family Rhinocerotidae have been studied by several scholars and have been investigated in several papers. Here we report the main studies and hypotheses published during the last forty years, while earlier phylogeny reconstructions involving the family Rhinocerotidae were exhaustively described by Prothero *et al*.[162] and references therein.

According to Guérin[29,134], the European Pleistocene fossil rhinoceroses are phylogenetically related with the living *Dicerorhinus sumatrensis*, and they emerged from Miocene-early Pliocene taxa belonging to the same genus. Here, *D. sumatrensis* and the Pleistocene European rhinoceroses should have a common ancestor at the end of the Miocene. According to the cladogram proposed by Guérin[29,134], *Coelodonta* emerged from the European dicerorhine and it is related to *Dicerorhinus* (*Brandtorhinus*) *etruscus* (=*Stephanorhinus etruscus*) and *D. (B.) hemitoechus*, and only slightly more distantly to *D. (B.) mercki* and *D. (B.) jeanvireti*. In the cladogram proposed by Guérin[134], the lineage composed by European Pleistocene dicerorhines and *Coelodonta* differs from that of *Dicerorhinus* by showing a reduction of anterior teeth and the ossification of the nasal septum. This latter character, however, is not exclusive to the genera *Coelodonta* and *Stephanorhinus* (= European dicerorhine in Guérin[29]) and occurs in several other groups. The presence of an ossification of the nasal septum has been firstly reported by Pocock[163] for the species *R. sondaicus* and *D. sumatrensis* and was later confirmed by Groves[136]. The ossification of the nasal septum has also been observed in *R. unicornis*[142].

Heissig, who recognized *Stephanorhinus* as a valid genus, proposed two alternative cladograms[153]. On both, he suggested a phyletic division between *Coelodonta* and *Stephanorhinus* and these two genera are differently located within the cladograms. On the first cladogram *Stephanorhinus* is sister to the clade composed by *Diceros* and *Ceratotherium*, whereas *Coelodonta* is sister to [*Eurhinoceros sondaicus* (= *Rhinoceros sondaicus*), (*Stephanorhinus* (*Diceros*, *Ceratotherium*)]. On the second cladogram *Stephanorhinus* is sister to *Didermocerus* (= *Dicerorhinus*) and this small clade is phylogenetically related with *Rhinoceros* (= *Rhinoceros unicornis*) and *Eurhinoceros* (= *Rhinoceros sondaicus*). In the latter cladogram, *Coelodonta* is sister to the clade composed by (*Rhinoceros*, *Eurhinoceros*) and (*Didermocerus* = *Dicerorhinus*, *Stephanorhinus*). According to these works, *Coelodonta* displays a reduced premolar cingulum, a more elongated skull with backwardly inclined occipital crest, and very hypsodont teeth, whereas *Stephanorhinus* shows more moralized premolars, implying that any similarities between these two taxa should be related to convergence.

Following Groves[136], *Stephanorhinus* shares many features with Rhinocerotini (*Rhinoceros* and *Dicerorhinus sensu* Groves[136]) such as an ossified nasal septum, firmly closed subaural channel (external auditory pseudomeatus), and developed metacone fold. Here, a few characters in *Stephanorhinus* suggest a divergence from Rhinocerotini such as a slope back occipital crest, an elongated radius, and an primitively oval foramen magnum. In addition, according to Groves[136], *Stephanorhinus* is closer to *Rhinoceros* than to *Dicerorhinus* in particular due to the presence of strong molarisation of the premolars, firm fusion of postglenoid and posttympanic processes and the great mastoid inflation. The Kretzoi's genus, however, also displays strong autapomorphic

features such as elongated nasal bones, backward-movement of the nasal notch, completely molarised premolars, total loss of incisors etc[136]. *Coelodonta* shares several morphological features with *Stephanorhinus* but lacks the strong moralization of the premolars, the foramen magnum is pear-shaped, the premaxilla is horizontal and the radius is fairly short. According to Groves[136], the parallelism between *Coelodonta* and *Stephanorhinus* suggested by Heissig[153] may be correct but further investigations were needed to confirm the hypothesis.

In the cladogram proposed by Prothero et al.[162], *Stephanorhinus* and *Coelodonta* are sister taxa sharing the following two characters: ossification of the nasal septum and loss of upper and lower incisors. These two taxa are closely related with the clade (*Gaindatherium*, *Punjabitherium*, *Rhinoceros*) than with *Dicerorhinus*. Nevertheless, the presence of *Elasmotherium* as sister of (*Stephanorhinus*, *Coelodonta*) and (*Gaindatherium*, *Punjabitherium*, *Rhinoceros*) would suggest some issues in the cladogram proposed by Prothero et al.[162]. *Elasmotherium* is indeed an extinct genus of large-sized rhinoceros belonging to the subfamily Elasmotheriinae[30,164], which is distantly related from the investigated group (the latter belonging to the subfamily Rhinocerotinae).

The cladogram proposed by Cerdeño[159] was based on a cladistic analysis of the family Rhinocerotidae encompassing 45 taxa and 72 craniomandibular, dental and postcranial characters. The cladogram shows *Stephanorhinus* as sister of *Coelodonta*, and the elasmotheres *Elasmotherium* and *Ninxiatherium* as sister group of the (*Stephanorhinus*, *Coelodonta*) clade. The African clade (*Paradiceros* (*Ceratotherium*, *Diceros*)) is the sister group of the clade composed by the previous two. The genus *Dicerorhinus* is instead the sister taxon of the clade composed by *Punjabitherium* and *Rhinoceros* and the major one composed by the previously listed taxa. The main synapomorphy shared by (*Ninxiatherium*, *Elasmotherium*) and (*Stephanorhinus*, *Coelodonta*) refers to the nasal septum partially or totally ossified, while other synapomorphies are also shared by *Iranotherium* and relatives[159]. An early acquisition of the nasal septum and loss of incisors occurred in the Miocene Elasmothere *Ninxiatherium*[159,165,166], which implies that these characteristics are not exclusive of *Stephanorhinus* and are independent acquisitions within different evolutionary lineages. However, according to Cerdeño[159], the closer relationship of *Ninxiatherium* with (*Stephanorhinus*, *Coelodonta*) is weakly supported, more than that of *Elasmotherium*. An inclusion of the *Stephanorhinus*-*Coelodonta* group in the subtribe Elasmotheriina together with *Elasmotherium*-*Ninxiatherium* group was therefore suggested.

The cladistic analysis performed by Deng et al.[161] included 17 taxa and 46 characters. The consensus tree shows that *Stephanorhinus* is sister to *Coelodonta* and that *Dihoplus* is sister to the clade composed by these two taxa. *Stephanorhinus etruscus* and *S. hundsheimensis* diverged early within the *Stephanorhinus*-*Coelodonta* clade. The extant Asian rhinoceroses (*D. sumatrensis* and the small clade of *R. unicornis* and *R. sondaicus*), are located in a trichotomy with the (*Dihoplus*, *Stephanorhinus*, *Coelodonta*) clade.

Welker et al.[167] recently performed proteomic analysis on collagen type I sequences and obtained a phylogenetic tree where the clade composed by the extant *R. unicornis* and *R. sondaicus* is sister to the clade composed by *Stephanorhinus* sp., *Coelodonta antiquitatis* and *Dicerorhinus sumatrensis*. Nevertheless, the relationships between *Stephanorhinus*, *Coelodonta* and *Dicerorhinus* were not resolved, indicating the limited phylogenetic reach of collagen type I sequencing.

Phylogenetic inference performed by Kirillova *et al*.[69], based on a complete mitochondrial genome sequence, places *Stephanorhinus kirchbergensis* as sister taxon to *Coelodonta*, and *Dicerorhinus sumatrensis* as sister of the clade composed by *Stephanorhinus* and *Coelodonta*. The *Rhinoceros* clade is here sister to (*Dicerorhinus* (*Stephanorhinus*, *Coelodonta*).

DATA DEPOSITION NOTE

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

SUPPLEMENTARY REFERENCES

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

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

3 Lordkipanidze, D. *et al.* Postcranial evidence from early Homo from Dmanisi, Georgia. *Nature* **449**, 305-310, doi:10.1038/nature06134 (2007).

4 Messager, E. *et al.* 40Ar/39Ar dating and phytolith analysis of the Early Pleistocene sequence of Kvemo-Orozmani (Republic of Georgia): chronological and palaeoecological implications for the hominin site of Dmanisi. *Quaternary Science Reviews* **30**, 3099-3108, doi:10.1016/j.quascirev.2011.07.008 (2011).

5 Tappen, M., Lordkipanidze, D., Bukhsianidze, M., Ferring, R. & Vekua, A. in *African Taphonomy: A Tribute to the Career of C.K. "Bob" Brain* (eds T. R. Pickering, K. Schick, & N. Toth) 119-135 (Oxford University Press, 2007).

6 Vekua, A. Die Wirbeltierfauna des Villafranchium von Dmanisi und ihre biostratigraphische Bedeutung. *Jahrbuch des Römisch-Germanischen Zentralmuseums Mainz* **42**, 77-180, pls.187-154 (1995).

7 Bukhsianidze, M. *The fossil Bovidae of Dmanisi* PhD thesis, Università degli Studi di Ferrara, (2005).

8 Bukhsianidze, M. in *Dmanisi V* (ed M. Tetradze) 132-160 (in Georgian, with Russian and English summaries) (Oazisi, 2006).

9 Bukhsianidze, M. & Vekua, A. Capra dalii nov. sp. (Caprinae, Bovidae, Mammalia) at the limit of Plio-Pleistocene from Dmanisi (Georgia). *Courier Forschungsinstitut Senckenberg* **256**, 159-171 (2006).

10 Palmqvist, P., Torregrosa, V., Pérez-Claros, J. A., Martínez-Navarro, B. & Turner, A. A re-evaluation of the diversity of Megantereon (Mammalia, Carnivora, Machairodontinae) and the problem of species identification in extinct carnivores. *Journal of Vertebrate Paleontology* **27**, 160-175, doi:10.1671/0272-4634(2007)27[160:AROTDO]2.0.CO;2 (2007).

11 Vekua, A. New Spiral-Horned Antelope in Dmanisi Fauna. *Bulletin of the Georgian National Academy of Sciences* **6**, 139-144 (2012).

12 Vekua, A., Lordkipanidze, D., Vanishvili, N. & Magradze, G. in *Problems of Paleobiology* Vol. III, 104-114 (in Georgian, with English summary) (2008).

13 Furió, M., Agustí, J., Mouskhelishvili, A., Sanisidro, Ó. & Santos-Cubedo, A. The paleobiology of the extinct venomous shrew Beremendia (Soricidae, Insectivora, Mammalia) in relation to the geology and paleoenvironment of Dmanisi (Early Pleistocene, Georgia). *Journal of Vertebrate Paleontology* **30**, 928-942, doi:10.1080/02724631003762930 (2010).

14 Hemmer, H., Kahlke, R. D. & Vekua, A. K. The cheetah Acinonyx pardinensis (Croizet et Jobert, 1828) s.l. at the hominin site of Dmanisi (Georgia) - A potential prime meat supplier in Early Pleistocene ecosystems. *Quaternary Science Reviews* **30**, 2703-2714, doi:10.1016/j.quascirev.2011.05.024 (2011).

15 Hemmer, H., Kahlke, R.-D. & Vekua, A. K. Panthera onca georgica ssp. nov. from the Early Pleistocene of Dmanisi (Republic of Georgia) and the phylogeography of jaguars (Mammalia, Carnivora, Felidae). *Neues Jahrbuch für Geologie und Paläontologie* **257**, 115-127 (2010).
16 Sotnikova, M. & Rook, L. Dispersal of the Canini (Mammalia, Canidae: Caninae) across Eurasia during the Late Miocene to Early Pleistocene. *Quat. Int.* **212**, 86-97, doi:10.1016/j.quaint.2009.06.008 (2010).
17 Rook, L. & Martínez-Navarro, B. Villafranchian: The long story of a Plio-Pleistocene European large mammal biochronologic unit. *Quaternary International* **219**, 134-144, doi:10.1016/j.quaint.2010.01.007 (2010).
18 Kahlke, R. D. *et al.* Western Palaearctic palaeoenvironmental conditions during the Early and early Middle Pleistocene inferred from large mammal communities, and implications for hominin dispersal in Europe. *Quaternary Science Reviews* **30**, 1368-1395, doi:10.1016/j.quascirev.2010.07.020 (2011).
19 Fejfar, O., Heinrich, W.-D. & Lindsay, E. H. in *The Dawn of the Quaternary* (eds T. Van Kolfschoten & P.L. Gibbard) 533-554 (Mededelingen Nederlands Instituut voor Toegepaste Geowetenschappen TNO 60, 1998).
20 Ali-Zade, A. A. *et al.* in *International colloquium on the Neogene Quaternary boundary.* 104 (INQUA Subcommission on the Pliocene-Quaternary Boundary, IUGS Subcommission on Neogene Stratigraphy, Geological Institute of the Academy of Sciences of the USSR, Department of Paleontology and Stratigrapby of the Academy of Sciences of Moldavian SSR, Institute of Paleobiology of the Academy of Sciences of Georgian SSR, Institute of Geology of the Academy of Sciences of Azerbaijan SSR, Azerbaijan Research Design Institute of Oil Prospecting).
21 Lebedeva, N. A. *Korrelyatsiya antropogenovyh tolshch Ponto-Kaspiya*. 135p. (in Russian). (Nauka, 1978).
22 Sotnikova, M. V. & Sablin, M. V. Late Villafranchian association of predators mammals from the locality Palan-Tukan (East Transcaucases, Azerbaijan). *Proceedings of Zoological Institute of the Russian Academy of Sciences* **249**, 134-135 (In Russian) (1993).
23 Espigares, M. P. *et al.* Homo vs. Pachycrocuta: Earliest evidence of competition for an elephant carcass between scavengers at Fuente Nueva-3 (Orce, Spain). *Quaternary International* **295**, 113-125, doi:10.1016/j.quaint.2012.09.032 (2013).
24 Martínez-Navarro, B. in *Out of Africa I: the First Hominin Colonization of Eurasia* (eds J.G. Fleagle *et al.*) Ch. 13, 207-224 (Springer, 2010).
25 Palmqvist, P. *et al.* The giant hyena Pachycrocuta brevirostris: Modelling the bone-cracking behavior of an extinct carnivore. *Quaternary International* **243**, 61-79, doi:10.1016/j.quaint.2010.12.035 (2011).
26 Behrensmeyer, A. K. Taphonomic and Ecologic Information from Bone Weathering. *Paleobiology* **4**, 150-162, doi:10.1017/S0094837300005820 (1978).
27 Pandolfi, L. & Erten, H. Stephanorhinus hundsheimensis (Mammalia, Rhinocerotidae) from the late early Pleistocene deposits of the Denizli Basin (Anatolia, Turkey). *Geobios* **50**, 65-73, doi:10.1016/j.geobios.2016.10.002 (2017).
28 Pandolfi, L. & Tagliacozzo, A. Stephanorhinus hemitoechus (Mammalia, Rhinocerotidae) from the Late Pleistocene of Valle Radice (Sora, Central Italy) and re-evaluation of the morphometric variability of the species in Europe. *Geobios* **48**, 169-191, doi:10.1016/j.geobios.2015.02.002 (2015).
29 Guérin, C. Les rhinocéros (Mammalia, Perissodactyla) du Miocène terminal au Pleistocène supérieur en Europe occidentale, comparaison avec les espèces actuelles. *Documents du Laboratoire de Geologie de la Faculte des Sciences de Lyon* **79**, 3-1183 (1980).
30 Antoine, P.-O. in *Mémoires du Muséum National d'Histoire Naturelle* Vol. 188 1-359 (Publications Scientifiques du Muséum, 2002).
31 Lacombat, F. Morphological and biometrical differentiation of the teeth from Pleistocene species of Stephanorhinus (Mammalia, Perissodactyla, Rhinocerotidae) in Mediterranean Europe and the Massif Central, France. *Palaeontographica A* **274**, 71-111 (2006).

32 Kahlke, H. D. Die Rhinocerotiden-Reste aus den Tonen von Voigtstedt in Thüringen. *Paläontologische Abhandlungen A* **2**, 451-520 (1965).
33 Kahlke, H. D. Die Rhinocerotiden-Reste aus den Kiesen von Süssenborn bei Weimar. *Paläontologische Abhandlungen A* **3**, 667-709 (1969).
34 Koenigswald, W., Smith, B. & Keller, T. Supernumerary teeth in a subadult rhino mandible (Stephanorhinus hundsheimensis) from the middle Pleistocene of Mosbach in Wiesbaden (Germany). *Paläontologische Zeitschrift* **81**, 416-428, doi:10.1007/BF02990253 (2007).
35 Lacombat, F. Les Rhinoceros fossiles des sites prehistoriques de l'Europe Mediterraneenne et du Massif Central - Paleontologie et implications biochronologiques. *British Archeological Reports* **1419**, 1-175 (2005).
36 Pandolfi, L., Grossi, F. & Frezza, V. New insight into the Pleistocene deposits of Monte delle Piche, Rome, and remarks on the biochronology of Hippopotamus (Mammalia, Hippopotamidae) and Stephanorhinus etruscus (Mammalia, Rhinocerotidae) in Italy. *Estudios Geologicos* **71**, e026, doi:10.3989/egeol.41796.337 (2015).
37 Penkman, K. E. H., Kaufman, D. S., Maddy, D. & Collins, M. J. Closed-system behaviour of the intra-crystalline fraction of amino acids in mollusc shells. *Quaternary Geochronology* **3**, 2-25, doi:10.1016/j.quageo.2007.07.001 (2008).
38 Kaufman, D. S. & Manley, W. F. A new procedure for determining DL amino acid ratios in fossils using reverse phase liquid chromatography. *Quaternary Science Reviews* **17**, 987-1000, doi:10.1016/S0277-3791(97)00086-3 (1998).
39 Sykes, G. A., Collins, M. J. & Walton, D. I. The significance of a geochemically isolated intracrystalline organic fraction within biominerals. *Organic Geochemistry* **23**, 1059-1065, doi:10.1016/0146-6380(95)00086-0 (1995).
40 Dickinson, M. L., A.; Penkman, K. A new method for enamel amino acid racemization dating: a closed system approach. *Quaternary Geochronology* **50**, 29-46, doi:10.1016/j.quageo.2018.11.005 (2019).
41 Hill, R. L. Hydrolysis of proteins. *Advances in Protein Chemistry* **20**, 37-107 (1965).
42 Preece, R. C. & Penkman, K. E. H. New faunal analyses and amino acid dating of the Lower Palaeolithic site at East Farm, Barnham, Suffolk. *Proceedings of the Geologists' Association* **116**, 363-377, doi:10.1016/S0016-7878(05)80053-7 (2005).
43 Hendy, J. *et al.* A guide to ancient protein studies. *Nature Ecology & Evolution* **2**, 791-799, doi:10.1038/s41559-018-0510-x (2018).
44 Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nature Methods* **6**, 359-362, doi:10.1038/nmeth.1322 (2009).
45 Cappellini, E. *et al.* Resolution of the type material of the Asian elephant, Elephas maximus Linnaeus, 1758 (Proboscidea, Elephantidae. *Zoological Journal of the Linnean Society* **170**, 222-232, doi:10.1111/zoj.12084 (2014).
46 Cappellini, E. *et al.* Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins. *Journal of Proteome Research* **11**, 917-926, doi:10.1021/pr200721u (2012).
47 Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663-670, doi:10.1021/ac0261171 (2003).
48 Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature Methods* **11**, 319-324, doi:10.1038/nmeth.2834 (2014).
49 Jersie-Christensen, R. R., Sultan, A. & Olsen, J. V. in *Phospho-Proteomics* Vol. 1355 *Methods in Molecular Biology* (ed L. von Stechow) 251-260 (Springer, 2016).
50 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).
51 Zhang, J. *et al.* PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular and Cellular Proteomics* **11**, M111.010587, doi:10.1074/mcp.M111.010587 (2012).

52 Welker, F. *et al.* Palaeoproteomic evidence identifies archaic hominins associated with the Châtelperronian at the Grotte du Renne. *Proceedings of the National Academy of Sciences* **113**, 11162-11167, doi:10.1073/pnas.1605834113 (2016).
53 Welker, F. Elucidation of cross-species proteomic effects in human and hominin bone proteome identification through a bioinformatics experiment. *BMC Evolutionary Biology* **18**, 23, doi:10.1186/s12862-018-1141-1 (2018).
54 Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649, doi:10.1093/bioinformatics/bts199 (2012).
55 Gabriels, R., Martens, L. & Degroeve, S. Updated MS2PIP web server delivers fast and accurate MS2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *bioRxiv*, 544965, doi:10.1101/544965 (2019).
56 Brademan, D. R., Riley, N. M., Kwiecien, N. W. & Coon, J. J. Interactive Peptide Spectral Annotator: A Versatile Web-Based Tool for Proteomic Applications. *Molecular & Cellular Proteomics*, mcp.TIR118.001209, doi:10.1074/mcp.TIR118.001209 (2019).
57 Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **30**, 918-920, doi:10.1038/nbt.2377 (2012).
58 Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved visualization of protein consensus sequences by iceLogo. *Nature Methods* **6**, 786-787, doi:10.1038/nmeth1109-786 (2009).
59 Mackie, M. *et al.* Palaeoproteomic Profiling of Conservation Layers on a 14th Century Italian Wall Painting. *Angewandte Chemie (International ed.)* **57**, 7369-7374, doi:10.1002/anie.201713020 (2018).
60 Han, X., He, L., Xin, L., Shan, B. & Ma, B. PeaksPTM: Mass Spectrometry-Based Identification of Peptides with Unspecified Modifications. *J. Proteome Res.* **10**, 2930-2936, doi:10.1021/pr200153k (2011).
61 Savitski, M., Nielsen, M. & Zubarev, R. ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures. *Molecular and Cellular Proteomics* **5**, 935-948, doi:10.1074/mcp.T500034-MCP200 (2006).
62 Pitulko, V. *et al.* The Yana RHS site: Humans in the Arctic before the Last Glacial Maximum. *Science* **303**, 52-56 (2004).
63 Yang, D. Y., Eng, B., Waye, J. S., Dudar, J. C. & Saunders, S. R. Improved DNA extraction from ancient bones using silica-based spin columns. *American Journal of Physical Anthropology* **105**, 539-543, doi:10.1002/(SICI)1096-8644(199804)105:4<539::AID-AJPA10>3.0.CO;2-1 (1998).
64 Brace, S. *et al.* Serial population extinctions in a small mammal indicate Late Pleistocene ecosystem instability. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 20532-20536, doi:10.1073/pnas.1213322109 (2012).
65 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, doi:10.1101/pdb.prot5448 (2010).
66 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows– Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
67 Kircher, M. in *Ancient DNA* Vol. 840 *Methods in Molecular Biology (Methods and Protocols)* (eds B. Shapiro & M. Hofreiter) 197-228 (Humana Press, 2012).
68 Li, H. *et al.* The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **25**, doi:10.1093/bioinformatics/btp352 (2009).
69 Kirillova, I. *et al.* Discovery of the skull of Stephanorhinus kirchbergensis (Jäger, 1839) above the Arctic Circle. *Quaternary Research* **88**, 537-550, doi:10.1017/qua.2017.53 (2017).
70 Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 15758, doi:10.1073/pnas.1314445110 (2013).

71 Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864, doi:10.1093/bioinformatics/btr026 (2011).
72 Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178, doi:10.1186/1471-2164-13-178 (2012).
73 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
74 Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682-1684, doi:10.1093/bioinformatics/btt193 (2013).
75 Orlando, L. *et al.* True single-molecule DNA sequencing of a pleistocene horse bone. *Genome Research* **21**, 1705-1719, doi:10.1101/gr.122747.111 (2011).
76 Gilbert, M. T. P. *et al.* Whole-Genome Shotgun Sequencing of Mitochondria from Ancient Hair Shafts. *Science* **317**, 1927-1930 (2007).
77 Ersmark, E. *et al.* Population Demography and Genetic Diversity in the Pleistocene Cave Lion. *Open Quaternary* **1**, 1-14, doi:10.5334/oq.aa (2015).
78 Carøe, C. *et al.* Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution* **9**, 410-419, doi:10.1111/2041-210X.12871 (2018).
79 Mak, S. S. T. *et al.* Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *GigaScience* **6**, 1-13, doi:10.1093/gigascience/gix049 (2017).
80 Chiu, T.-P. *et al.* GBshape: a genome browser database for DNA shape annotations. *Nucleic acids research* **43**, D103-D109, doi:10.1093/nar/gku977 (2015).
81 Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31, doi:10.1186/1471-2105-6-31 (2005).
82 Korneliussen, T., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356-356, doi:10.1186/s12859-014-0356-4 (2014).
83 SeaUrchinGenomeSequencingConsortium. The Genome of the Sea Urchin Strongylocentrotus purpuratus. *Science* **314**, 941-952 (2006).
84 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733-D745, doi:10.1093/nar/gkv1189 (2016).
85 Katoh, K. & Frith, M. C. Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics* **28**, 3144-3146, doi:10.1093/bioinformatics/bts578 (2012).
86 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593, doi:10.1093/bioinformatics/btq706 (2011).
87 Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).
88 Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology* **61**, 539-542, doi:10.1093/sysbio/sys029 (2012).
89 Willerslev, E. & Cooper, A. Ancient DNA. *Proceedings Of The Royal Society B* **272**, 3-16, doi:10.1098/rspb.2004.2813 (2005).
90 Gilbert, M. T. P., Bandelt, H. J., Hofreiter, M. & Barnes, I. Assessing ancient DNA studies. *Trends in Ecology and Evolution* **20**, 541-544, doi:10.1016/j.tree.2005.07.005 (2005).
91 Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74-78, doi:10.1038/nature12323 (2013).
92 Rohland, N. & Hofreiter, M. Comparison and optimization of ancient DNA extraction. *BioTechniques* **42**, 343-352, doi:10.2144/000112383 (2007).
93 Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols* **9**, 1056-1082, doi:10.1038/nprot.2014.063 (2014).

94 Varrette, S., Bouvry, P., Cartiaux, H. & Georgatos, F. in *2014 International Conference on High Performance Computing & Simulation.* (eds Waleed W. Smari & Vesna. Zeljkovic) 959-967 (IEEE and The Printing House, Inc.).
95 Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Research Notes* **5**, 337, doi:10.1186/1756-0500-5-337 (2012).
96 Kosnik, M. A. & Kaufman, D. S. Identifying outliers and assessing the accuracy of amino acid racemization measurements for geochronology: II. Data screening. *Quaternary Geochronology* **3**, 328-341, doi:10.1016/j.quageo.2008.04.001 (2008).
97 Tomiak, P. J. *et al.* Testing the limitations of artificial protein degradation kinetics using known-age massive Porites coral skeletons. *Quaternary Geochronology* **16**, 87-109, doi:10.1016/j.quageo.2012.07.001 (2013).
98 Penkman, K. E. H. *et al.* Testing the aminostratigraphy of fluvial archives: the evidence from intra-crystalline proteins within freshwater shells. *Quaternary Science Reviews* **26**, 2958-2969, doi:10.1016/j.quascirev.2007.06.034 (2007).
99 Fincham, A. G., Moradian-Oldak, J. & Simmer, J. P. The structural biology of the developing dental enamel matrix. *Journal Of Structural Biology* **126**, 270-299, doi:10.1006/jsbi.1999.4130 (1999).
100 Moradian-Oldak, J. Protein- mediated enamel mineralization. *Frontiers in bioscience* **17**, 1996-2023 (2012).
101 Hu, J. C. C., Chun, Y. H. P., Al Hazzazzi, T. & Simmer, J. P. Enamel Formation and Amelogenesis Imperfecta. *Cells Tissues Organs* **186**, 78-85 (2007).
102 Lu, Y. *et al.* Functions of KLK4 and MMP-20 in dental enamel formation. *Biological chemistry* **389**, 695-700, doi:10.1515/BC.2008.080 (2008).
103 Castiblanco, G. A. *et al.* Identification of proteins from human permanent erupted enamel. *European Journal of Oral Sciences* **123**, 390-395, doi:10.1111/eos.12214 (2015).
104 Nielsen-Marsh, C. M. *et al.* Extraction and sequencing of human and Neanderthal mature enamel proteins using MALDI-TOF/TOF MS. *Journal of Archaeological Science* **36**, 1758-1763, doi:10.1016/j.jas.2009.04.004 (2009).
105 Stewart, N. A., Gerlach, R. F., Gowland, R. L., Gron, K. J. & Montgomery, J. Sex determination of human remains from peptides in tooth enamel. *Proceedings of the National Academy of Sciences of the United States* **114**, 13649-13654, doi:10.1073/pnas.1714926115 (2017).
106 Santos, F. R., Pandya, A. & Tyler-Smith, C. Reliability of DNA-based sex tests. *Nature Genetics* **18**, 103, doi:10.1038/ng0298-103 (1998).
107 Skoglund, P., Storå, J., Götherström, A. & Jakobsson, M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *Journal of Archaeological Science* **40**, 4477-4482, doi:10.1016/j.jas.2013.07.004 (2013).
108 Stewart, N. A. *et al.* The identification of peptides by nanoLC-MS/MS from human surface tooth enamel following a simple acid etch extraction. *RSC Advances* **6**, 61673-61679, doi:10.1039/c6ra05120k (2016).
109 Salido, E. C., Yen, P. H., Koprivnikar, K., Yu, L. C. & Shapiro, L. J. The human enamel protein gene amelogenin is expressed from both the X and the Y chromosomes. *American Journal of Human Genetics* **50**, 303-316 (1992).
110 Lasa-Benito, M., Marin, O., Meggio, F. & Pinna, L. A. Golgi apparatus mammary gland casein kinase: monitoring by a specific peptide substrate and definition of specificity determinants. *FEBS Letters* **382**, 149-152, doi:10.1016/0014-5793(96)00136-6 (1996).
111 Tagliabracci, V. S. *et al.* Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. *Science* **336**, 1150-1153, doi:10.1126/science.1217817 (2012).
112 Catak, S., Monard, G., Aviyente, V. & Ruiz-López, M. F. Computational Study on Nonenzymatic Peptide Bond Cleavage at Asparagine and Aspartic Acid. *The Journal of Physical Chemistry A* **112**, 8752-8761, doi:10.1021/jp8015497 (2008).
113 Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **11**, 2301-2319, doi:10.1038/nprot.2016.136 (2016).

114 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207-214, doi:10.1038/nmeth1019 (2007).
115 Neilson, K. A. *et al.* Less label, more free: Approaches in label-free quantitative mass spectrometry. *Proteomics* **11**, 535-553, doi:10.1002/pmic.201000553 (2011).
116 Azmitia, E. C. in *Handbook of Behavioral Neuroscience* Vol. 21 (eds Christian P. Müller & Barry L. Jacobs) Ch. 1.1, 3-22 (Elsevier, 2010).
117 Enache, T. A. & Oliveira-Brett, A. M. Pathways of Electrochemical Oxidation of Indolic Compounds. *Electroanalysis* **23**, 1337-1344, doi:10.1002/elan.201000671 (2011).
118 Puddick, J. *et al.* Structural Characterization of New Microcystins Containing Tryptophan and Oxidized Tryptophan Residues. *Marine Drugs* **11**, 3025-3045, doi:10.3390/md11083025 (2013).
119 Plowman, J. E., Deb-Choudhury, S., Grosvenor, A. J. & Dyer, J. M. Protein oxidation: identification and utilisation of molecular markers to differentiate singlet oxygen and hydroxyl radical-mediated oxidative pathways. *Photochemical & Photobiological Sciences* **12**, 1960-1967, doi:10.1039/C3PP50182E (2013).
120 Grosvenor, A. J., Morton, J. D. & Dyer, J. M. Profiling of residue-level photo-oxidative damage in peptides. *Amino Acids* **39**, 285-296, doi:10.1007/s00726-009-0440-7 (2010).
121 Solazzo, C. *et al.* Application of redox proteomics to the study of oxidative degradation products in archaeological wool. *Journal of Cultural Heritage* **16**, 896-903, doi:10.1016/j.culher.2015.02.006 (2015).
122 Vestling, M. M., Kelly, M. A., Fenselau, C. & Costello, C. E. Optimization by mass spectrometry of a tryptophan-specific protein cleavage reaction. *Rapid Communications in Mass Spectrometry* **8**, 786-790, doi:10.1002/rcm.1290080925 (1994).
123 Kerwin, B. A. & Remmele, R. L., Jr. Protect from Light: Photodegradation and Protein Biologics. *Journal of Pharmaceutical Sciences* **96**, 1468-1479, doi:10.1002/jps.20815 (2007).
124 Schöneich, C. Novel chemical degradation pathways of proteins mediated by tryptophan oxidation: tryptophan side chain fragmentation. *Journal of Pharmacy and Pharmacology* **70**, 655-665, doi:10.1111/jphp.12688 (2017).
125 Rabbani, N. & Thornalley, Paul J. Dicarbonyls linked to damage in the powerhouse: glycation of mitochondrial proteins and oxidative stress. *Biochemical Society Transactions* **36**, 1045-1050 (2008).
126 Sell, D. R. & Monnier, V. M. Conversion of arginine into ornithine by advanced glycation in senescent human collagen and lens crystallins. *The Journal of biological chemistry* **279**, 54173-54184, doi:10.1074/jbc.M408946200 (2004).
127 Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical Reviews* **113**, 2343-2394, doi:10.1021/cr3003533 (2013).
128 Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 14616-14621, doi:10.1073/pnas.0704665104 (2007).
129 Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences* **279**, 4724-4733, doi:10.1098/rspb.2012.1745 (2012).
130 Lex, A., Strobelt, H., Vuillemot, R., Pfister, H. & Gehlenborg, N. UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics* **20**, 1983-1992, doi:10.1109/TVCG.2014.2346248 (2014).
131 Kretzoi, M. Präokkupierte und durch ältere zu ersetzende Säugetiernamen. *Földtani Közlöni, Budapest* **LXXII (4-12)**, 345-349 (1942).
132 Kretzoi, M. Bemerkungen zur System der Nachmiozänen Nashorn-Gattungen. *Földtani Közlöni, Budapest* **LXXII (4-12)**, 309-318 (1942).
133 Heissig, K. Die Unterfamilien und Tribus der rezenten und fossilen Rhinocerotidae (Mammalia). *Saugetierkundliche Mitteilungen* **21**, 25-30 (1973).

134 Guérin, C. Premiere biozonation du Pléistocène Européen, principal resultat biostratigraphique de l'etude des Rhinocerotidae (Mammalia, Perissodactyla) du Miocene Terminal au Pléistocène supérieur d'Europe occidentale. *Geobios* **15**, 593-598 (1982).
135 Guérin, C. La famille des Rhinocerotidae. *Cranium* **6**, 3-14 (1989).
136 Groves, C. P. Phylogeny of the living species of Rhinoceros. *Journal of Zoological Systematics and Evolutionary Research* **21**, 293-313, doi:10.1111/j.1439-0469.1983.tb00297.x (1983).
137 Fortelius, M., Mazza, P. & Sala, B. Stephanorhinus (Mammalia: Rhinocerotidae) of the Western European Pleistocene, with a revision of S. etruscus (Falconer, 1868). *Palaeontographia Italica, Pisa* **80**, 63-155 (1993).
138 MacKenna, M. C. & Bell, S. K. *Classification of mammals above the species level*. (Columbia University Press, 1997).
139 van der Made, J. The rhinos from the Middle Pleistocene of Neumark-Nord (Saxony-Anhalt). *Veröffentlichungen des Landesamtes für Archäologie* **62**, 432-527 (2010).
140 Pandolfi, L. Rhinocerotidae (Mammalia, Perissodactyla) from the Middle Pleistocene site of Ponte Milvio, central Italy. *Bollettino della Società Paleontologica Italiana, Modena* **52**, 219-229, doi:10.4435/BSPI.2013.25 (2013).
141 Pandolfi, L. & Petronio, C. Stephanorhinus etruscus (Falconer, 1868) from Pirro Nord (Apricena, Foggia, Southern Italy) with notes on the late Early Pleistocene rhinoceroses of Italy. *Rivista Italiana di Paleontologia e Stratigrafia* **117**, 173-187, doi:10.13130/2039-4942/5969 (2011).
142 Antoine, P. O. & Saraç, G. Rhinocerotidae (Mammalia, Perissodactyla) from the late Miocene of Akkaşdağı, Turkey. *Geodiversitas* **27**, 601-632 (2005).
143 Falconer, H. in *Palaeontological Memoirs and Notes of the late Hugh Falconer* Vol. 2 (ed Charles Murchison) 309-403 (Spottiswoode and Co., 1868).
144 Vialli, V. *Sul rinoceronte e l'elefante dei livelli superiori della serie lacustre di Leffe (Bergamo)*. Vol. Atti del Museo Civico di Storia Naturale di Milano, 12 fasc. 1 1-71 (Museo civico di storia naturale di Milano, 1956).
145 Osborn, H. F. Phylogeny of the rhinoceroses of Europe. *Bulletin of the American Museum of Natural History* **13**, 229-267 (1900).
146 Toula, F. Das Nashörn von Hundsheim: Rhinoceros (Ceratorhinus Osborn) hundsheimensis nov. form.: mit Ausfuhrungen uber die Verhaltnisse von elf Schadeln von Rhinoceros (Ceratorhinus) sumatrensis. *Abhandlungen der Geologischen Reichsanstalt* **19**, 1-92 (1902).
147 Schroeder, H. *Die Wirbelthier-Fauna des Mosbacher Sandes. – 1. Gattung Rhinoceros*. Vol. nF, Heft 18 1-143 (Königliche Geologischen Landesanstalt, 1903).
148 Schroeder, H. *Über Rhinoceros mercki und seine nord- und mitteldeutschen Fundstellen*. Vol. nF, Heft 124 1-114 (Königliche Geologischen Landesanstalt, 1930).
149 Staesche, K. *Nashörner der Gattung Dicerorhinus aus Diluvium dem Württembergs*. Vol. 200 1-148 (Königliche Geologischen Landesanstalt, 1941).
150 Simpson, G. G. The principles of classification and a classification of mammals [Suborder Ceratomorpha]. *Bulletin of the American Museum of Natural History* **85**, 14-258 (1962).
151 Dzhafarov, R. D. in *Trudy Estest.-Istoricheskogo Muzeya im. G. Zardaby* Vol. IV 65-88 (Izd-vo AN AzerbSSR, 1955).
152 Tong, H. & Wu, X. Stephanorhinus kirchbergensis (Rhinocerotidae, Mammalia) from the Rhino Cave in Shennongjia, Hubei. *Chinese Science Bulletin* **55**, 1157-1168, doi:10.1007/s11434-010-0050-5 (2010).
153 Heissig, K. Probleme bei der cladistichen Analyse einer Gruppe mit wenigen eindeutigen Apomorphien – Rhinocerotidae [Problems with the cladistic analysis of a group with few unequivocal apomorphies – Rhinocerotidae] [in German, English summ.]. *Paläontologische Zeitschriften* **55**, 117-123 (1981).
154 Sickenberg, O. Eine Saugertierfauna des tieferen Biharius aus dem Becken von Megalopolis (Peloponnes, Greichenland). *Annales Geologiques des Pays Helleniques* **27**, 25-73 (1976).

155 Prothero, D. R. & Schoch, R. M. in *The Evolution of perissodactyls* (eds D.R. Prothero & R.M. Schoch) 530-537 (Claredon Press and Oxford University Press, 1989).
156 Guérin, C. Les rhinocéros (Mammalia, Perissodactyla) du gisement villafranchien moyen de Saint-Vallier (Drôme). *Geobios* **37**, S259-S278, doi:10.1016/S0016-6995(04)80018-4 (2004).
157 Geraads, D. Révision des Rhinocerotinae (Mammalia) du Turolien de Pikermi: comparaison avec les formes voisines. *Annales de Paléontologie* **74**, 13-41 (1988).
158 Cerdeño, E. Remarks on the spanish Plio-Pleistocene Stephanorhinus etruscus (Rhinocerotidae). *Comptes Rendus de l'Académie des Sciences* **317**, 1363-1367 (1993).
159 Cerdeño, E. Cladistic analysis of the family Rhinocerotidae (Perissodactyla). *American Museum Novitates* **3143**, 1-25 (1995).
160 Cerdeño, E. Diversity and evolutionary trends of the Family Rhinocerotidae (Perissodactyla). *Palaeogeography, Palaeoclimatology, Palaeoecology* **141**, 13-34, doi:10.1016/S0031-0182(98)00003-0 (1998).
161 Deng, T. *et al.* Out of Tibet: pliocene woolly rhino suggests high-plateau origin of Ice Age megaherbivores. *Science* **333**, 1285-1288, doi:10.1126/science.1206594 (2011).
162 Prothero, D. R., Manning, E. & Hanson, C. B. The phylogeny of the Rhinocerotoidea (Mammalia, Perissodactyla). *Zoological Journal of the Linnean Society* **87**, 341-366, doi:10.1111/j.1096-3642.1986.tb01340.x (1986).
163 Pocock, R. I. The nasal septum in existing Asiatic rhinoceroses. *Annals and Magazine of Natural History* **11**, 341-344, doi:10.1080/00222934508654729 (1945).
164 Antoine, P. O. Middle Miocene elasmotheriine Rhinocerotidae from China and Mongolia: taxonomic revision and phylogenetic relationships. *Zoologica Scripta* **32**, 95-118, doi:10.1046/j.1463-6409.2003.00106.x (2003).
165 Chen, G. A new genus of Iranotheriinae of Ningxia. *Vertebrata PalAsiatica* **15**, 143-147 (1977).
166 Deng, T. A new elasmothere (Perissodactyla, Rhinocerotidae) from the late Miocene of the Linxia Basin in Gansu, China. *Geobios* **41**, 719-728, doi:10.1016/j.geobios.2008.01.006 (2008).
167 Welker, F. *et al.* Middle Pleistocene protein sequences from the rhinoceros genus Stephanorhinus and the phylogeny of extant and extinct Middle/Late Pleistocene Rhinocerotidae. *PeerJ* **5**, e3033, doi:10.7717/peerj.3033 (2017).