

Phylogenetically informative proteins from an Early Miocene rhinocerotid

<https://doi.org/10.1038/s41586-025-09231-4>

Received: 7 June 2024

Accepted: 3 June 2025

Published online: 9 July 2025

Open access

 Check for updates

Ryan S. Paterson^{1✉}, Meaghan Mackie^{1,2}, Alessio Capobianco^{3,4}, Nicola S. Heckeberg^{3,4}, Danielle Fraser^{5,6,7,8✉}, Beatrice Demarchi⁹, Fazeelah Munir¹⁰, Ioannis Patramanis¹, Jazmin Ramos-Madrigal¹, Shanlin Liu¹¹, Abigail D. Ramsøe¹, Marc R. Dickinson¹⁰, Chloë Baldreki¹⁰, Marisa Gilbert⁵, Raffaele Sardella¹², Luca Bellucci¹³, Gabriele Scorrano^{11,14}, Michela Leonardi^{15,16}, Andrea Manica¹⁵, Fernando Racimo¹, Eske Willerslev^{1,16,17}, Kirsty E. H. Penkman¹⁰, Jesper V. Olsen², Ross D. E. MacPhee¹⁸, Natalia Rybczynski^{5✉}, Sebastian Höhna^{3,4} & Enrico Cappellini^{1✉}

In the past decade, ancient protein sequences have emerged as a valuable source of data for deep-time phylogenetic inference^{1–4}. Still, even though ancient proteins have been reported from the Middle–Late Miocene^{5,6}, the recovery of protein sequences providing subordinal-level phylogenetic insights does not exceed 3.7 million years ago (Pliocene)¹. Here, we push this boundary back to 21–24 million years ago (Early Miocene) by retrieving enamel protein sequences of a rhinocerotid (*Epiaceratherium* sp.; CMNFV59632) from Canada’s High Arctic. We recover partial sequences of seven enamel proteins and more than 1,000 peptide–spectrum matches, spanning at least 251 amino acids. Endogeneity is in line with thermal age estimates and is supported by indicators of protein damage, including several spontaneous and irreversible chemical modifications accumulated during prolonged diagenesis. Bayesian tip-dating places the divergence time of CMNFV59632 in the Middle Eocene–Oligocene, coinciding with a phase of high rhinocerotid diversification⁷. This analysis identifies a later Oligocene divergence for Elasmotheriinae, weakening alternative models suggesting a deep basal split between Elasmotheriinae and Rhinocerotinae^{8,9}. The findings are consistent with hypotheses on the origin of the enigmatic fauna of the Houghton Crater, which, in spite of considerable endemism, has similarity to distant Eurasian faunas^{10,11}. Our findings demonstrate the potential of palaeoproteomics in obtaining phylogenetic information from a specimen that is approximately ten times older than any sample from which endogenous DNA has been obtained so far.

Phylogenetic placement of deep-time (>1 million years ago (Ma)) fossils has typically relied on morphological observations, because the recovery of sufficiently extensive genetic evidence has not been proven to be possible before the Pleistocene¹². Although ancient DNA (aDNA) sequences are often a valuable source of data for inferring phylogenies and population dynamics in the Middle–Late Pleistocene^{13–15}, the oldest authentic aDNA from macrofossils has been extracted from Arctic-situated specimens dated to no more than 1.2 Ma (ref. 16). By contrast, palaeoproteomic data have been recovered from Middle–Late Miocene, Pliocene and Early Pleistocene fossils, even in localities that are warm, humid and/or at low latitudes^{5,6,17}. Although protein sequences from the Early Pleistocene have been used successfully

to infer the phylogenetic placement of various fossil mammals^{2–4}, the precise limit of proteomic survival has not been systematically characterized yet, because it depends on a complex interplay of time, temperature and environmental factors driving chemical breakdown mechanisms^{18,19}. At present, the oldest confirmed palaeoproteomic data successfully used to infer subordinal taxonomic relationships derive from bone collagen of camelids from the 3.7-million-year-old Fyles Leaf Bed site of Canada’s High Arctic^{1,20}. Beyond this time frame and latitude, only peptide sequences too short to be genetically informative^{5,6} and the products of advanced diagenesis^{21–25} have been reported.

Rhinocerotidae is a family that includes only five extant species, but a wide diversity of fossil members^{7,26}. It remains debated as to where

¹Globe Institute, University of Copenhagen, Copenhagen, Denmark. ²Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark. ³GeoBio-Center LMU, Ludwig-Maximilians-Universität München, Munich, Germany. ⁴Department of Earth and Environmental Sciences, Palaeontology & Geobiology, Ludwig-Maximilians-Universität München, Munich, Germany. ⁵Palaeobiology, Canadian Museum of Nature, Ottawa, Ontario, Canada. ⁶Department of Earth Sciences, Carleton University, Ottawa, Ontario, Canada. ⁷Department of Biology, Carleton University, Ottawa, Ontario, Canada. ⁸Department of Paleobiology, Smithsonian Institution, National Museum of Natural History, Washington, DC, USA. ⁹Department of Life Sciences and Systems Biology, University of Turin, Turin, Italy. ¹⁰Department of Chemistry, University of York, York, UK. ¹¹Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. ¹²Department of Earth Sciences (PaleoFactory lab.), Sapienza University of Rome, Rome, Italy. ¹³Museo di Geologia e Paleontologia, Sistema Museale di Ateneo, Università di Firenze, Firenze, Italy. ¹⁴Center for Molecular Anthropology for the Study of Ancient DNA, Department of Biology, University of Rome “Tor Vergata”, Rome, Italy. ¹⁵Department of Zoology, University of Cambridge, Cambridge, UK. ¹⁶Department of Genetics, University of Cambridge, Cambridge, UK. ¹⁷MARUM, University of Bremen, Bremen, Germany. ¹⁸Department of Mammalogy, American Museum of Natural History, New York, NY, USA. ✉e-mail: ryan.paterson@sund.ku.dk; dfraser@nature.ca; nrybczynski@nature.ca; ecappellini@sund.ku.dk

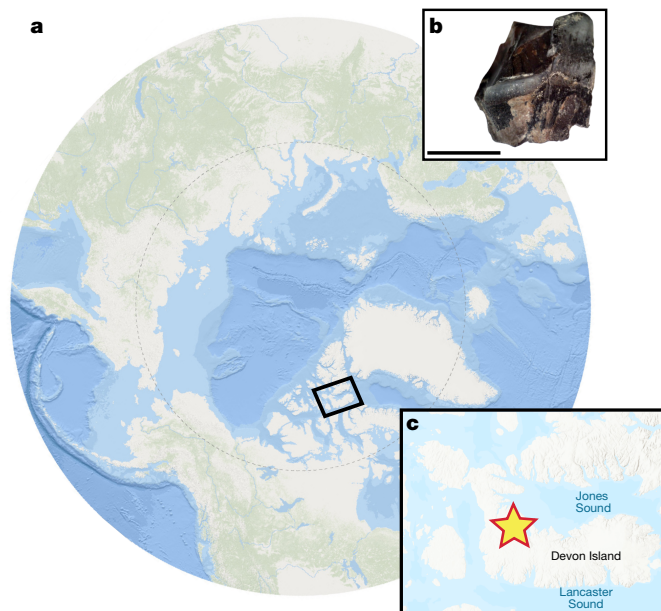


Fig. 1 | The high-latitude Houghton Crater on Devon Island has produced a highly endemic vertebrate fauna. **a**, Location of Devon Island in the circumpolar North (map data from International Bathymetric Chart of the Arctic Ocean, accessed via ArcGIS on 8 May 2024). **b**, Anterolingual view of specimen CMNFV59632 after destructive palaeoproteomic analysis. **c**, Location of Houghton Crater (75° 22' N, 89° 40' W) on Devon Island. Scale bar, 1 cm (**b**). Panel **a** adapted from ref. 48, Springer Nature Limited, under a Creative Commons Licence CC BY 4.0. Basemap from Natural Earth (<https://www.naturalearthdata.com/>).

and when the radiation of this group occurred²⁷. For most of the past two decades, the group was defined by a deep ‘basal split’ between two clades—Rhinoerotinae and Elasmotheriinae—before episodes of rhinocerotid diversification in the Late Eocene^{8,9,28,29}. This paradigm contrasts earlier hypotheses of a close relationship between two extinct rhinocerotids that survived into the Late Pleistocene—the Siberian unicorn (Elasmotheriinae, *Elasmotherium sibiricum*) and the woolly rhinoceros (Rhinoerotinae, *Coelodonta antiquitatis*)³⁰. Recently, the sequenced genomes of *Coelodonta* and *Elasmotherium*³¹ were used to confirm hypotheses on the basis of morphological data that suggest they have distinct phylogenetic affinities⁸, but also allowed for the recognition of a split between these two groups during the Late Eocene (36 Ma). This suggests that the deep-divergence hypothesis based on the morphological analysis of fossils is not supported by molecular evidence. However, the lack of available genetic sequence data from other early-diverging rhinocerotid lineages (for example, Aceratheriinae³²), makes it difficult to assess the timing of the Rhinoerotinae–Elasmotheriinae split in relation to other radiations that occurred in the group. For these reasons, the ancient radiations of the group remain obscured.

To investigate the timing of Rhinoerotidae divergence and the potential for evolutionarily informative protein sequences to persist in deep time, we targeted dental enamel deriving from the Houghton Crater (75° N, Nunavut) in Canada’s High Arctic (Fig. 1). The Houghton Crater is an impact structure with its stratigraphy including post-impact fossiliferous lacustrine sediments dated to 21–24 Ma (ref. 33). Fossils from these sediments are found in a polar landscape, at present characterized by permafrost. Compared with similarly aged material from lower latitudes, this creates a temperature regime favourable for biomolecular preservation, sparing these fossils from the harshest effects of diagenesis, and potentially paralleling those mechanisms underlying the remarkable soft-tissue preservation of Konservat-Lagerstätten³⁴. To maximize the likelihood of proteomic recovery, we focused on dental

enamel, following recent successful extractions of pre-Pliocene peptides from highly biomineralized tissues^{5,6}. The prismatic enamel of placental mammals, in which tightly packed enamel prisms typically extend from the dentine–enamel junction to the tooth surface, presents a suitable scaffold for protecting biomolecules³⁵.

The digestion-free palaeoproteomic workflow^{2,3} applied to an Early Miocene rhinocerotid (*Epiaceratherium* sp.) specimen (CMNFV59632) (Fig. 1b and Supplementary Information) of dental enamel¹³⁶ from the Houghton Formation (21.8 Ma) allowed for the recovery of an enamel proteome covering 1,163 confident peptide–spectrum matches (PSMs), at least seven proteins (AHSG, ALB, AMBN, AMELX, AMTN, ENAM and MMP20) and spanning at least 251 amino acids (Fig. 2a and Extended Data Fig. 1). At present, the enamel proteome of CMNFV59632 represents both the oldest mammalian skeletal proteome reported, confirming the predicted deep-time persistence of ancient mammalian proteins from high latitudes^{3,5}, and the first biomolecular characterization of the extinct genus *Epiaceratherium*. The survival of a relatively rich enamel proteome from such ancient deposits is representative of the specimen’s excellent state of preservation.

Protein diagenesis in closed systems such as enamel is driven by the combined effect of time and temperature. Therefore, thermal age³⁷ can be used to assess expected and observed molecular degradation, typically by normalizing the thermal age to a mean annual temperature of 10 °C, and to predict survival into deep time at different geographic locations^{3,17}. Rather than using global signals of climate change³⁸ to estimate the temperature history of the sample, we extrapolated location-specific palaeotemperature values from the HadCM3 model^{39,40} (Extended Data Fig. 2 and Supplementary Table 1). We calculate the equivalent thermal age at 10 °C (Ma at 10 °C) for CMNFV59632, obtaining a value of 2.5 ± 2.5 Ma at 10 °C. Despite the broad confidence interval, caused by the wide seasonal temperature fluctuations extrapolated from the model for the Houghton Crater location in the High Arctic (Extended Data Fig. 3), this result is entirely consistent with previously reported protein survival over Pleistocene–Pliocene timescales in temperate climates³. Consequently, it reinforces the idea that the climatic history of Miocene high-latitude sites is compatible with protein preservation, a crucial requirement for extending molecular-based phylogenetic reconstructions into deep time.

To better appreciate the preservation state of the Houghton Crater enamel proteome, we compared it with those of two other rhinocerotids, the Early Pleistocene *Stephanorhinus* from the site of Dmanisi (Georgia), dated at 1.77 Ma (ref. 2), and a Middle Pleistocene *Stephanorhinus* (about 0.4 Ma) from the site of Fontana Ranuccio (Italy). A medieval ovicaprid enamel control sample² was also re-analysed to illustrate preservation differences. Although the set of proteins retrieved from the CMNFV59632 enamel specimen is similar to that of the other two Pleistocene rhinocerotids used for comparison, fewer peptides and a shorter reconstructed amino acid sequence were recovered from the Arctic specimen (Fig. 2).

As expected, diagenetic modifications are extensive in the enamel proteome of CMNFV59632 (Fig. 2b). Average peptide lengths are similar, although slightly shorter than those of the Dmanisi Early Pleistocene specimen (9.64 amino acids versus 10.42 amino acids, respectively), and further reduced in comparison with the Fontana Ranuccio Middle Pleistocene *Stephanorhinus* (10.99 amino acids), indicating a greater degree of peptide bond hydrolysis (Fig. 2b). We also observe high deamidation rates in CMNFV59632, although no more so than in the Pleistocene rhinocerotids (Extended Data Fig. 3), or other previously sequenced mammalian specimens from low-latitude sites^{2–4}. Although high deamidation rates can be useful for confirming proteome authenticity, they can be highly variable in samples^{41,42}, and can plateau relatively quickly in fossil proteomes, reducing their utility in characterizing degradation patterns in deep time (Fig. 2c). Instead, we identify a suite of informative spontaneous modifications indicative of advanced diagenesis that are observed at a higher rate in

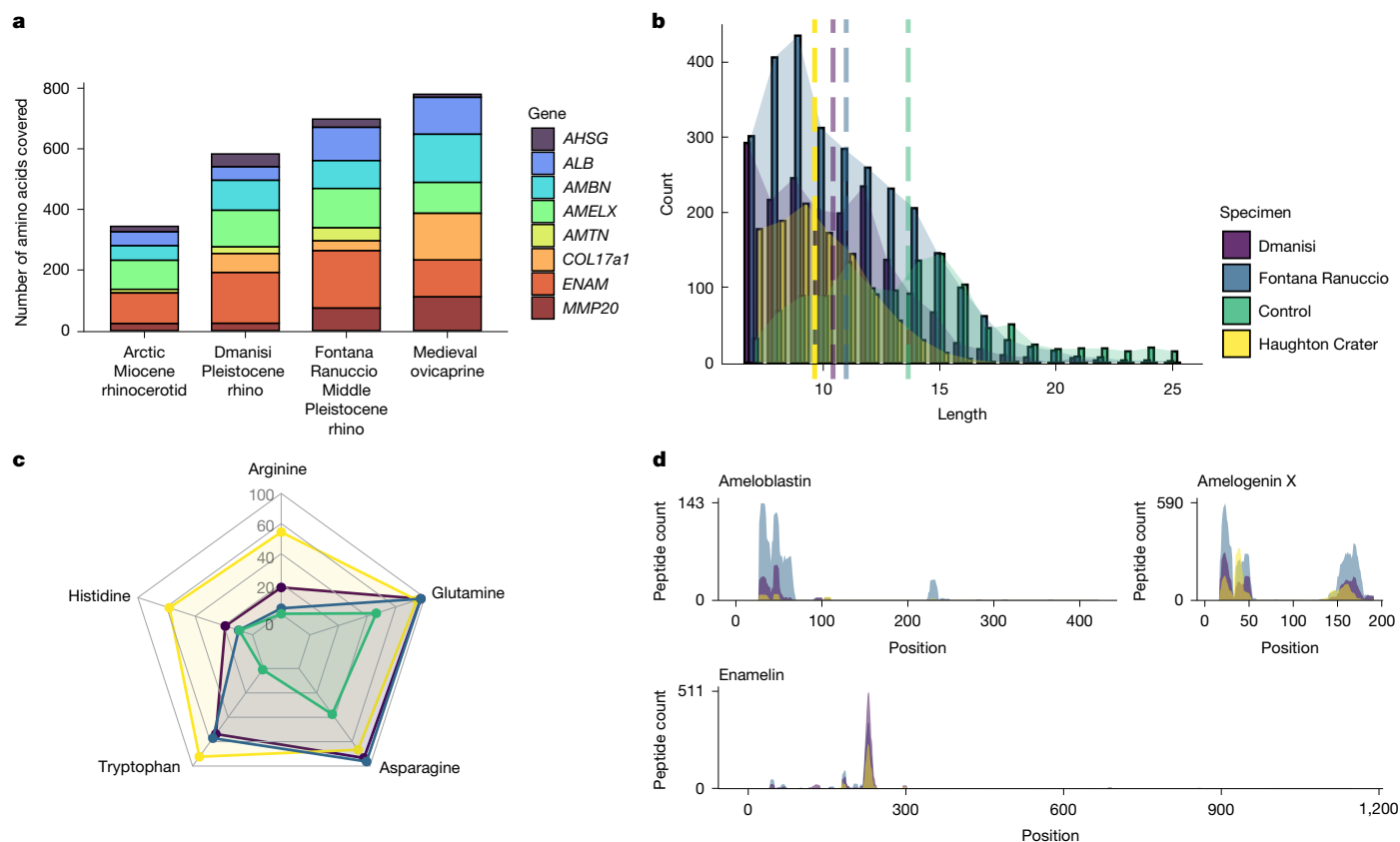


Fig. 2 | Proteome preservation in the enamel specimen of the Early Miocene rhinocerotid (CMNFV59632, Haughton Crater, Canada). Preservation is compared with enamel proteomes from an Early Pleistocene (1.77 Ma) *Stephanorhinus* (rhino) (DM.5/157, Dmanisi, Georgia), a Middle Pleistocene (0.4 Ma) *Stephanorhinus* (CGG1_023342, Fontana Ranuccio, Italy) and a medieval ovicaprine (Control, Aarhus, Denmark)². All plots exclude contaminants and reverse hits. **a**, Comparison of amino acid sequence coverage for each identified protein between samples, showing that coverage decreases over time. **b**, Comparison of the peptide length distributions for each sample. Dashed bars represent average peptide length for each specimen, showing that older samples have shorter average lengths. **c**, Comparison of the modification rate (0%–100%) of selected amino acids recovered from each sample that are often modified in ancient enamel proteomes. Colours are the same as in **b**. Results derive from modification-specific searches as described

in the Methods. ‘Arginine’ includes arginine to ornithine conversion (–42.02 Da); ‘Glutamine’ includes glutamine deamidation (+0.98 Da); ‘Asparagine’ includes asparagine deamidation (+0.98 Da); ‘Tryptophan’ includes advanced tryptophan oxidation to kynurenine (+3.99 Da), oxylactone (+13.98 Da) and tryptophandione (+29.97 Da); ‘Histidine’ includes oxidation (+15.99 Da) and dioxidation (+31.99 Da) of histidine, as well as histidine conversion to hydroxyglutamate (+7.98 Da). The overall average modification rate of these amino acids (excluding deamidation) ranges from 6.2% in the control to 72.3% in the Haughton Crater specimen, showing an increase in oxidative damage over time, especially for histidine and arginine. **d**, Sequence coverage plots for the three most abundant EMPs (AMBN, AMELX and ENAM), recording relative number of PSMs (coloured areas). Colours are the same as in **b**. Raw data used to create these figures are given in the Supplementary Information.

the Arctic Miocene rhinocerotid, providing support for their utility as markers of advanced diagenesis and authenticity in deep time² (Fig. 2c). These include arginine to ornithine conversion (Fig. 2c) and advanced forms of tryptophan (Extended Data Fig. 1b) and histidine oxidation (Extended Data Fig. 1c). Intra-crystalline protein decomposition analysis further confirms the advanced degradation state of CMNFV59632. The concentration of free amino acids (FAA) and total hydrolysable amino acids (THAA) is around half of those in the Early Pleistocene *Stephanorhinus* sample from Dmanisi (Extended Data Fig. 4a), and the percentage of FAA in CMNFV59632 (about 75%) is higher than in the Pleistocene *Stephanorhinus* from Dmanisi (about 50%) (Extended Data Fig. 5b), supporting increased peptide bond hydrolysis. Furthermore, these analyses confirm that the enamel of CMNFV59632 behaves as a closed system, because the racemization values for CMNFV59632 fall along the expected FAA versus THAA trends for both fossil enamel and experimentally heated enamel samples (300 °C for 10 min) (Extended Data Fig. 6). On a peptide level, endogeneity is supported by the similar patterns and levels of across-sequence degradation shown by sequence coverage plots for CMNFV59632 and the experimentally heated samples (Extended Data Fig. 7).

Peptide sequences recovered from CMNFV59632 also derive from sequence regions similar to those previously identified in the Dmanisi Pleistocene *Stephanorhinus* proteome (Fig. 2d), particularly for the three most abundant enamel matrix proteins (EMPs). ENAM and AMBN present broadly similar sequence coverage patterns in both specimens, although with fewer PSMs covering most positions in the Miocene sample. AMELX, the most abundant EMP, is instead covered by a similar number of PSMs in both the Miocene and Pleistocene samples. The depth of coverage is also similar for the most abundantly covered AMELX sequences, including those spanning the deletion observed in the leucine-rich amelogenin peptide².

Despite a relatively limited breadth of coverage across the sequenced proteins, a high depth of coverage allows for the confident reconstruction of sequences in CMNFV59632, including positions variable in Perissodactyla. At least ten single amino acid polymorphisms (SAPs) support the placement of CMNFV59632 in Rhinocerotidae. A smaller number (two or more) of SAPs are shared between CMNFV59632 and other perissodactyls, to the exclusion of later-diverging rhinocerotids. No new variants are uncovered in CMNFV59632, because the aforementioned SAPs represent character states retained from ancestors in

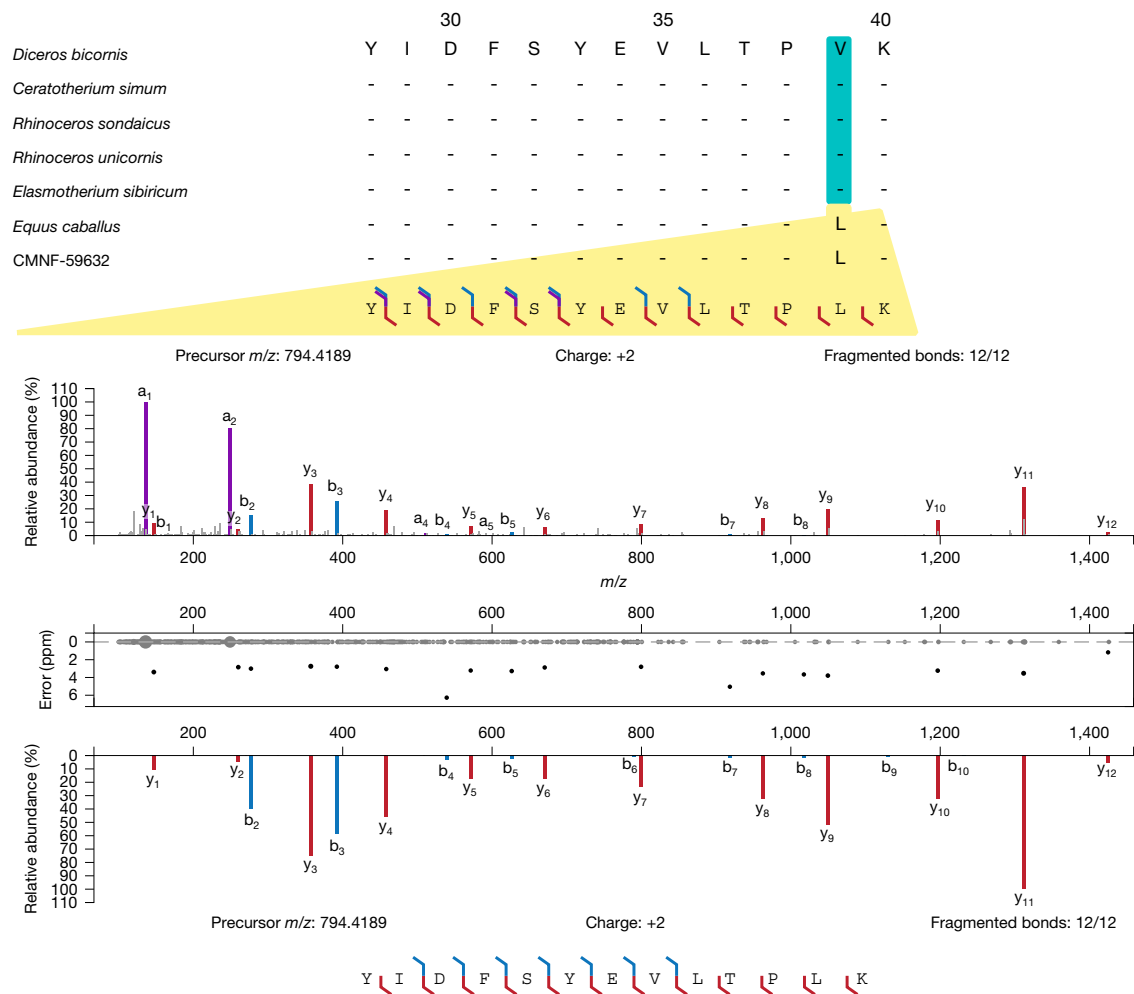


Fig. 3 | Abridged alignment and mirror plots of a phylogenetically informative SAP at AMELX-39. Residue numbering (depicted above the alignment) follows the UniProt reference sequence F6QHS4 (F6QHS4_HORSE), corresponding to AMELX isoform 1 of *Equus caballus*. The upper spectrum is experimentally derived, whereas the lower one is predicted using the ‘Original mode’ with the Prosit tool, available online via the Universal Spectrum

Explorer⁴⁹. This spectrum is the highest scoring PSM (with Andromeda) for AMELX sequence positions spanning the most abundantly covered SAP differentiating between CMNFV59632 and all other rhinocerotids for which sequences are available. Instead, the more ancestral variant (YIDFSYEVLTPK), shared with horses and others, is recovered.

Perissodactyla and Mammalia more broadly. The identification of these SAPs is supported by several unique PSMs showing almost complete ion series (Fig. 3 and Supplementary Section 4).

Regardless of the mechanisms behind preferential mass spectrometric and data analysis identification of specific sequence regions, biases favouring the recovery¹⁷ and identification⁴³ of conserved peptide sequences can ultimately lead to underestimates of divergence times in taxa represented by empirically derived protein sequences. To accurately estimate the phylogenetic position of CMNFV59632 and estimate divergence times in the group, we completed a phylogenetic analysis of a suite of extinct and extant perissodactyls. In addition to the perissodactyl taxa previously used², we incorporated whole-genome sequence data to predict enamel protein sequences from the Siberian unicorn (*Elasmotherium sibiricum*) and a pair of extant tapirs (*Tapirus terrestris* and *Tapirus indicus*).

The time-calibrated phylogenetic analysis of enamel protein sequences under a fossilized birth–death (FBD) model infers CMNFV59632 as the earliest diverging rhinocerotid in the analysis, with *Elasmotherium sibiricum* being more closely related to Rhinocerotina (crown rhinoceroses) than to CMNFV59632 (Fig. 4). This phylogenetic hypothesis has also been supported by previous total-evidence analysis³⁶. Also, our FBD analysis resolves the Early

Pleistocene *Stephanorhinus* from Dmanisi as a sampled ancestor of the Middle Pleistocene *Stephanorhinus* from Fontana Ranuccio. Divergence time estimates place the split between CMNFV59632 and all other rhinocerotids during the Middle Eocene–Oligocene (around 41–25 Ma). The divergence between Elasmotheriinae and Rhinocerotina is reconstructed to have probably occurred in the Oligocene (around 34–22 Ma), which is younger than previous molecular clock estimates³¹.

The Late Eocene and the Early Oligocene represent dynamic periods in the evolution of rhinocerotids, particularly in North America. After appearing in the Middle Eocene (37–34 Ma), North American rhinocerotids diversify during the Late Eocene, evolving a variety of body sizes and ecologies as several new clades arise, before rhinocerotid diversity experiences a significant drop in the Early Oligocene (34–32 Ma)⁴⁴. During this time frame, other early-diverging lineages are also appearing in Asia^{27,45}, eventually spreading as far as western Europe²⁷. Morphologically, the Houghton Crater rhinocerotid shares closer affinities with these early-diverging lineages from Eurasia¹⁰, particularly those in the genus *Epiaceratherium*³⁶. Similarly, some other vertebrates in the highly endemic fauna of the Houghton Formation have their closest relatives in Eurasia. These include the transitional pinniped *Puijila darwini*, sister to the Oligocene *Potamotherium* of Europe⁴¹, and a swan, family Anatidae, a group that is otherwise restricted to the Oligocene and

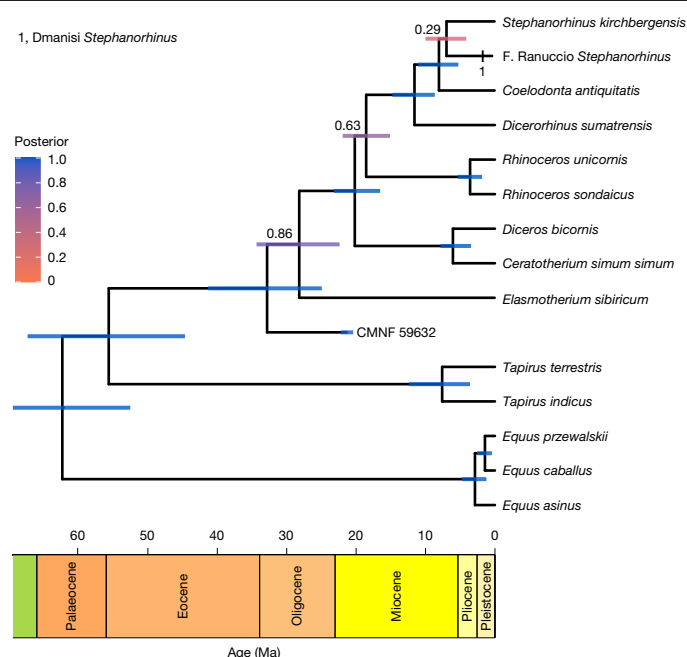


Fig. 4 | Time-calibrated phylogeny of Rhinocerotidae enamel proteomes. The maximum a posteriori (MAP) tree was produced using RevBayes v.1.2.1 (ref. 50) (<https://revbayes.github.io/>) with a FBD model. Coloured bars at nodes represent 95% height posterior density age interval estimates. Specimen CMNFV59632 represents the Early Miocene rhinocerotid from the Haughton Crater.

Miocene of Europe¹⁰. Overall, these patterns, in conjunction with the recovered divergence times, suggest the Haughton Crater rhinocerotid represents a migrant from eastern Asia or western Europe, derived from one of the early-diverging lineages that arose in the Late Eocene or early Oligocene of East Asia.

We provide molecular evidence that this lineage falls outside Rhinocerotinae, because it diverges before the Rhinocerotinae–Elasmotheriinae split. We also reject a deep divergence (basal split) between Elasmotheriinae and Rhinocerotinae^{8,9,29} and find moderate support for their branching event after the divergence of *Epiaceratherium*. Our analysis disagrees with that in ref. 9, which noted a deep divergence for Elasmotheriinae (47.3 Ma), and an early divergence for Rhinocerotinae (almost 30.8 Ma). The later divergence times for these nodes in our analysis are despite equivalently old ages for crown Ceratomorpha (earliest Eocene). Among other timetrees, our dates are generally most consistent with those reported in ref. 31. Our recovered topologies are also broadly similar to trees derived from previous morphology-based phylogenetic analyses^{27,32}, identifying Elasmotheriinae and Rhinocerotinae as deeply nested in Rhinocerotidae. Discrepancies between the genomic³¹ and proteomic trees arise probably because of different calibration points. The more ancient age of Elasmotheriinae in the analysis in ref. 31 is constrained by a high minimum bound for the Elasmotheriinae–Rhinocerotinae split (35 Ma). However, this date is based on the earliest age of *Epiaceratherium naduongense* and its allocation to Rhinocerotinae. Assuming monophyly of *Epiaceratherium*, the present proteomic evidence refutes the assignment of this genus to Rhinocerotinae, because it falls as earlier-diverging than Elasmotheriinae without such topological constraints in our phylogenetic analysis.

In sum, these findings highlight the importance of integrating palaeoproteomic sequence data into phylogenetic analyses to infer topologies and estimate divergence times. Ancient proteomic sequence data allow for robustly supported timetrees, and can serve to develop phylogenetic frameworks in deep time, particularly from specimens too old to preserve aDNA. For example, the present data allow for firm placement

of the Haughton Crater rhinocerotid outside Rhinocerotina, and probably outside the Elasmotheriinae–Rhinocerotinae clade, a fact that has significant implications for both morphological and molecular studies integrating fossil calibration times from the fossil record. In the future, fully characterizing these deep divergences in Rhinocerotidae requires accessing protein sequence data from Aceratheriinae, a group that includes the late-surviving Pliocene *Shansirhinus* from high-altitude deposits in the Linxia Basin⁴⁶, a region that has shown to be amenable to ancient protein survival into the Miocene⁵.

Our experimental results and thermal age calculations firmly indicate that at least some high-latitude fossiliferous deposits preserve not only the tangible remains of extinct organisms, but also ancient biomolecules. Further experiments on a broader sample of fossils from this site can reveal whether this exceptional preservation is an isolated case or extends across the Haughton Formation. The latter scenario would support the notion that the Haughton Formation could potentially represent a new type of lagerstätte³⁴—a palaeomolecular lagerstätte—showing preservation of subfossil peptides from a time range when they are otherwise not known. These findings should encourage further vertebrate palaeontological field work in the High Arctic, and other cold-temperature sites, with a goal that includes identifying taphonomic conditions favourable to such remarkable biomolecular preservation. The survival of an extended set of mammalian enamel peptides in the Early Miocene demonstrates that the research scopes of palaeoproteomics and palaeobiogeochemical analyses focusing on proteogens from the Palaeocene and beyond^{19,21} can finally overlap. The complementary integration^{25,47} of these two approaches can ultimately lead us to better define a unifying framework for understanding the degradation pathways from intact biological proteins and polypeptide chains to very short, unsequenceable, oligopeptides and isolated amino acids. More broadly, this work illustrates the power of palaeoproteomics in elucidating phylogeny and taxonomy of extinct vertebrates in deep time.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-025-09231-4>.

- Buckley, M., Lawless, C. & Rybczynski, N. Collagen sequence analysis of fossil camels, *Camelops* and c.f. *Paracamelus*, from the Arctic and sub-Arctic of Plio-Pleistocene North America. *J. Proteomics* **194**, 218–225 (2019).
- Cappellini, E. et al. Early Pleistocene enamel proteome from Dmanisi resolves *Stephanorhinus* phylogeny. *Nature* **574**, 103–107 (2019).
- Welker, F. et al. Enamel proteome shows that *Gigantopithecus* was an early diverging pongine. *Nature* **576**, 262–265 (2019).
- Welker, F. et al. The dental proteome of *Homo antecessor*. *Nature* **580**, 235–238 (2020).
- Demarchi, B. et al. Survival of mineral-bound peptides into the Miocene. *eLife* **11**, e82849 (2022).
- Stolarski, J. et al. First paleoproteome study of fossil fish otoliths and the pristine preservation of the biomineral crystal host. *Sci. Rep.* **13**, 3822 (2023).
- Prothero, D. R. in *Evolution of Tertiary Mammals of North America* (ed. Janis, C. M. et al.) 595 (Cambridge Univ. Press, 1998).
- Antoine, P. O. *Phylogénie et évolution des Elasmotheriina (Mammalia, Rhinocerotidae)* (Muséum national d'Histoire naturelle, 2002).
- Kosintsev, P. et al. Evolution and extinction of the giant rhinoceros *Elasmotherium sibiricum* sheds light on late Quaternary megafaunal extinctions. *Nat. Ecol. Evol.* **3**, 31–38 (2019).
- Whitlock, C. & Dawson, M. R. Pollen and vertebrates of the early Neogene Haughton Formation, Devon Island, Arctic Canada. *Arctic* **43**, 324–330 (1990).
- Rybczynski, N., Dawson, M. R. & Tedford, R. H. A semi-aquatic Arctic mammalian carnivore from the Miocene epoch and origin of Pinnipedia. *Nature* **458**, 1021–1024 (2009).
- Kjær, K. H. et al. A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature* **612**, 283–291 (2022).
- Orlando, L. et al. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78 (2013).
- Dabney, J. et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110**, 15758–15763 (2013).

15. Barlow, A. et al. Middle Pleistocene genome calibrates a revised evolutionary history of extinct cave bears. *Curr. Biol.* **31**, 1771–1779.e7 (2021).
16. van der Valk, T. et al. Million-year-old DNA sheds light on the genomic history of mammoths. *Nature* **591**, 265–269 (2021).
17. Demarchi, B. et al. Protein sequences bound to mineral surfaces persist into deep time. *eLife* **5**, e17092 (2016).
18. Buckley, M., Warwood, S., van Dongen, B., Kitchener, A. C. & Manning, P. L. A fossil protein chimera; difficulties in discriminating dinosaur peptide sequences from modern cross-contamination. *Proc. Biol. Sci.* **284**, 20170544 (2017).
19. Umamaheswaran, R. & Dutta, S. Preservation of proteins in the geosphere. *Nat. Ecol. Evol.* **8**, 858–865 (2024).
20. Rybczynski, N. et al. Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution. *Nat. Commun.* **4**, 1550 (2013).
21. Wiemann, J. et al. Fossilization transforms vertebrate hard tissue proteins into N-heterocyclic polymers. *Nat. Commun.* **9**, 4741 (2018).
22. Dutta, S. et al. Chemical evidence of preserved collagen in 54-million-year-old fish vertebrae. *Palaeontology* **63**, 195–202 (2020).
23. Boskovic, D. S. et al. Structural and protein preservation in fossil whale bones from the Pisco Formation (middle-upper Miocene), Peru. *Palaio* **36**, 155–164 (2021).
24. Umamaheswaran, R. et al. *Proteins in the Geological Record Under a Diagenetic Framework: Insights From PY-GC × GC-Tofms* (European Association of Geoscientists & Engineers, 2023).
25. Saitta, E. T., Vinther, J., Crisp, M. K. & Abbott, G. D. Non-avian dinosaur eggshell calcite can contain ancient, endogenous amino acids. *Geochim. Cosmochim. Acta* **365**, 1–20 (2024).
26. Pandolfi, L. Evolutionary history of Rhinocerotina (Mammalia, Perissodactyla). *Fossilia* **2018**, 27–32 (2018).
27. Tissier, J., Antoine, P.-O. & Becker, D. New material of *Epiacetherium* and a new species of *Mesacetherium* clear up the phylogeny of early Rhinocerotidae (Perissodactyla). *R. Soc. Open Sci.* **7**, 200633 (2020).
28. Antoine, P.-O. et al. A revision of *Aceratherium blanfordi* Lydekker, 1884 (Mammalia: Rhinocerotidae) from the Early Miocene of Pakistan: postcranials as a key. *Zool. J. Linn. Soc.* **160**, 139–194 (2010).
29. Becker, D., Antoine, P.-O. & Maridet, O. A new genus of Rhinocerotidae (Mammalia, Perissodactyla) from the Oligocene of Europe. *J. Syst. Palaeontol.* **11**, 947–972 (2013).
30. Cerdeño, E. *Cladistic Analysis of the Family Rhinocerotidae (Perissodactyla)* (American Museum of Natural History, 1995).
31. Liu, S. et al. Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. *Cell* **184**, 4874–4885.e16 (2021).
32. Lu, X.-K., Deng, T. & Pandolfi, L. Reconstructing the phylogeny of the hornless rhinoceros *Aceratheriinae*. *Front. Ecol. Evol.* <https://doi.org/10.3389/fevo.2023.1005126> (2023).
33. Jessberger, E. K. ^{40}Ar – ^{39}Ar dating of the Haughton impact structure. *Meteoritics* **23**, 233–234 (1988).
34. Kimmig, J. & Schiffbauer, J. D. A modern definition of Fossil-Lagerstätten. *Trends Ecol. Evol.* **39**, 621–624 (2024).
35. Taurozzi, A. J. et al. Deep-time phylogenetic inference by paleoproteomic analysis of dental enamel. *Nat. Protoc.* **19**, 2085–2116 (2024).
36. Fraser, D., Rybczynski, N., Gilbert, M. & Dawson, M. R. Post-Eocene Rhinocerotid dispersal via the North Atlantic. Preprint at *BioRxiv* <https://doi.org/10.1101/2024.06.04.597351> (2024).
37. Wehmiller, J. F. Amino acid studies of the Del Mar, California, midden site: apparent rate constants, ground temperature models, and chronological implications. *Earth Planet. Sci. Lett.* **37**, 184–196 (1977).
38. Hansen, J., Sato, M., Russell, G. & Kharecha, P. Climate sensitivity, sea level and atmospheric carbon dioxide. *Philos. Trans. A Math. Phys. Eng. Sci.* **371**, 20120294 (2013).
39. Valdes, P. J., Scotese, C. R. & Lunt, D. J. Deep ocean temperatures through time. *Clim. Past* **17**, 1483–1506 (2021).
40. Leonardi, M., Hallett, E. Y., Beyer, R., Krapp, M. & Manica, A. *pastclim*: an R package to easily access and use paleoclimatic reconstructions. *Ecography* **2023**, e06481 (2023).
41. Simpson, J. P. et al. The effects of demineralisation and sampling point variability on the measurement of glutamine deamidation in type I collagen extracted from bone. *J. Archaeol. Sci.* **69**, 29–38 (2016).
42. Pal Chowdhury, M. & Buckley, M. Trends in deamidation across archaeological bones, ceramics and dental calculus. *Methods* **200**, 67–79 (2022).
43. Welker, F. Elucidation of cross-species proteomic effects in human and hominin bone proteome identification through a bioinformatics experiment. *BMC Evol. Biol.* **18**, 23 (2018).
44. Prothero, D. R. *The Evolution of North American Rhinoceroses* (Cambridge Univ. Press, 2005).
45. Antoine, P.-O. et al. Early rhinocerotids (Mammalia: Perissodactyla) from South Asia and a review of the Holarctic Paleogene rhinocerotid record. *Can. J. Earth Sci.* **40**, 365–374 (2003).
46. Deng, T. New cranial material of *Shansirhinus* (Rhinocerotidae, Perissodactyla) from the Lower Pliocene of the Linxia Basin in Gansu, China. *Geobios* **38**, 301–313 (2005).
47. Schweitzer, M. H., Schroeter, E. R. & Goshe, M. B. Protein molecular data from ancient (>1 million years old) fossil material: pitfalls, possibilities and grand challenges. *Anal. Chem.* **86**, 6731–6740 (2014).
48. Jakobsson, M. et al. The International Bathymetric Chart of the Arctic Ocean Version 5.0. *Sci. Data* **11**, 1420 (2024).
49. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
50. Höhna, S. et al. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65**, 726–736 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons

Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

Methods

Site and specimen

Located in the Houghton impact crater (75° N, Nunavut, Canada), the Houghton Formation comprises the remnants of a large, post-impact lacustrine deposit, dated to the Early Miocene. Previous dating estimates, using fission-track and ^{40}Ar – ^{39}Ar furnace step-heating dating, identified an age of 24–21 Ma (refs. 33,51). An Early Miocene age has also been corroborated by (U-Th)/He thermochronology⁵². Although older age estimates between 30 and 40 Ma have also been suggested^{53–55}, there have been no age estimates younger than the Early Miocene. Therefore, we conservatively use the younger Early Miocene age estimates in our analysis and interpretation.

The highly endemic fauna of the Houghton Formation consists of several vertebrate taxa, including a transitional pinniped¹¹, a pair of salmoniform fishes, a swan-like anatid, a small artiodactyl, a leporid rabbit, a heterosocid shrew and a well-preserved rhinocerotid^{10,36}. Although the megafloral assemblage is not particularly rich, the palynofloral assemblage is well-characterized, allowing for reconstruction of local climatic conditions. In the Early Miocene, the Houghton Crater lake and its surrounding environs experienced a significantly warmer annual temperature (8–12 °C) than the present day^{10,56}.

Specimen CMNFV59632 is a nearly complete rhinocerotid skeleton, including skull and dentition, uncovered 10.8 m above the base of the formation³⁶. Our analysis focuses on a single tooth fragment from a lower left m1 (Fig. 1b and Supplementary Information) that was already separated from the rest of its tooth row because of the fragmenting effectings of cryoturbation³⁶. The dental specimen's rhinocerotid affinities are further supported by its size and morphology (Supplementary Information), most notably the presence of vertical Hunter–Schreger bands on its enamel, a defining feature of rhinocerotids and found in few other mammals⁵⁷. A single tusk fragment (left i2) derived from CMNFV59632 was also selected for proteomic extraction. Owing to its thin enamel, only limited peptides were recovered from this tusk fragment, and the sample is thus excluded from further analysis and discussion.

Proteomic extraction

The laboratory workflow for the CMNFV59632 teeth and the Fontana Ranuccio *Stephanorhinus* tooth (for comparison) generally follows that reported in refs. 2,35. Using a sterilized drill, flakes of enamel were removed from the fragmentary teeth, with care taken to avoid sampling the dentine. The CMNFV59632 tooth enamel sample, weighing 154 mg, was then ground to a fine powder and demineralized overnight using 10% high-performance liquid chromatography (HPLC)-grade trifluoroacetic acid (TFA) (Merck, Sigma-Aldrich) in high-purity liquid chromatography–mass spectrometry (LC–MS) grade water. The CMNFV59632 tusk enamel sample, weighing 90 mg, was processed in the same way. The Fontana Ranuccio (FR sd-295) enamel sample was divided into three subsamples—FR2, FR3 and FR4—weighing 202, 243 and 205 mg, respectively, which were similarly ground to a fine powder, and demineralized using 10% TFA (FR3, FR4) or 10% HCl (FR2) in high-purity LC–MS grade water. For each sample, the demineralization step was repeated a second time to ensure complete demineralization. No enzymatic digestion was performed. Subsequently, peptides were collected and desalted on C18 StageTips⁵⁸ produced in-house. An extraction blank for each sample set was processed alongside the samples for every step, to control for contamination.

Mass spectrometry

StageTips were eluted with 30 μl of 40% acetonitrile (ACN) and 0.1% formic acid in high-purity LC–MS grade water, into a 96-well plate. To remove ACN and concentrate the samples, the plate was vacuum-centrifuged until approximately 3 μl of sample remained in each well. Next, samples were resuspended in 6 μl of 5% ACN, 0.1%

formic acid in high-purity LC–MS grade water, and 4 μl (CMNFV59632) or 5 μl (FR sd-295 *Stephanorhinus*), of sample were injected.

Liquid chromatography coupled with tandem mass spectrometry was used to analyse the samples, on the basis of previously published protocols^{2,59}. Samples were separated on a 15-cm column (75 μm inner diameter in-house laser pulled and packed with 1.9- μm C18 beads (Dr Maisch)) on an EASY-nLC1200 (Proxeon) connected to an Exploris 480 (CMNFV59632) or a Q-Exactive HF-X (Fontana Ranuccio *Stephanorhinus*) mass spectrometer (both Thermo Fisher Scientific), with an integrated column oven. Buffer A, containing 0.1% formic acid in MilliQ water, and the peptides were separated with increasing buffer B (80% ACN, 0.1% formic acid in MilliQ water) with a 77-min gradient, increasing buffer B concentration from 5% to 30% in 50 min, 30% to 45% in 10 min, 45% to 80% in 2 min, and maintained at 80% for 5 min before decreasing to 5% in 5 min, and finally held for 5 min at 5%. Flow rate was 250 nl min^{-1} . An integrated column oven was used to maintain the temperature at 40 °C.

The two mass spectrometers were run using the same parameters except where specified, owing to changes in running software. Spray voltage was set to 2 kV, the S-lens RF (radio frequency) level was set to 40%, and the heated capillary was set to 275 °C. Full-scan mass spectra (MS1) were recorded at a resolution of 120,000 at m/z 200 over the m/z range 350–1,400. The AGC (automatic gain control) target value was set to 300% (Exploris) or 3×10^6 (HF-X) with a maximum injection time of 25 ms. HCD (higher-energy collisional dissociation)-generated product ions (MS2) were recorded in data-dependent top-10 mode and recorded at a resolution of 60,000. The maximum ion injection time was 118 ms (Exploris) or 108 ms (HF-X), with an AGC target value of 200% (Exploris) or 2×10^5 (HF-X). Normalized collision energy was set at 30% (Exploris) or 28% (HF-X). The isolation window was set to 1.2 m/z with a dynamic exclusion of 20 s. A wash-blank, using 5% ACN, 0.5% TFA, was run between each sample and laboratory blank to limit cross-contamination.

Database construction

The protein reference alignment given in ref. 2 was used as a starting point to construct a database for sequence reconstruction. Owing to the vast evolutionary distance between CMNFV59632 and any extant taxa (>20 Myr), a broader database was constructed to identify sequence variants that may be known in other mammals. To construct a broader database, we searched UniProt and the National Center for Biotechnology Information for each enamel protein, specifying the taxonomic grouping of 'Theria' to include all therian mammals. To supplement available sequences, others were manually extracted from available genomes, following the methodology reported previously⁶⁰.

To investigate the relationships at the base of Rhinocerotidae, protein sequences translated from *Elasmotherium sibiricum* genomic data³¹ were generated. To obtain the corresponding amino acid sequences, we first collapsed the paired-end reads and masked the conflict bases as 'N' using adapterRemoval⁶¹. We then mapped the collapsed reads against the reference genome of the white rhinoceros (GCF_000283155.1_CerSimSim1) using the BWA MEM function⁶² with the shorter split hits being abandoned. After that, we removed duplicates using an in-house Perl script following ref. 31. Finally, we extracted the gene sequences according to their locations on the reference genome.

The remaining steps generally follow the workflow outlined in ref. 2. We used ANGSD⁶³ to generate consensus sequences from BAM files corresponding to chromosomes that include genes of interest. To reduce the effects of post mortem aDNA damage, we trimmed the first and last five nucleotides from each DNA fragment. We formatted each consensus sequence as a blast nucleotide database. To recover translated protein sequences, we performed a tblastn alignment⁶⁴, with the corresponding *Ceratotherium simum* sequences as queries. Finally, we used ProSpLign to recover the spliced alignments, and ultimately, the translated protein sequences⁶⁵.

Protein identification

Thermo Fisher Scientific .raw files generated using the mass spectrometers were searched with various software using an iterative search strategy to interpret spectra, characterize modifications and ultimately, reconstruct protein sequences. For comparison, .raw files from a medieval ovicaprine (control) and an Early Pleistocene *Stephanorhinus* generated previously² were also analysed. Among samples from the Fontana Ranuccio *Stephanorhinus*, only FR4 was analysed. Although it is possible that some inter-sample variation between CMNFV59632 and these samples is caused by analysis using different mass spectrometer models, according to benchmarking studies^{66,67}, it is unlikely that the overall trends and interpretations would change.

We primarily used MaxQuant⁶⁸ for sequence reconstruction and other downstream aspects of data analysis. We performed two initial runs: (1) a more focused run using the database we modified from ref. 2, and (2) a broad run using the 'Theria'-wide database we constructed from publicly available sequences.

In all runs, an Andromeda score threshold of 40 and a delta score of 0 were set for both unmodified and modified peptides. Minimum and maximum peptide lengths were specified as 7 and 25, respectively. The default peptide false discovery rate was used (0.01), whereas the protein false discovery rate was increased to 1 to show possible low-abundance proteins. Error tolerances were kept at the default settings for Orbitrap MS instruments: 20 ppm for the first search, 4.5 ppm for the final search and 20 ppm for the fragment ion. 'Unspecific' digestion was specified. No fixed post-translational modifications were set. Several modifications were set as variable modifications in our initial runs: glutamine and asparagine deamidation (delta mass (ΔM) = +0.984016), methionine and proline oxidation (ΔM = +15.9949), N-terminal pyroglutamic acid from glutamine (ΔM = -17.026549) and glutamic acid (ΔM = -18.010565), phosphorylation of serine, threonine and tyrosine (ΔM = +79.966331), and the conversion of arginine to ornithine (ΔM = -42.021798).

Proteins included in the database of common contaminants provided by MaxQuant (for example, proteinaceous laboratory reagents and human skin keratins), as well as reverse sequences, were removed manually and not examined further. In addition, proteins detected in the laboratory blank were also treated as contaminants, and not considered further.

To discover new SAPs and peptide variants not included in our database, we used more search tools. Peaks v.7.0 was used to attempt de novo sequencing and an homology search was performed using the SPIDER algorithm^{69–71}. The open search capabilities of openPFind⁷² and MSFragger⁷³ were also used. When possible, the same settings were selected as in the MaxQuant runs.

With our iterative search strategy, we integrated possible sequence variants from the results of our de novo, homology searches and open searches into hypothetical sequences from closely related taxa, to produce artificial sequences. These artificial sequences were included in a subsequent MaxQuant search, and only incorporated into reconstructed sequences if identified and validated using MaxQuant.

Sequence reconstruction and filtering

Before sequence reconstruction, all non-redundant PSMs were filtered using three criteria to reconstruct only those peptide sequences and amino acid residues that we can confidently assign. Sequences were accepted at two levels, resulting in two different datasets: (1) a minimally filtered dataset, and (2) a strictly filtered dataset. This filtering starts with using Basic Local Alignment Search Tool (BLAST)⁷⁴ to determine whether peptides match any contaminants, beyond those included in MaxQuant by default, such as soil bacteria and fungi. At this stage, PSMs are discarded if they present a match to any reasonable candidate contaminants, if they are also identified in the blank, or if

they present poorly covered ion series. The resulting PSMs are used to reconstruct sequences for the 'minimally filtered dataset'.

Next, ion series coverage is examined for each PSM. At this stage, peptide sequences are accepted for the strictly filtered dataset only if each amino acid residue is covered (for example, at least y-, b- or a- ion designates the mass of that specific amino acid, plus any identified modifications) by at least two spectra, following the approach outlined previously⁷⁵. Also, for both strictly and minimally filtered datasets, poorly supported spectra are removed at this stage, and proteins are only submitted for phylogenetic analysis if they are covered by at least two non-overlapping peptides. Finally, under the strict filtering criteria, BLAST is used again on any trimmed sequences, to remove any that match contaminants.

Intra-crystalline protein decomposition analysis

We analysed chiral amino acids on CMNFV59632 to evaluate the overall extent of amino acid degradation in the intra-crystalline fraction of the enamel, enabling comparison with previously analysed specimens², and samples that had been heated experimentally to between 60 and 80 °C for up to 17,520 h and with samples heated to 200–500 °C for up to 25 min. Enamel chips were drilled using a Dremel 4000 (kit 4000-1/45) drill with a diamond wheel point (4.4 mm (7105) by Dremel) to remove any dentine, which could be identified under a microscope (ZEISS Stemi 305, Axiocam 105 R2). Samples were processed following the methods in ref. 76. To remove excess powders, enamel chips were washed in deionized water and ethanol (analytical grade) before being powdered in an agate pestle and mortar. Powdered samples were weighed into a single plastic microcentrifuge tube and bleached (NaOCl, 12%, 50 $\mu\text{L mg}^{-1}$ of enamel) for 72 h to remove the inter-crystalline amino acids and any contamination. This bleached sample was washed five times with deionized water and then once with methanol (HPLC grade), before being left to dry overnight.

The dried bleached sample was then divided into four subsamples: two for technical replicate analysis of the FAA and two for replicate analysis of the THAA. The THAA subsamples were dissolved in HCl (7 M, 20 $\mu\text{L mg}^{-1}$, analytical grade) in a sterile 2-ml glass vial (Wheaton), purged with N₂ to reduce oxidation and heated at 110 °C for 24 h in an oven (BINDER series). The acid was then removed by centrifugal evaporation (Christ RVC2-25). THAA and FAA fractions were subjected to a biphasic separation procedure^{76,77} to remove inorganic phosphate from the enamel samples. HCl was added to both FAA (1 M, 25 $\mu\text{L mg}^{-1}$) and THAA (1 M, 20 $\mu\text{L mg}^{-1}$) fractions in separate 0.5-ml plastic microcentrifuge tubes (Eppendorf), and KOH (1 M, 28 $\mu\text{L mg}^{-1}$) was added into the acidified solutions, which then formed monophasic cloudy suspensions. Samples were agitated and then centrifuged (13,000 rpm for 10 min, Progen Scientific GenFuge 24D) to form a clear supernatant above a gel. The supernatant was removed and dried by vacuum centrifugation. The concentration of the intra-crystalline amino acids and their extent of racemization (D/L value) were then quantified using reverse-phase HPLC (Agilent 1100 series HPLC fitted with HyperSil C18 base deactivated silica column (5 μm , 250 $\times$ 3 mm) and fluorescence detector) following a modified method from ref. 78.

For the reverse-phase HPLC analysis, samples were rehydrated with an internal standard solution (L-homo-arginine (0.01 mM), sodium azide (1.5 mM) and HCl (0.01 M)), and run alongside standards and blanks. A tertiary mobile phase system (HPLC grade ACN–methanol–sodium buffer; 21 mM sodium acetate trihydrate, sodium azide, 1.3 μM EDTA, pH adjusted to 6.00 $\pm$ 0.01 with 10% acetic acid and sodium hydroxide) was used for analysis. The D and L peaks of the following amino acids were separated: aspartic acid and asparagine; glutamic acid and glutamine; serine, alanine, valine, phenylalanine, isoleucine, leucine, threonine, arginine, tyrosine and glycine. During preparation, asparagine and glutamine undergo rapid irreversible deamidation to aspartic acid and glutamic acid respectively⁷⁹ and hence they are reported together as aspartic acid and asparagine, and glutamic acid and glutamine.

One of the experimentally heated samples (300 °C for 10 min) was also analysed using liquid chromatography coupled with tandem mass spectrometry with minor changes to the protocol reported in ref. 2 (Extended Data Fig. 7 and Supplementary Information).

Phylogenetic analysis

A time-calibrated phylogenetic tree was inferred with the Bayesian phylogenetic software RevBayes v.1.2.1 (ref. 50) (<https://revbayes.github.io/>) under a constant-rate FBD model^{80,81}. The dataset consisted of enamel proteome data for 16 perissodactyl species (10 extant and 6 extinct), totalling 7 proteins and 3,446 amino acids. Phylogenetic analyses were performed with both the strictly filtered and minimally filtered sequences for CMNFV59632, to observe any topological differences between the two datasets and assess whether filtering is warranted. Because no main differences were observed, only the results from the 'strictly filtered' dataset are discussed. The proteome dataset was partitioned by protein. A General Time Reversible + Invariant sites (GTR + I) amino acid substitution model—in which stationary frequencies of the 20 amino acids and exchangeability rates among amino acids are free to vary and estimated from data—was applied to each partition. Preliminary unrooted phylogenetic analyses performed on each protein showed evidence for within-protein Γ -distributed rate variation only for MMP20, hence Γ -distributed rate variation was modelled only for the MMP20 partition. A relaxed clock model with uncorrelated lognormal-distributed rates was applied to allow rate variation across branches. The prior on the average clock rate was set as a log uniform distribution (min = 10^{-8} , max = 10^{-2} substitutions per lineage per million years). The prior on the clock rate standard deviation was set as an exponential distribution with mean equal to 0.587405, corresponding to one order of magnitude of clock rate variation among branches. The FBD tree model allows for placement of extinct species in a phylogenetic tree while simultaneously estimating the rates of speciation, extinction and fossilization (sampling of species in the past). The priors on speciation, extinction and fossilization parameters were set as uniform distributions bounded between 0 and 10. The sampling probability for extant species was fixed to 0.5882353 ($\frac{10}{17}$), corresponding to the fraction of extant perissodactyl species included in the analysis, and assuming uniform sampling of extant taxa. The three species of Equidae in the analysis (*Equus caballus*, *Equus przewalskii* and *Equus asinus*) were constrained as outgroup to other perissodactyls (Tapiridae and Rhinocerotidae). Tip ages of fossil taxa were given a uniform prior distribution ranging from the minimum to maximum age of the deposit in which each fossil has been found. The prior on the origin age of the tree was set as a uniform distribution with minimum = 54 Ma, corresponding to the oldest fossil that can be unequivocally assigned to crown Perissodactyla (*Cambaylophus vastanensis* from the Early Ypresian Cambay Shale⁸²), and maximum = 100 Ma, corresponding to the beginning of the Late Cretaceous and a very lax upper boundary on the origin of placental mammals⁸³. Further constraints on node ages on the basis of the fossil record of perissodactyls were set to improve the precision of divergence age estimates. Each node calibration was set up as a soft-bounded uniform distribution with normally distributed tails, with 2.5% of the distribution younger than the minimum age (allowing for potential misattribution of the oldest fossil of a clade) and 2.5% of the distribution older than the maximum age. Monophyly was not enforced when setting up these node calibrations. The following age constraints have been applied to five nodes: (1) Node = crown Perissodactyla; soft minimum = 54 Ma, with the same justification as the minimum on the origin age prior; soft maximum = 66 Ma, corresponding to the Cretaceous/Palaeogene boundary, before which no unambiguous crown placental fossils are known. (2) Node = Rhinocerotina (crown rhinoceroses); soft minimum = 22.6 Ma, corresponding to the earliest putative appearance of a crown rhinoceros in the fossil record (*Gaindatherium cf. browni* from the Aquitanian upper member of the Chitarwata Formation^{84,85}); soft

maximum = 44 Ma, corresponding to the minimum age of Rhinocerotidae as supported by fossil and phylogenetic evidence³¹. (3) Node = Diceroti (*Ceratotherium* + *Diceros*); soft minimum = 5.3 Ma, corresponding to the minimum age of the oldest deposits yielding *Diceros bicornis* fossils (Lothagam and Albertine^{86,87}); soft maximum = 7.3 Ma, as in ref. 31. (4) Node = *Rhinoceros unicornis* + *Rhinoceros sondaicus*; soft minimum = 1.9 Ma, corresponding to the Early Pleistocene appearance of *Rhinoceros unicornis* in the fossil record^{88,89}; soft maximum = 5.3 Ma, as in ref. 31. (5) Node = *Dicerorhinus* + *Stephanorhinus* + *Coelodonta*; soft minimum = 13 Ma, corresponding to Middle Miocene remains of *Dicerorhinus* from the Middle Siwaliks of Pakistan^{31,90}; soft maximum = 22.6 Ma, corresponding to the oldest crown rhinoceros fossil as in the soft minimum of calibration 2.

The Markov chain Monte Carlo was set up as four independent runs, running for 50,000 iterations and sampling every 10, averaging between 262.2 and 279.2 moves per iterations. Convergence between runs was checked by visually inspecting and calculating effective sample sizes of parameter estimates on Tracer v.1.7.2 (ref. 91). A MAP tree was calculated to summarize the posterior distribution of trees, with 20% burn-in. In the analysis of the minimally filtered dataset, one of the four runs was discarded from the MAP tree calculation, as it converged only in the last 10% of the Markov chain Monte Carlo.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The proteomics datasets have been deposited to the ProteomeXchange Consortium via the Proteomics Identifications Database (PRIDE) partner repository with the dataset identifier PXD052635. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission. Source data are provided with this paper.

Code availability

In-house scripts for the phylogenetic analyses are available at Figshare (<https://doi.org/10.6084/m9.figshare.25991269>).

- Omar, G. et al. Fission-track dating of Haughton Astrobleme and included biota, Devon Island, Canada. *Science* **237**, 1603–1605 (1987).
- Young, K. E. et al. Impact thermochronology and the age of Haughton impact structure, Canada. *Geophys. Res. Lett.* **40**, 3836–3840 (2013).
- Stephan, T. & Jessberger, E. K. Isotope systematics and shock-wave metamorphism: III. K–Ar in experimentally and naturally shocked rocks: the Haughton impact structure, Canada. *Geochim. Cosmochim. Acta* **56**, 1591–1605 (1992).
- Sherlock, S. C. et al. Re-evaluating the age of the Haughton impact event. *Meteorit. Planet. Sci.* **40**, 1777–1787 (2005).
- Erickson, T. M. et al. Resolving the age of the Haughton impact structure using coupled ⁴⁰Ar/³⁹Ar and U–Pb geochronology. *Geochim. Cosmochim. Acta* **304**, 68–82 (2021).
- Hickey, L. J., Johnson, K. R. & Dawson, M. R. The stratigraphy, sedimentology, and fossils of the Haughton formation: a post-impact crater-fill, Devon Island, N.W.T., Canada. *Meteoritics* **23**, 221–231 (1988).
- Von Koenigswald, W., Holbrook, L. T. & Rose, K. D. Diversity and evolution of Hunter Schreger band configuration in tooth enamel of perissodactyl mammals. *Acta Palaeontol. Pol.* **56**, 11–32 (2011).
- Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
- Mackie, M. et al. Palaeoproteomic profiling of conservation layers on a 14th century Italian wall painting. *Agnew. Chem. Int. Ed. Engl.* **57**, 7369–7374 (2018).
- Rüther, P. L. et al. SPIN enables high throughput species identification of archaeological bone by proteomics. *Nat. Commun.* **13**, 2458 (2022).
- Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **5**, 337 (2012).
- Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
- Korneliusson, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
- Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

65. Kuznetsov, A. & Bollin, C. J. NCBI Genome Workbench: desktop software for comparative genomics, visualization, and GenBank data submission. *Methods Mol. Biol.* **2231**, 261–295 (2021).
66. Bekker-Jensen, D. B. et al. A compact quadrupole-orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients. *Mol. Cell. Proteomics* **19**, 716–729 (2020).
67. Kelstrup, C. D. et al. Performance evaluation of the Q Exactive HF-X for shotgun proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
68. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
69. Han, Y., Ma, B. & Zhang, K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. In *Proc. 2004 IEEE Computational Systems Bioinformatics Conf. (CSB 2004)* (eds. Dunker, A. K. et al.) 206–215 (Institute of Electrical and Electronics Engineers, 2004).
70. Zhang, J. et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **11**, M111.010587 (2012).
71. Ma, B. & Johnson, R. De novo sequencing and homology searching. *Mol. Cell. Proteomics* **11**, O111.014902 (2012).
72. Chi, H. et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059–1061 (2018).
73. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
74. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
75. Coutu, A. N. et al. Palaeoproteomics confirm earliest domesticated sheep in southern Africa ca. 2000 BP. *Sci. Rep.* **11**, 6631 (2021).
76. Dickinson, M. R., Lister, A. M. & Penkman, K. E. H. A new method for enamel amino acid racemization dating: a closed system approach. *Quat. Geochronol.* **50**, 29–46 (2019).
77. Dickinson, M. R. *Enamel Amino Acid Racemisation Dating and its Application to Building Proboscidean Geochronologies*. PhD thesis, Univ. of York (2018).
78. Kaufman, D. S. & Manley, W. F. A new procedure for determining DL amino acid ratios in fossils using reverse phase liquid chromatography. *Quat. Sci. Rev.* **17**, 987–1000 (1998).
79. Hill, R. L. Hydrolysis of proteins. *Adv. Protein Chem.* **20**, 37–107 (1965).
80. Heath, T. A., Huelsenbeck, J. P. & Stadler, T. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proc. Natl Acad. Sci. USA* **111**, E2957–E2966 (2014).
81. Zhang, C., Stadler, T., Klopstein, S., Heath, T. A. & Ronquist, F. Total-evidence dating under the fossilized birth–death process. *Syst. Biol.* **65**, 228–249 (2016).
82. Kapur, V. V. & Bajpai, S. Oldest South Asian tapiromorph (*Perissodactyla*, *Mammalia*) from the Cambay Shale Formation, western India, with comments on its phylogenetic position and biogeographic implications. *Palaeobotanist* **64**, 95–103 (2015).
83. Carlisle, E., Janis, C. M., Pisani, D., Donoghue, P. C. J. & Silvestro, D. A timescale for placental mammal diversification based on Bayesian modeling of the fossil record. *Curr. Biol.* **33**, 3073–3082.e3 (2023).
84. Métais, G. et al. Lithofacies, depositional environments, regional biostratigraphy and age of the Chitarwata Formation in the Bugti Hills, Balochistan, Pakistan. *J. Asian Earth Sci.* **34**, 154–167 (2009).
85. Antoine, P.-O. et al. in *Fossil Mammals of Asia* (eds Fortelius, M. et al.) 400–422 (Columbia Univ. Press, 2013).
86. Pickford, M., Senut, B. & Hadoto, D. *Geology and Palaeobiology of the Albertine Rift Valley, Uganda-Zaire. Volume I: Geology* (Centre international pour la formation et les échanges géologiques, 1993).
87. Brown, F. H. & McDougall, I. Geochronology of the Turkana depression of northern Kenya and southern Ethiopia. *Evol. Anthropol.* **20**, 217–227 (2011).
88. Tong, H. & Moigne, A.-M. Quaternary rhinoceros of China. *Acta Anthropol. sinica* **19**, 257–263 (2000).
89. Antoine, P.-O. et al. A new rhinoceros clade from the Pleistocene of Asia sheds light on mammal dispersals to the Philippines. *Zool. J. Linn. Soc.* **194**, 416–430 (2022).
90. Heissig, K. *Palaeontologische und geologische untersuchungen im Tertiär von Pakistan. V. Rhinocerotidae (Mamm.) aus den unteren und mittleren Siwalik-Schichten* (Verlag der Bayerischen Akademie der Wissenschaften, in Kommission bei C.H. Beck, 1972).
91. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement (no. 861389 - PUSHH). E.C., I.P., K.E.H.P., F.R., J.V.O., M.M. and R.S.P. are supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 101021361 - BACKWARD). M.M. was also supported by the Danish National Research Foundation grant PROTEIOS (DNRF128), and M.D., K.E.H.P. and C.B. by the UK NERC (NE/S010211/1 & NE/S00713X/1). D.F. was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN-2018-05305) and a Research Activity Grant through the Canadian Museum of Nature. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether-Programme (Award HO 6201/1-1 to SH) and by the European Union (ERC, MacDrive, GA 101043187). B.D. was supported by the European Union (ERC-2023-COG HORIZON AviArch GA 101125532). M.L. received funding from the Leverhulme Research Grant RPG-2020-317. E.W. is supported by the Lundbeck Foundation (grant nos. R302-2018-2155, R155-2013-16338), the Novo Nordisk Foundation (grant no. NNF18SA0035006), the Wellcome Trust (grant no. UNS69906), Carlsberg Foundation (grant no. CF18-0024), the Danish National Research Foundation (grant nos. DNRF94, DNRF174), the University of Copenhagen (KU2016 programme), Ferring Pharmaceuticals A/S, and Germany's Excellence Strategy (EXC-2077), project no. 390741603. L.B. and R.S. thank M. Rubini of the Soprintendenza Archeologia Belle Arti e Paesaggio per le province di Frosinone e Latina for the permission to study *Stephanorhinus* samples from Fontana Ranuccio. L.B. and R.S. also acknowledge S. Grimaldi and F. Parenti (Istituto Italiano di Paleontologia Umana) for allowing us access to the Fontana Ranuccio collections and F. Strani for her valuable support in the selection of *Stephanorhinus* samples from Fontana Ranuccio. The authors thank L. Pandolfi for valuable constructive criticism on aspects of the manuscript. Field work in Canada's High Arctic was supported by a palaeontology permit to N.R. from the Government of Nunavut, Department of Culture, Language, Elders and Youth (D.R. Stenton, J. Ross), with the permission of the Qikiqtani Inuit Association, particularly Grise Fiord. M. Dawson's field work, which resulted in the collection of most of the rhinocerotid, in 1986, was supported by National Geographic. Associated fieldwork led by N.R. was also supported by the Canadian Museum of Nature and the Polar Continental Shelf Program.

Author contributions R.S.P., R.D.E.M., N.R., D.F. and E.C. designed the study. N.R., D.F. and M.G. conducted field work at the Houghton Crater site. R.D.E.M., N.R., D.F., M.G., R.S. and L.B. provided ancient samples. R.S.P., M.M., A.C., N.S.H., I.P., S.L., J.R.-M., A.D.R., F.M., M.R.D., C.B., G.S., M.L. and B.D. performed data generation and analysed data with support from E.C., S.H., E.W., A.M., N.R., R.D.E.M., F.R., J.V.O. and D.F. R.S.P., M.M. and E.C. wrote the manuscript with contributions from all authors.

Competing interests The authors declare no competing interests.

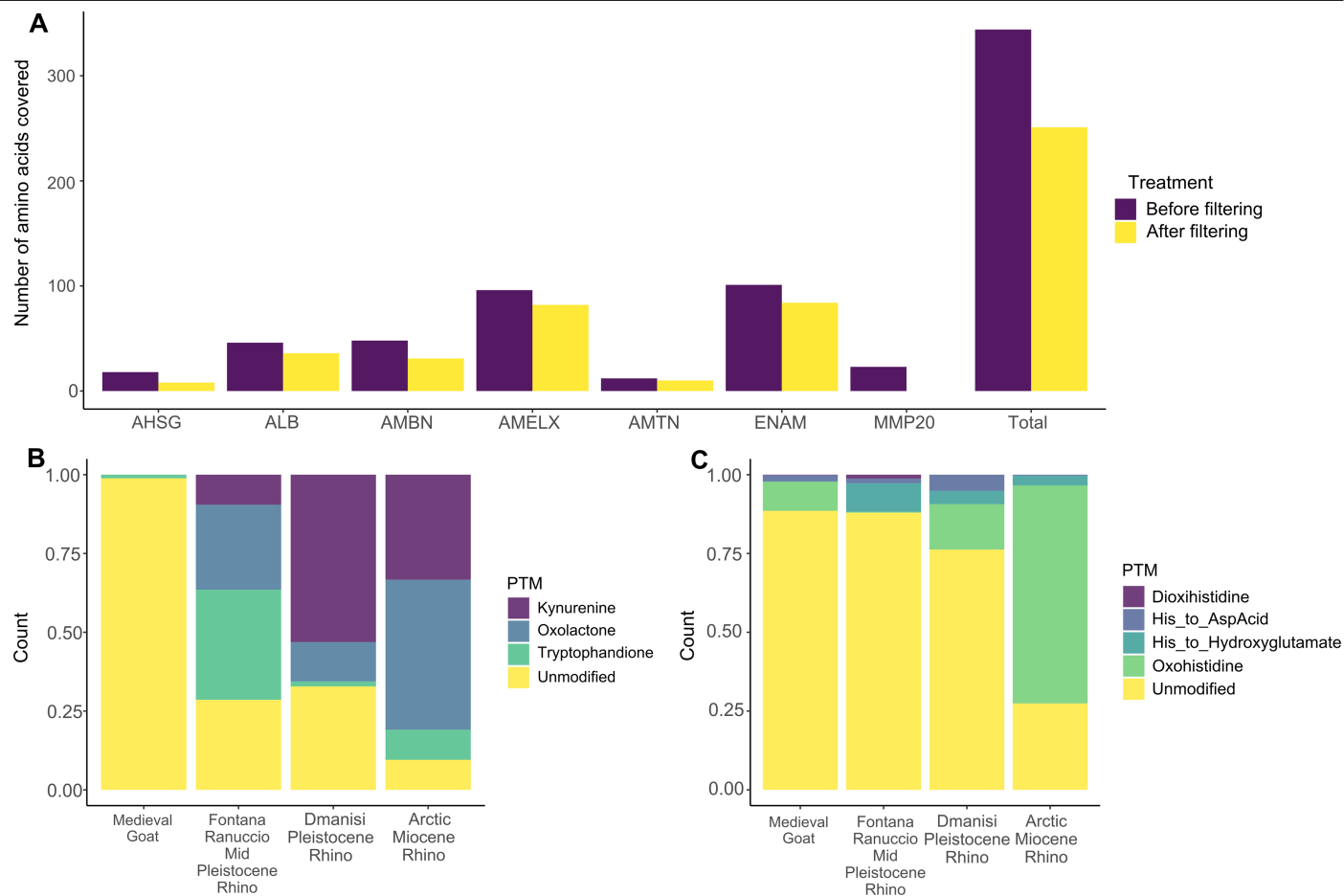
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-025-09231-4>.

Correspondence and requests for materials should be addressed to Ryan S. Paterson, Danielle Fraser, Natalia Rybczynski or Enrico Cappellini.

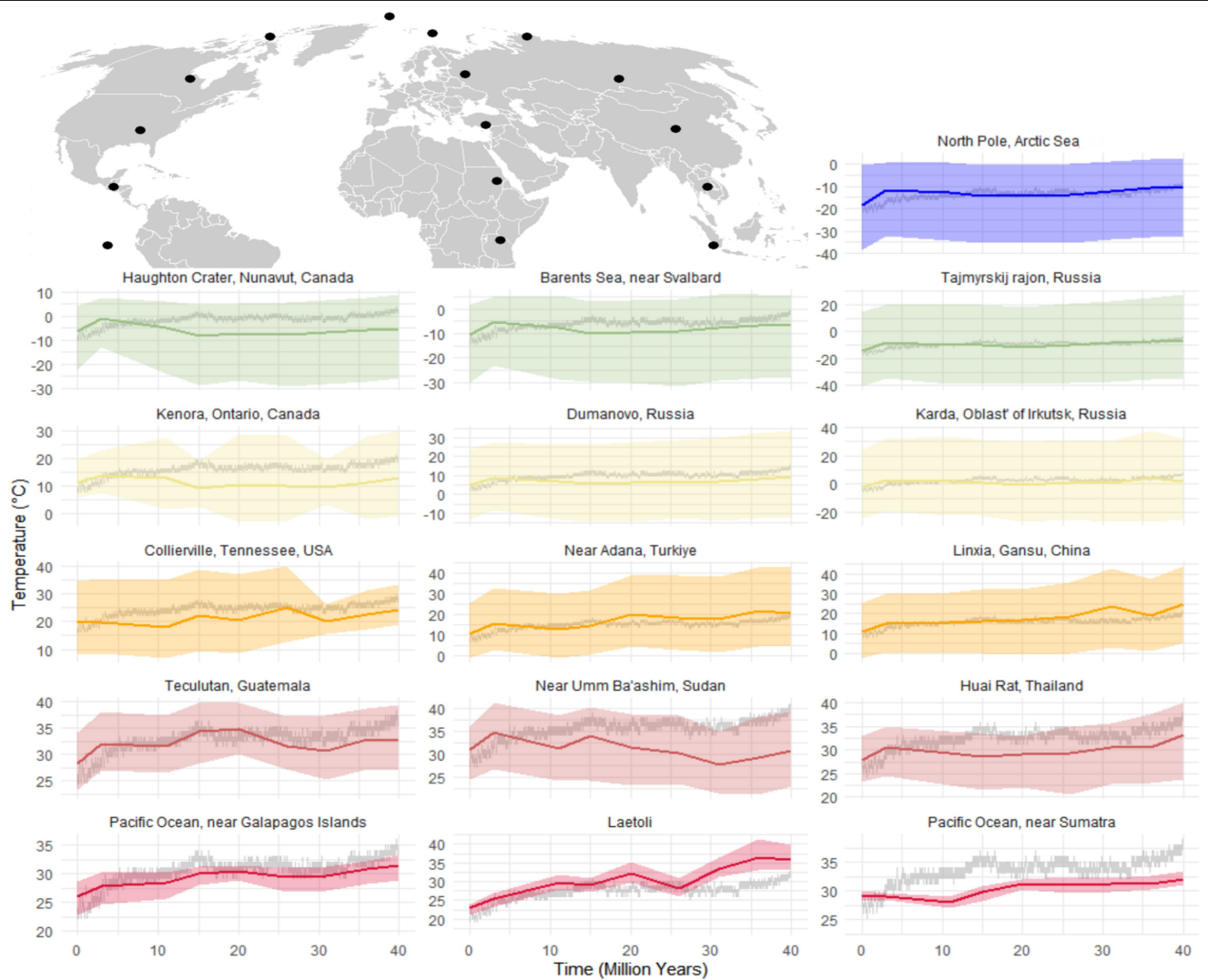
Peer review information *Nature* thanks Tao Deng, Jorune Sakalauskaite and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



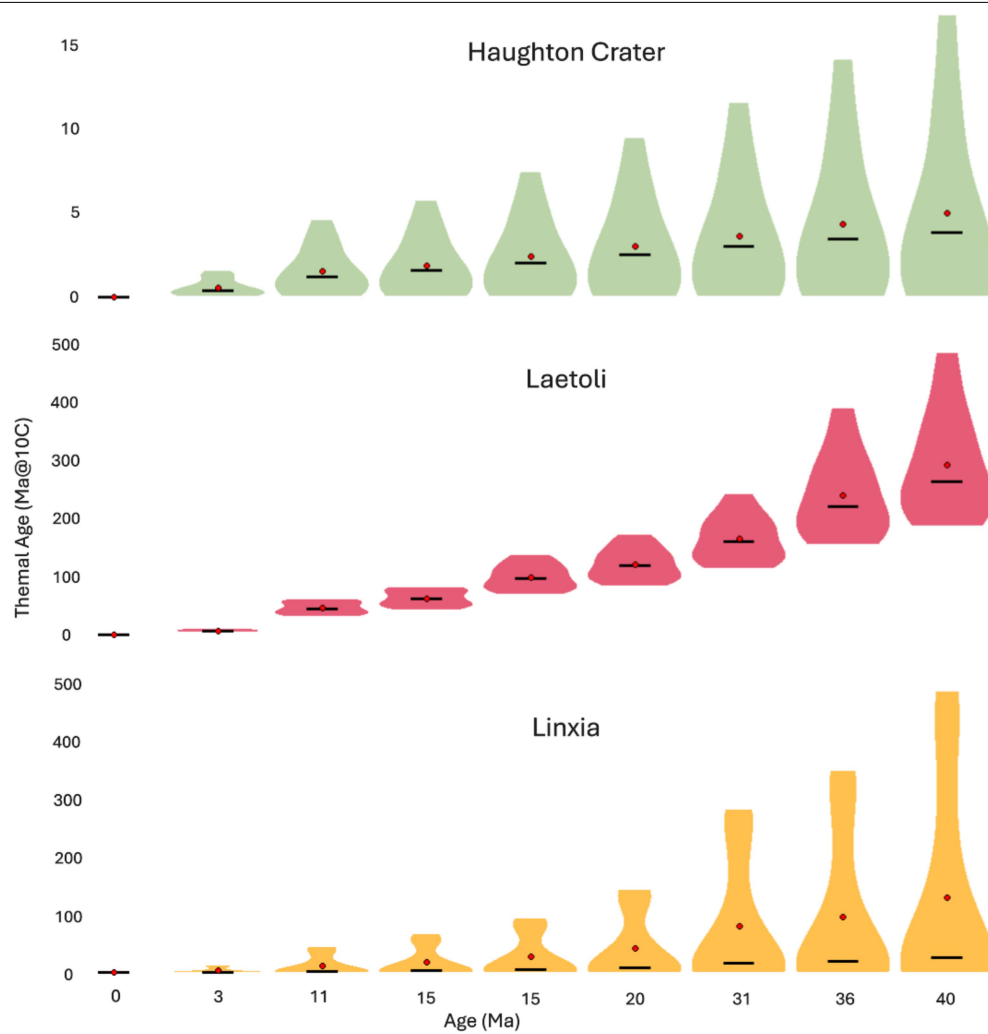
Extended Data Fig. 1 | Proteome preservation in CMNFV59632. A) Amino acid count for each identified protein, before and after our strict filtering step (see Methods), B) DIMs (diagenetically-induced modifications) related to oxidative degradation of tryptophan for CMNFV59632, compared to enamel proteomes from other ancient rhinos and a mediaeval ovicaprine, C) DIMs

related to oxidative degradation of histidine for same taxon set. The moderate protein preservation in CMNFV59632, indicated by the lower amino acid coverage compared to other ancient enamel proteomes, is further supported by the high incidence of modifications related to oxidative degradation, compared to other fossil rhinocerotids.



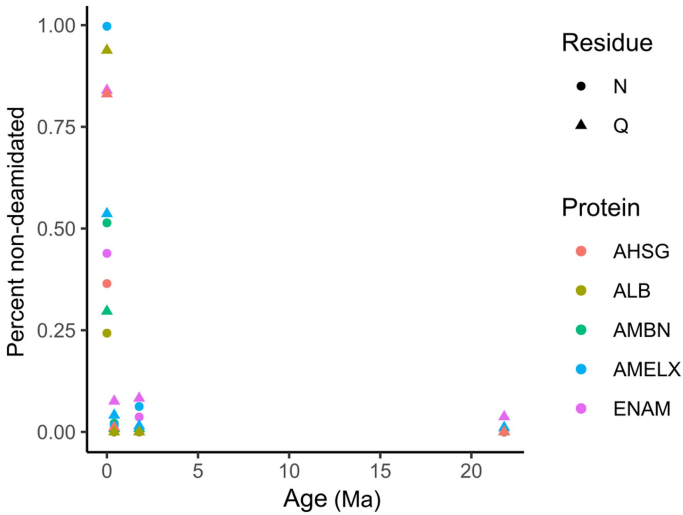
Extended Data Fig. 2 | Palaeotemperature values extracted from Valdes et al.³⁹. Monthly temperatures were used to calculate maximum and minimum values at 3, 11, 15, 20, 26, 31, 36, and 40 Ma. Note the wide yearly ranges for temperate and cold environments, reflecting seasonal maxima and minima.

The Hansen³⁸ model, scaled to current mean annual temperature values for each location, is shown as a grey curve. The basemap was generated using the maps package in R using ggplot2.

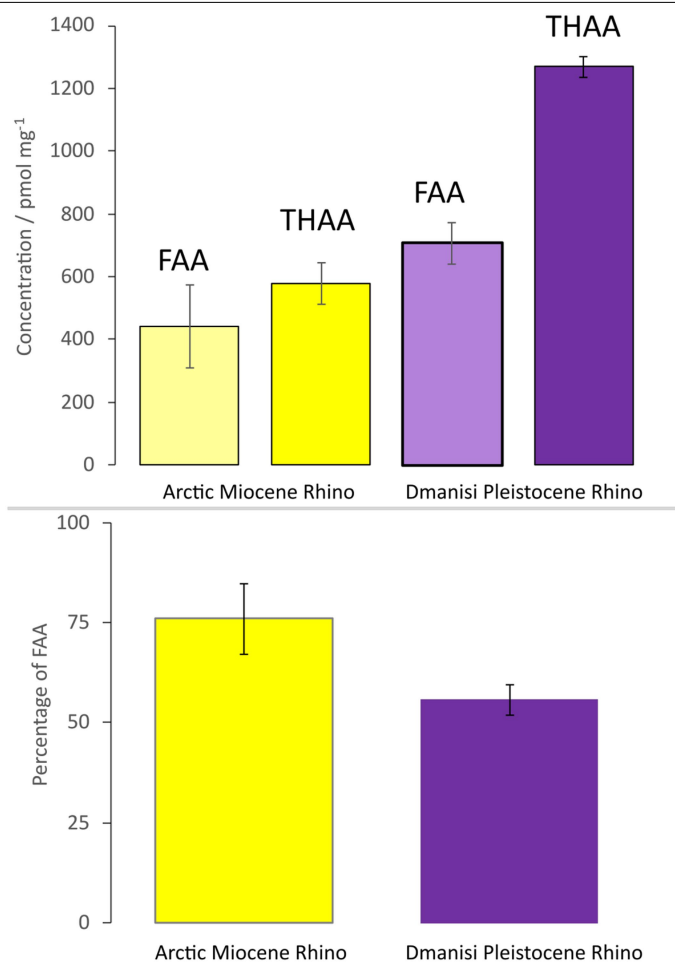


Extended Data Fig. 3 | Thermal age ranges estimated for selected sites preserving ancient peptides. These violin plots display the range of monthly thermal ages, mean and median (Ma@10 °C) calculated for Haughton Crater,

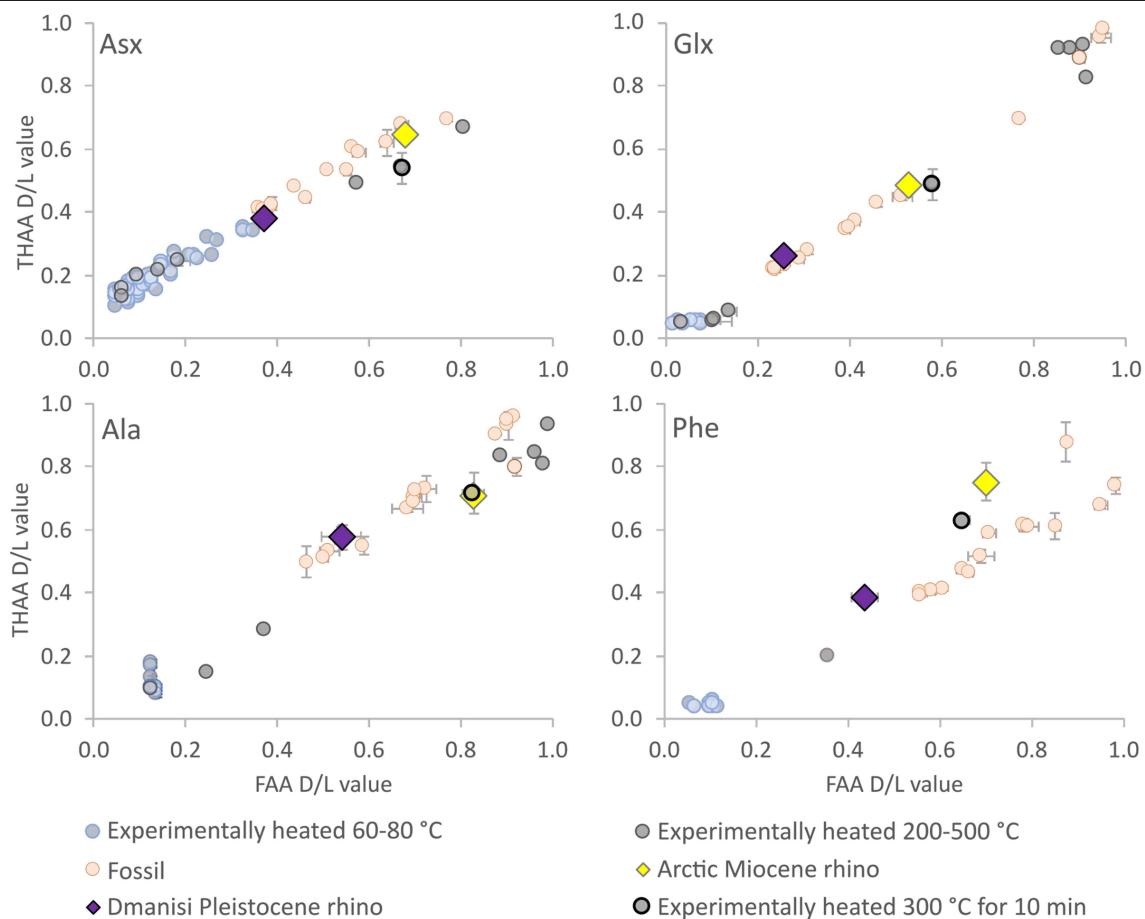
Laetoli and Linxia, for each time point. Note the scale difference between Haughton Crater and the other sites, and the high standard deviations for Haughton Crater and Linxia.



Extended Data Fig. 4 | Deamidation rates in fossil rhinocerotid enamel proteomes, plotted against geological age. Data used is CMFN-59632 from Haughton Crater (21.8 Ma), DM.5/157 from Dmanisi (1.77 Ma), CGG1_023342 from Fontana Ranuccio (0.4 Ma), and a mediaeval control sample (0.005 Ma). While useful for establishing authenticity of an ancient proteome, deamidation rates plateau relatively quickly, so they are not reliable for assessing relative degradative state in ancient proteomes from deep geological timescales.

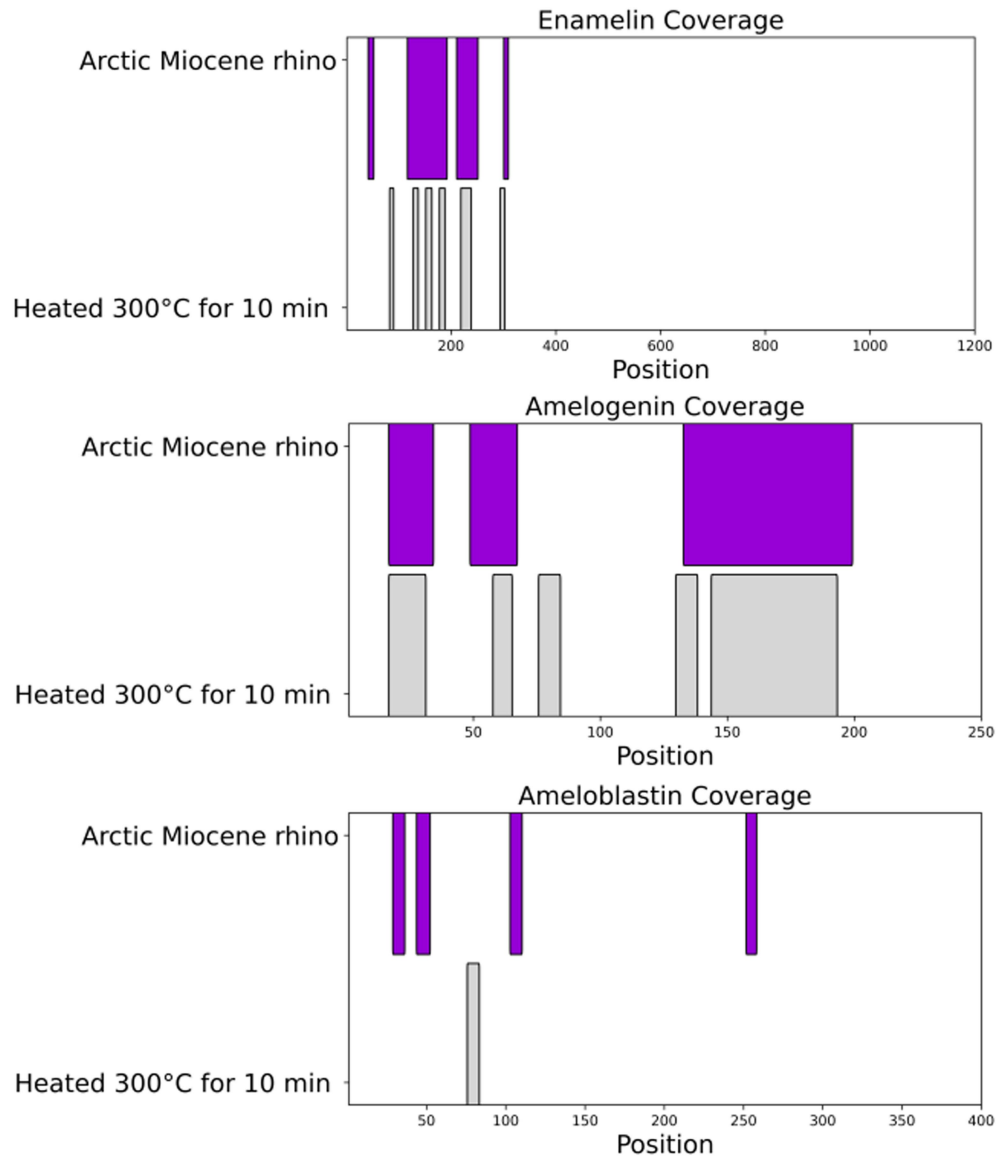


Extended Data Fig. 5 | Comparison of FAA & THAA concentration (top) and %FAA (bottom) of the Arctic Miocene rhinocerotid with the Dmanisi Pleistocene *Stephanorhinus*⁴. The lower overall concentration, higher %FAA and yet incomplete hydrolysis in the Arctic Miocene rhino is consistent with endogenous peptides in the tooth enamel. Error bars represent 1 standard deviation about the mean for preparative replicates.



Extended Data Fig. 6 | Asx, Glx, Ala and Phe FAA vs THAA D/L values for tooth enamel from Arctic Miocene rhino from Ellesmere Island Canada, and the Dmanisi Pleistocene rhino. A data set consisting of published and unpublished enamel data from other rhino palaeontological and experimental data has been included for comparison; peptides were also successfully retrieved from the experimentally heated sample in bold (300 °C for 10 min).

The good correlation between FAA & THAA for the Arctic Miocene rhino (CMNFV59632) sample supports the presence of closed system original peptides and their constituent amino acids in this Miocene sample. Error bars represent 1 standard deviation about the mean for preparative/experimental replicates.



Extended Data Fig. 7 | Similarity of protein sequences recovered between CMNFV59632 and experimentally-heated samples. Plots show sequence coverage for the three most abundant enamel matrix proteins (ENAM, AMELX & AMBN) for the Arctic Miocene rhino (CMNFV59632) tooth enamel and tooth enamel from *Coelodonta antiquitatis* that had been heated at 300 °C for 10 min, which showed similar levels of degradation (Extended Data Fig. 6). Peptides

were retrieved from similar regions for ENAM and AMELX. The Cappellini et al. (2019) method² was used to prepare *Coelodonta antiquitatis* enamel MS analysis, with some minor changes: HCl was used for demineralization instead of TFA, ZipTips (Millipore P10 with 0.6 µL C18 resin) were used instead of in-house StageTips and an Orbitrap Fusion Tribrid™ was used instead of a Q-Exactive for the proteomic analysis.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|---|
| n/a | Confirmed |
| ☒ | ☐ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ A description of all covariates tested |
| ☒ | ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☒ | ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☐ | ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Mass spectrometric data were acquired using the Xcalibur™ Software, controlling the Thermo Scientific™ LC-MS systems.
Data analysis	Custom code for a) phylogenetic analyses via RevBayes) and b) protein translation, is available in the supplementary data via FigShare. Xcalibur™ (version 4.1) MaxQuant (version 1.5.3.30) PEAKS (version 7.0) Python (version 3.5.4) RevBayes (version 1.2.1) blastall (version 2.2.26) mafft (version 7.205) ANGSD (version 0.913) BWA-MEM (version 0.7.7) BLAST+ (version 2.13.0) ProSplign pFind (version 3.1.6) FragPipe (version 17.1) ArcGIS(version 2.8.0) R package pastclim R, package ggplot2 (version 3.3.6) R, package stringr (version 1.4.0)

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The proteomics datasets have been deposited to the ProteomeXchange Consortium via the Proteomics Identifications Database (PRIDE) partner repository with the dataset identifier PXD052635. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender [Sex of specimens was not considered a priori due to the fragmentary nature of the fossil specimens.](#)

Reporting on race, ethnicity, or other socially relevant groupings [The study did not involve human participants.](#)

Population characteristics [See above](#)

Recruitment [Not relevant](#)

Ethics oversight [Not required](#)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Specimens of Rhinocerotidae were selected from multiple sites to test ancient protein preservation. Reconstructed sequence data from successful extractions were used in subsequent phylogenetic analyses. A replicate was performed for the protein extraction of the rhinocerotid tooth from Fontana Ranuccio. Controls were processed alongside samples (extraction and injection blanks) for every stage of data collection.
Research sample	Three specimens were sampled for ancient proteins - an enamel fragment from a tooth of Rhinocerotidae sp. from the Houghton Crater site, a tusk fragment from Rhinocerotidae sp. from the Houghton Crater site, and an enamel fragment from a molar tooth from Stephanorhinus sp. from the Fontana Ranuccio site. Previously-generated proteomic and genomic data also derives from Cappellini et al. (2019), generated from enamel fragments from Stephanorhinus sp. from Dmanisi, Georgia, and enamel fragments from a medieval ovicaprine, and Liu et al (2021), generated from modern and subfossilised tissues of various extant and extinct members of Rhinocerotidae.
Sampling strategy	No sample size calculation was performed. Samples deriving from only three palaeontological specimens were process, to reduce destruction of palaeontological materials.
Data collection	Wet labwork for protein extraction was performed by RP in dedicated clean labs at Globe Institute, University of Copenhagen, for the extraction of ancient biomolecules, wearing protective gear to prevent contamination. Acquisition of MS data was performed by MM at the Novo Nordisk Center for Protein Research, University of Copenhagen. Amino acid chirality analysis was performed by FM at Department of Chemistry, University of York.

Timing and spatial scale	Mass spectrometry data was collected on separate days for each specimen.
Data exclusions	No data were excluded from our analyses
Reproducibility	No attempts to repeat the experiment were made due to the destructive nature of sampling, but multiple samples were analysed for each specimen.
Randomization	Not relevant due to the sampling approach involving finite palaeontological materials.
Blinding	Not relevant to our study

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☐	☒ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Palaeontology and Archaeology

Specimen provenance	Specimens and their provenance are described separately in Fraser et al. (2024) ("Post-Eocene Rhinocerotid Dispersal via the North Atlantic"). Approval for destructive sampling was provided by stewards of the vertebrate palaeontology collections at the Canadian Museum of Nature.
Specimen deposition	Specimens were previously deposited at the Canadian Museum of Nature
Dating methods	No new dates are provided
☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.	
Ethics oversight	As these are palaeontological samples, no ethical approval is required, but the permit for destructive sampling is described above under "Specimen provenance"

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.
Novel plant genotypes	Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.