bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A 20+ Ma old enamel proteome from Canada's High Arctic reveals diversification of Rhinocerotidae in the middle Eocene-Oligocene

Authors

Ryan S. Paterson[1*], Meaghan Mackie[1,2], Alessio Capobianco[3,4], Nicola S. Heckeberg[3,4], Danielle Fraser[5,6,7,8], Fazeelah Munir[9], Ioannis Patramanis[1], Jazmín Ramos-Madrigal[1], Shanlin Liu[10], Abigail D. Ramsøe[1], Marc R. Dickinson[9], Chloë Baldreki[9], Marisa Gilbert[5], Raffaele Sardella[11], Luca Bellucci[12], Gabriele Scorrano[1,13], Fernando Racimo[1], Eske Willerslev[1,14,15], Kirsty E.H. Penkman[9], Jesper V. Olsen[2], Ross D.E. MacPhee[16], Natalia Rybczynski[5*], Sebastian Höhna[3,4], Enrico Cappellini[1*]

* Corresponding authors: Ryan Sinclair Paterson: ryan.paterson@sund.ku.dk; Natalia Rybczynski: nrybczynski@nature.ca; Enrico Cappellini: ecappellini@sund.ku.dk

Affiliations

[1]Globe Institute, University of Copenhagen, 5-7 Øster Voldgade, 1350, Copenhagen, Denmark, [2]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3b, 2200, Copenhagen, Denmark, [3]GeoBio-Center LMU, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, Munich 80333, Germany, [4]Department of Earth and Environmental Sciences, Palaeontology & Geobiology, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, Munich 80333, Germany, [5]Palaeobiology, Canadian Museum of Nature, PO Box 3443 Stn "D", Ottawa ON K1P 6P4, Canada, [6]Department of Earth Sciences, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada, [7]Department of Biology, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada, [8]Department of Paleobiology, Smithsonian Institution, National Museum of Natural History, 10th and Constitution NW, Washington, DC 20013-

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

7012, [9]Department of Chemistry, University of York, York, YO10 5DD, UK, [10]Department of Entomology, College of Plant Protection, China Agricultural University, Beijing, China, [11]Sapienza University of Rome, Department of Earth Sciences (PaleoFactory lab.), Piazzale Aldo Moro 5, 00185, Rome, Italy, [12]Museo di Geologia e Paleontologia, Sistema Museale di Ateneo, Università di Firenze, via Giorgio La Pira 4, 50121, Firenze, Italy, [13]Center for Molecular Anthropology for the Study of Ancient DNA, Department of Biology, University of Rome "Tor Vegata", via della Ricerca Scientifica n. 1, 173, Rome, Italy, [14]Department of Zoology, University of Cambridge, 17 Mill Lane, CB2 1RX, Cambridge, UK, [15]Marum Center for Marine Environmental Sciences and Faculty of Geosciences, University of Bremen, Leobener Str. 8, 28359, Bremen, Germany, [16]Department of Mammalogy, American Museum of Natural History, 200 Central Park West, 10024, New York City, United States.

In the past decade, ancient protein sequences have emerged as a valuable source of data for deep-time phylogenetic inference. Still, the recovery of protein sequences providing novel phylogenetic insights does not exceed 3.7 Ma (Pliocene). Here, we push this boundary back to 21-24 Ma (early Miocene), by retrieving enamel protein sequences of an early-diverging rhinocerotid (*Epiaceratherium* sp. - CMNF-59632) from the Canadian High Arctic. We recover partial sequences of seven enamel proteins (AHSG, ALB, AMBN, AMELX, AMTN, ENAM, MMP20) and over 1000 peptide-spectrum matches, spanning over at least 251 amino acids. Authentic endogeneity of these sequences is supported by indicators of protein damage, including several spontaneous and irreversible post-translational modifications accumulated during prolonged diagenesis and reaching near-complete occupancy at many sites. Bayesian tip-dating, across 15 extant and extinct perissodactyl taxa, places the divergence time of CMNF-59632 in the middle Eocene-Oligocene, and identifies a later divergence time for Elasmotheriinae in the Oligocene. The finding weakens alternative models suggesting a deep basal split between Elasmotheriinae and Rhinocerotinae. This divergence time of CMNF-59632 coincides with a phase of high diversification of rhinocerotids, and supports a Eurasian origin of

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

this clade in the late Eocene or Oligocene. The findings are consistent with previous hypotheses on the origin of the enigmatic fauna of the Haughton crater, which, in spite of their considerable degree of endemism, also display similarity to distant Eurasian faunas. Our findings demonstrate the potential of palaeoproteomics in obtaining phylogenetic information from a specimen that is ten times older than any sample from which endogenous DNA has been obtained.

Phylogenetic placement of deep-time (>1 Ma) fossils has typically relied on morphological observations, as the recovery of sufficiently extensive genetic evidence has not been thought to be possible beyond the Pleistocene[1]. While ancient DNA (aDNA) sequences are a valuable source of data for inferring phylogenies and population dynamics in the late Pleistocene[2–5], the oldest authentic aDNA from macrofossils has been extracted from Arctic-situated specimens dated to no more than 1.2 Ma[6]. In contrast, palaeoproteomic data have been recovered from late Miocene, Pliocene and early Pleistocene fossils, even in localities that are warm, humid, and/or at low latitudes[7,8]. Although protein sequences from the early Pleistocene have been successfully used to infer the phylogenetic placement of various fossil mammals[9–11], the precise limit of proteomic survival has not yet been systematically characterised[12,13]. Currently, the oldest confirmed palaeoproteomic data successfully used to infer sub-ordinal taxonomic relationships derive from bone collagen of camelids from the 3.7 Ma Fyles Leaf Bed site of Canada’s High Arctic[14,15]. This illustrates how our understanding of evolutionary relationships is currently limited by the preservation of biomolecules from extinct species.

Rhinocerotidae is a family that includes only five extant species, but a wide diversity of fossil members[16,17]. It remains debated as to where and when the radiation of this group occurred[18]. For most of the last two decades, the group was defined by a deep ‘basal split’ between two clades - Rhinocerotinae and Elasmotheriinae - prior to bursts of rhinocerotid diversification in the late Eocene[19–23]. 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 rhino (Rhinocerotinae, *Coelodonta antiquitatis*)[24]. Recently, the sequenced genomes of *Coelodont*a and *Elasmotherium*[25] were used to confirm hypotheses based

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

on morphological data that suggest they have distinct phylogenetic affinities[19], but also allowed for the recognition of a relatively-young 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 makes it difficult to assess the timing of the Rhinocerotinae-Elasmotheriinae split in relation to other radiations that occurred within the group. For these reasons, the ancient radiations of the group still remain obscured.

To investigate the timing of Rhinocerotidae divergence and the potential for evolutionarily informative protein sequences to persist in deep-time, we targeted vertebrate dental enamel deriving from the Haughton crater (75°N, Nunavut) in Canada's High Arctic (Figure 1). The Haughton crater is an impact structure with its stratigraphy including post-impact fossiliferous lacustrine sediments dated to 21-24 Ma[26]. While their geological age is advanced, fossils from these sediments are found in a polar landscape, currently characterised by permafrost. This creates a temperature regime favourable for biomolecular preservation, sparing these fossils from the harshest effects of diagenesis. Accordingly, specimens from this site serve as promising candidates for biomolecular preservation in deep-time.

The digestion-free palaeoproteomic workflow[9,10] applied to an early Miocene rhinocerotid (*Epiaceratherium* sp.) specimen of dental enamel[27] from the Haughton Formation (21.8 Ma) allowed for the recovery of an enamel proteome covering 1163 confident peptide-spectrum matches (PSMs), at least seven proteins (AHSG, ALB, AMBN, AMELX, AMTN, ENAM, MMP20), and spanning at least 251 amino acids (Figure 2a) (Ext. Figure 1a). The enamel proteome of CMNF-59632 currently represents both the oldest mammalian skeletal proteome currently reported, confirming the predicted deep-time persistence of ancient mammalian proteins from high latitudes[8,10], and the first biomolecular characterisation of the extinct genus *Epiaceratherium*. While the survival of a relatively rich enamel proteome from such ancient deposits is surprising, the age of the specimen belies its excellent state of preservation.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To better appreciate the preservation state of the Haughton Crater enamel proteome, we compared it to those of two other rhinocerotids, namely the early Pleistocene *Stephanorhinus* from the site of Dmanisi (Georgia), dated at 1.77 Ma[9], and a middle Pleistocene *Stephanorhinus* (~0.4 Ma) from the site of Fontana Ranuccio (Italy). While the set of proteins retrieved from the CMNF-59632 enamel specimen is similar to that of the other two Pleistocene rhinocerotids used for comparison (Figure 2a), fewer peptides and a shorter reconstructed amino acid sequence were recovered from the Arctic specimen. In addition to proteins previously found in deep-time enamel samples (AHSG, AMBN, AMELX, AMTN, ENAM, MMP20), serum albumin was also found in each rhinocerotid sample. Previously recovered from other fossil enamel specimens[11], serum albumin is phylogenetically more informative than enamel-specific proteins due to its higher amino acid sequence variability[28]. In the present analysis, while some albumin peptides are removed during initial data filtering (as they are identical to peptide sequences from bovine serum albumin, a potential laboratory contaminant), several others do not match contaminants. The vast majority of the spectra that confidently support their identification show post-translational modifications (PTMs) that derive from prolonged diagenesis, which supports their authenticity.

As expected, diagenetic modifications and PTMs are extensive in the enamel proteome of CMNF-59632 (Figure 2b). Average peptide lengths are similar, though slightly shorter than those of the Dmanisi early Pleistocene specimen, and further reduced in comparison to the Fontana Ranuccio middle Pleistocene *Stephanorhinus*, indicating a greater degree of peptide bond hydrolysis (Figure 2b). We also observe high deamidation rates in CMNF-59632, though no more so than in the Pleistocene rhinocerotids (Ext. Figure 2). While high deamidation rates can be useful for confirming proteome authenticity, they can be highly variable within samples[29–31], and can plateau relatively-quickly in fossil proteomes, reducing their utility in characterising degradation patterns in deep-time (Figure 2c). Instead, we identify a suite of informative spontaneous PTMs indicative of advanced diagenesis that are observed at a higher rate in the Arctic Miocene rhinocerotid, providing support for their utility as markers of advanced diagenesis and authenticity in deep-time (Figure 2c)[9]. These include arginine to ornithine conversion (Figure 2c) and advanced forms of tryptophan (Ext. Figure

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2b) and histidine oxidation (Ext. Figure 1c). Intra-crystalline protein decomposition analysis further confirms the advanced degradation state of CMNF-59632. The concentration of free and total hydrolysable amino acids (FAA and THAA, respectively) is around half of those in the early Pleistocene *Stephanorhinus* sample from Dmanisi (Extended Figure 2a), and the percentage of FAA in CMNF-59632 (~75%) is higher than in the Pleistocene rhino from Dmanisi (~50%) (Ext. Figure 3b), supporting increased peptide bond hydrolysis. Furthermore, the racemisation values for CMNF-59632 fall along the expected FAA vs THAA trends for both fossil enamel, and the experimentally-heated samples (Ext. Figure 4), confirming the closed system behaviour of CMNF-59632 enamel amino acids, and supporting the endogeneity of the peptides retrieved.

At least 10 single amino acid polymorphisms (SAPs) support the placement of CMNF-59632 within Rhinocerotidae. A smaller number (2+) of SAPs are shared between CMNF-59632 and other perissodactyls, to the exclusion of later-diverging rhinocerotids. No novel variants are uncovered in CMNF-59632, as the aforementioned SAPs represent character states retained from ancestors within Perissodactyla and Mammalia more broadly. The identification of these SAPs is supported by several unique PSMs displaying almost complete ion series (e.g., Ext. Figure 5).

Peptide sequences recovered from CMNF-59632 derive from similar sequence regions to those previously identified in the Dmanisi Pleistocene *Stephanorhinus* proteome (Figure 2d), particularly for the three most abundant enamel matrix proteins (EMPs). ENAM and AMBN present broadly similar sequence coverage patterns in both specimens, though 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 Peptide (LRAP)[9,32].

Regardless of the mechanisms behind preferential mass spectrometric and data analysis identification of specific sequence regions, biases favouring the recovery[7] and identification[33] of conserved peptide sequences can ultimately lead to underestimates of divergence times in taxa

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

represented by empirically-derived protein sequences. To accurately estimate the phylogenetic position of CMNF-59632 and estimate divergence times within the group, we completed a phylogenetic analysis of a suite of extinct and extant perissodactyls. In addition to the perissodactyl taxa used in Cappellini et al. (2019)[9], 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 Fossilised Birth Death (FBD) model infers CMNF-59632 as the earliest diverging rhinocerotid in the analysis, with *Elasmotherium sibiricum* being more closely related to Rhinocerotina (crown rhinoceroses) than to CMNF-59632 (Figure 4). This phylogenetic hypothesis is also supported by Fraser et al. (2024) in a total-evidence analysis. Additionally, 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 CMNF-59632 and all other rhinocerotids during the middle Eocene–Oligocene (around 41–25 Ma). The divergence between *Elasmotherium sibiricum* and Rhinocerotina is reconstructed to have likely occurred in the Oligocene (around 34–22 Ma), which is younger than previous molecular clock estimates[25].

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)[34], 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)[35]. During this timeframe, other early-diverging lineages are also appearing in Asia[18,36,37], and eventually spreading as far as Western Europe[18]. Morphologically, the Haughton crater rhinocerotid shares closer affinities with these early-diverging lineages from Eurasia[38], particularly those within the genus *Epiaceratherium*[27]. Similarly, some other vertebrates within the highly-endemic fauna of the Haughton Formation have their closest relatives in Eurasia. These include the transitional pinniped *Puijila darwini*, sister to the Oligocene *Potamotherium* of Europe[39], and the swan-like anatid, a group which is otherwise restricted to the Oligocene and Miocene of Europe[38]. Overall, these patterns,

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 of Rhinocerotinae, as it diverges before the Rhinocerotinae-Elasmotheriinae split. We also reject a deep-divergence (basal split) between Elasmotheriinae and Rhinocerotinae[19–23] and find moderate support for their branching event after the divergence of *Epiaceratherium*. Our analysis disagrees with that of Kosintsev et al. (2019)[23], who find 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 in spite of equivalently old ages for crown Ceratomorpha (earliest Eocene). Among other timetrees, our dates are generally most consistent with those of Liu et al. (2021)[25]. Our recovered topologies are also broadly similar to trees derived from the morphology-based phylogenetic analyses of Tissier et al. (2020)[18] and Lu et al. (2023)[40], identifying Elasmotheriinae and Rhinocerotinae as deeply-nested within Rhinocerotidae. Discrepancies between the genomic [25] and proteomic trees arise likely due to different calibration points. The more ancient age of Elasmotheriinae in the analysis of Liu et al. (2021)[25] 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, as 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 allows for robustly-supported timetrees, and can serve to develop phylogenetic frameworks in deep-time, particularly from specimens too old to preserve ancient DNA. For example, the present data allows for firm placement of the Haughton Crater rhinocerotid outside of Rhinocerotina, and likely outside the Elasmotheriinae-Rhinocerotinae clade, a fact which has significant implications for both morphological and molecular studies integrating fossil calibration times from the fossil record.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Furthermore, we demonstrate the deep-time survival of a rich set of peptides derived from proteins present in mammalian enamel, well beyond the previously known limits of survival. This work illustrates the power of palaeoproteomics in elucidating phylogeny and taxonomy of extinct vertebrates in deep-time. These findings should encourage further vertebrate palaeontological fieldwork in the High Arctic, and other cold-temperature sites with taphonomic conditions favourable to biomolecular preservation.

References

1. Kjær, K. H. *et al.* A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature* **612**, 283–291 (2022).
2. Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78 (2013).
3. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15758–15763 (2013).
4. Meyer, M. *et al.* A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* **505**, 403–406 (2014).
5. Barlow, A. *et al.* Middle Pleistocene genome calibrates a revised evolutionary history of extinct cave bears. *Curr. Biol.* **31**, 1771–1779.e7 (2021).
6. van der Valk, T. *et al.* Million-year-old DNA sheds light on the genomic history of mammoths. *Nature* **591**, 265–269 (2021).
7. Demarchi, B. *et al.* Protein sequences bound to mineral surfaces persist into deep time. *Elife* **5**, (2016).
8. Demarchi, B. *et al.* Survival of mineral-bound peptides into the Miocene. *Elife* **11**, (2022).
9. Cappellini, E. *et al.* Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny. *Nature* **574**, 103–107 (2019).
10. Welker, F. *et al.* Enamel proteome shows that Gigantopithecus was an early diverging pongine.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Nature **576**, 262–265 (2019).

11. Welker, F. *et al.* The dental proteome of Homo antecessor. *Nature* **580**, 235–238 (2020).
12. 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**, (2017).
13. Umamaheswaran, R. & Dutta, S. Preservation of proteins in the geosphere. *Nat Ecol Evol* (2024) doi:10.1038/s41559-024-02366-z.
14. Rybczynski, N. *et al.* Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution. *Nat. Commun.* **4**, 1550 (2013).
15. 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).
16. Prothero, D. R. Rhinocerotidae. in *Evolution of tertiary mammals of North America* 595 (1998).
17. Pandolfi, L. Evolutionary history of Rhinocerotina (Mammalia, Perissodactyla). *Fossilia* **2018**, 27–32 (2018).
18. Tissier, J., Antoine, P.-O. & Becker, D. New material of Epiaceratherium and a new species of Mesaceratherium clear up the phylogeny of early Rhinocerotidae (Perissodactyla). *R Soc Open Sci* **7**, 200633 (2020).
19. Antoine, P. O. & national d’histoire naturelle, M. Phylogénie et évolution des Elasmotheriina (Mammalia, Rhinocerotidae). (2002).
20. Antoine, P.-O. Middle Miocene elasmotheriine Rhinocerotidae from China and Mongolia: taxonomic revision and phylogenetic relationships. *Zool. Scr.* **32**, 95–118 (2003).
21. 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).
22. 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).
23. Kosintsev, P. *et al.* Evolution and extinction of the giant rhinoceros Elasmotherium sibiricum

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sheds light on late Quaternary megafaunal extinctions. *Nat Ecol Evol* **3**, 31–38 (2019).

24. Cerdeño, E. *Cladistic Analysis of the Family Rhinocerotidae (Perissodactyla)*. (American Museum of Natural History, 1995).

25. Liu, S. *et al.* Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. *Cell* **184**, 4874–4885.e16 (2021).

26. Jessberger, E. K. 40Ar-39Ar Dating of the Haughton Impact Structure. *Meteoritics* (1988).

27. Fraser D, Rybczynski N, Gilbert M, Dawson MR. Post-Eocene Rhinocerotid Dispersal via the North Atlantic. *BioRxiv* (2024) doi:10.1101/2024.06.04.597351.

28. Nei, M. *Molecular Evolutionary Genetics*. (Columbia University Press, 1987).

29. 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).

30. Brown, S. *et al.* Examining collagen preservation through glutamine deamidation at Denisova Cave. *J. Archaeol. Sci.* **133**, 105454 (2021).

31. Pal Chowdhury, M. & Buckley, M. Trends in deamidation across archaeological bones, ceramics and dental calculus. *Methods* **200**, 67–79 (2022).

32. Gibson, C. W. *et al.* Identification of the leucine-rich amelogenin peptide (LRAP) as the translation product of an alternatively spliced transcript. *Biochem. Biophys. Res. Commun.* **174**, 1306–1312 (1991).

33. 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).

34. Hanson, C. B. Teletaceras Radinskyi, a new primitive Rhinocerotid from the Late Eocene Clarno Formation of Oregon. in *The evolution of perissodactyls* 379–398 (1989).

35. Prothero, D. R. *The Evolution of North American Rhinoceroses*. (Cambridge University Press, 2005).

36. 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).

37. Böhme, M. *et al.* Na Duong (northern Vietnam)-an exceptional window into Eocene ecosystems

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from Southeast Asia. *Zitteliana. Reihe A, Mitteilungen der Bayerischen Staatssammlung für Paläontologie und Geologie* **53**, 121–167 (2013).

38. Whitlock, C. & Dawson, M. R. Pollen and Vertebrates of the Early Neogene Haughton Formation, Devon Island, Arctic Canada. *Arctic* **43**, 324–330 (1990).

39. 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).

40. Lu, X.-K., Deng, T. & Pandolfi, L. Reconstructing the phylogeny of the hornless rhinoceros Aceratheriinae. *Frontiers in Ecology and Evolution* **11**, (2023).

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

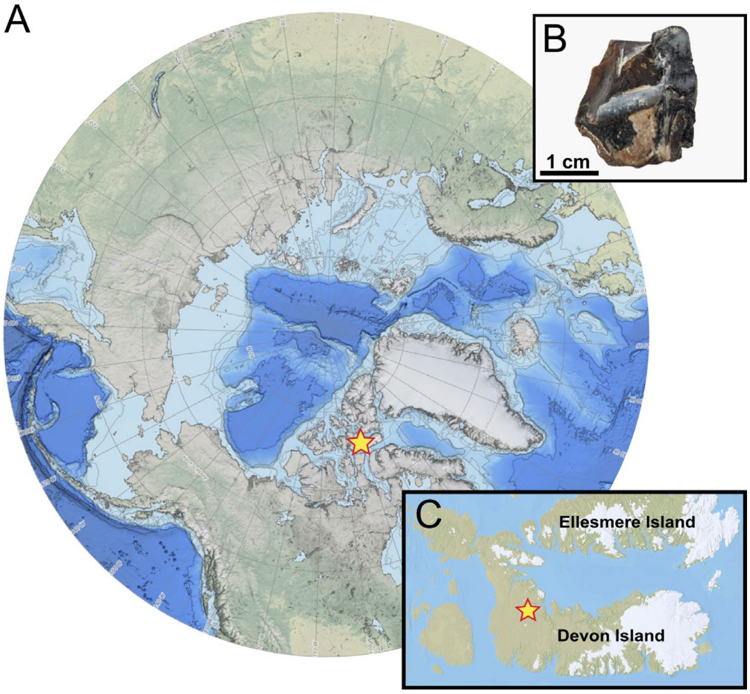


Figure 1) The high latitude Haughton crater on Devon Island has produced a highly endemic vertebrate fauna. A) Location of Devon Island within the circumpolar North. B) Anterolingual view of specimen CMNF-59632 after destructive palaeoproteomic analysis. C) Location of Haughton crater on Devon Island.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

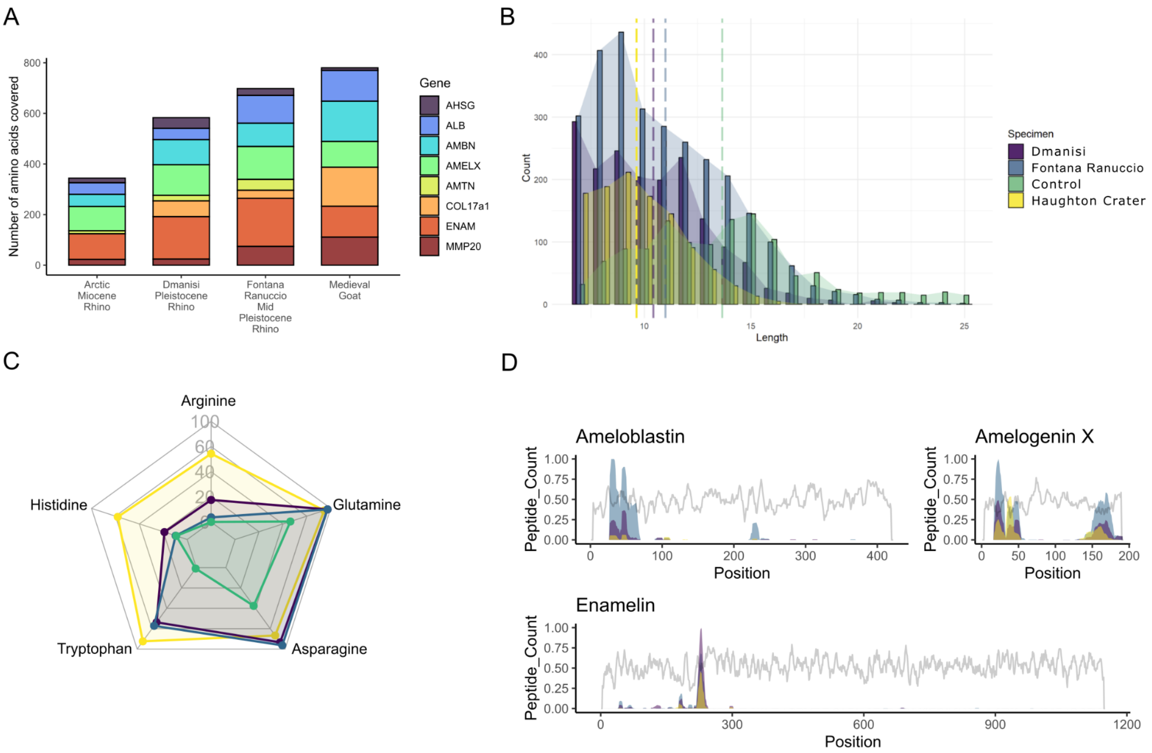


Figure 2) Proteome preservation in the enamel specimen of the early Miocene rhinocerotid (CMNF-59632). Preservation is compared to enamel proteomes from an early Pleistocene (1.77 Ma) *Stephanorhinus* (DM.5/157), a middle Pleistocene (0.4 Ma) *Stephanorhinus* (CGG 1_023342) and a medieval ovicaprine (Control)[9]. All plots exclude contaminants and reverse hits. A) Amino acid sequence coverage for each identified protein; B) peptide length distributions, dashed bars represent average peptide length for each specimen; C) proportion of a selected sample of amino acids that are often modified in ancient enamel proteomes. Results derive from PTM-specific searches described in methods. 'Arginine' includes arginine-to-ornithine conversion, 'Glutamine' includes glutamine deamidation, 'Asparagine' includes asparagine deamidation, 'Tryptophan' includes advanced tryptophan oxidation to kynurenine, oxylactone, and tryptophandione, 'Histidine' includes oxidation and dioxidation of histidine, histidine conversion to hydroxyglutamate. D) Sequence coverage plots for the three most abundant enamel matrix proteins (AMBN, AMELX, ENAM), recording number of PSMs (coloured areas) and mutability (grey line).

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

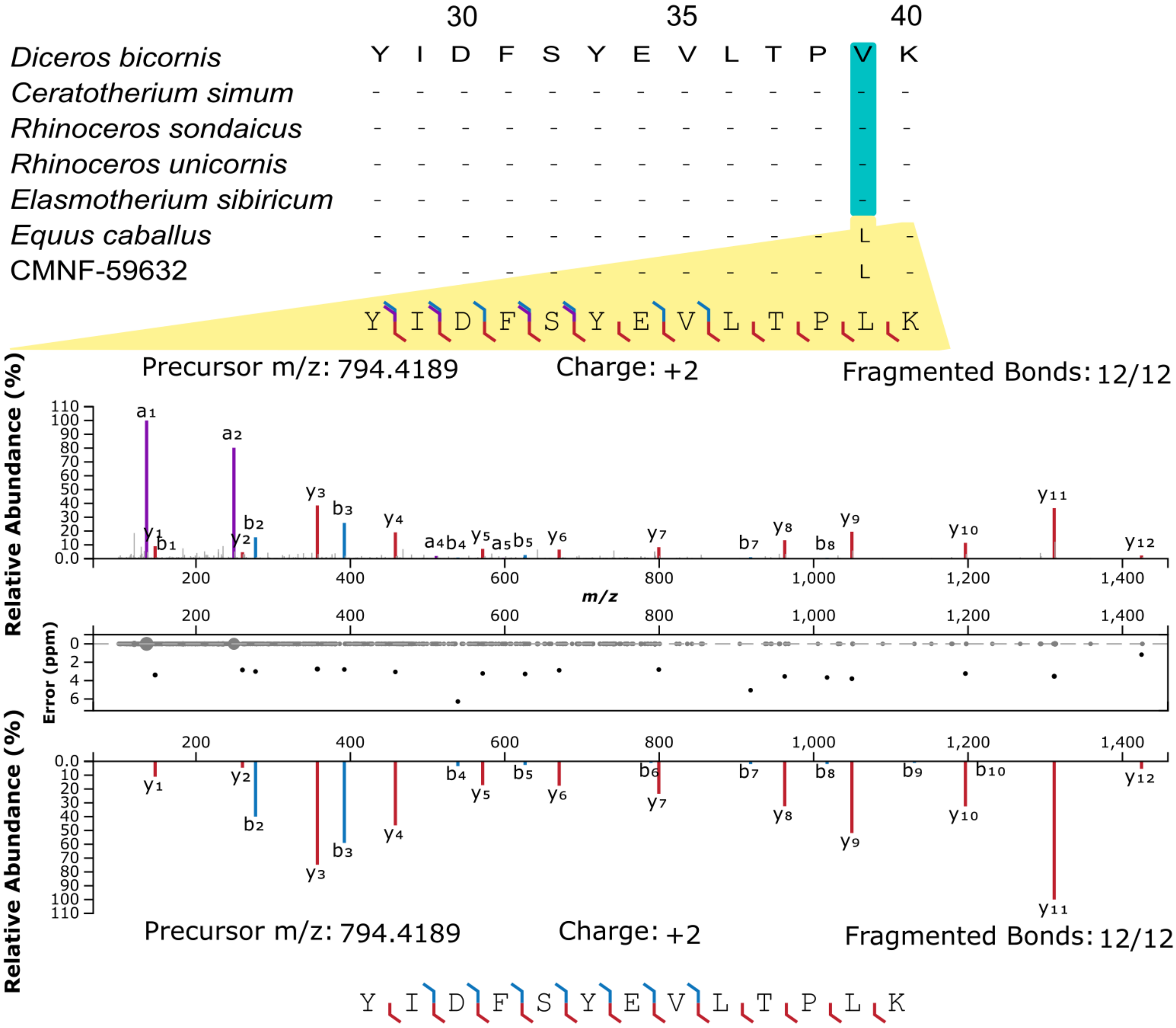


Figure 3) Abridged alignment and mirror plots of a phylogenetically-informative single amino acid polymorphism (SAP) at AMELX-39. The top spectrum is experimentally-derived, while the bottom one is predicted using the 'Original mode' with the Prosit tool, available online via the Universal Spectrum Explorer[83]. This spectrum is the highest scoring peptide-spectrum match (with Andromeda) for AMELX sequence positions spanning the most abundantly-covered SAP differentiating between CMNF-59632, and all other rhinocerotids for which sequences are available.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

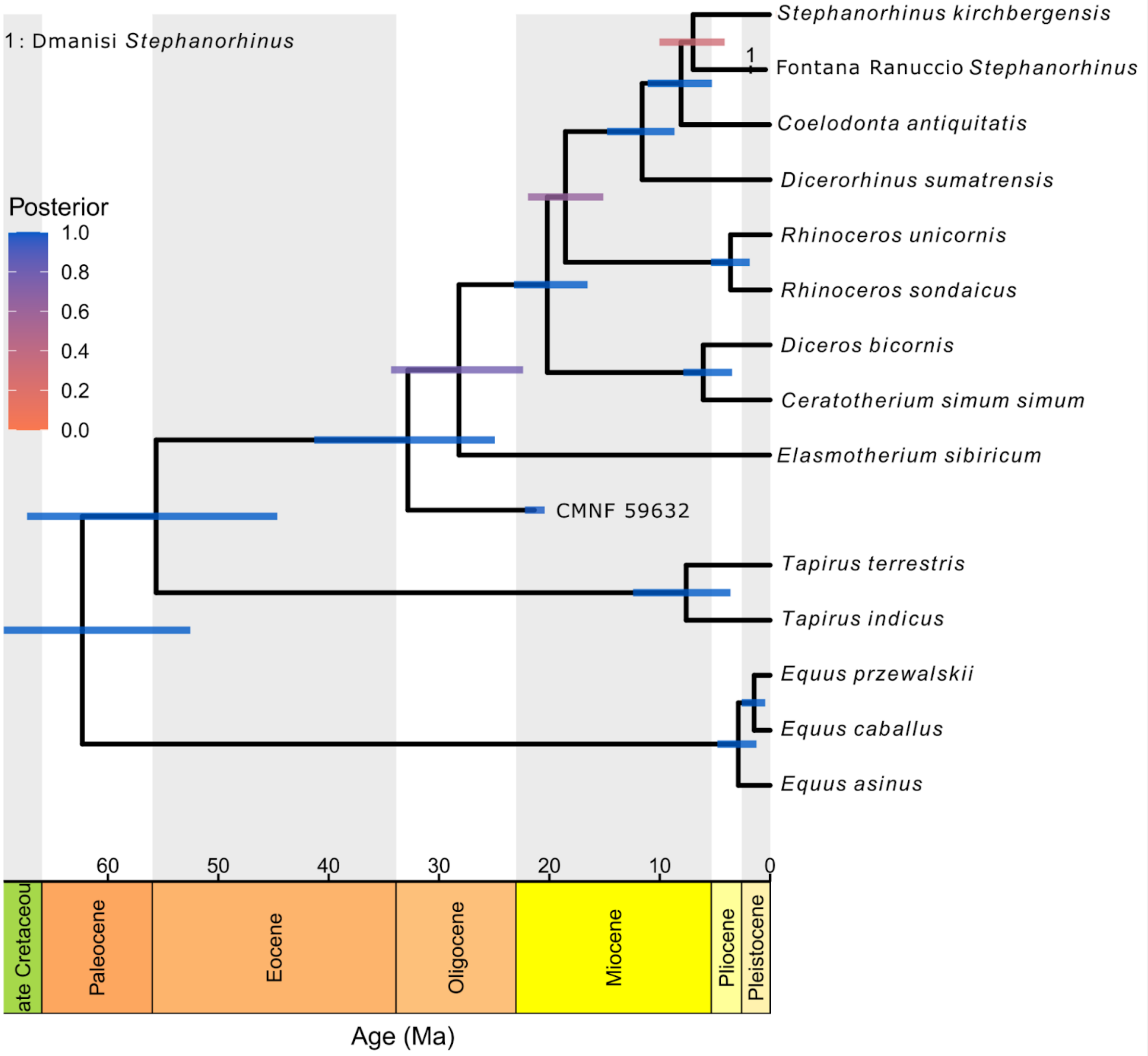


Figure 4) Time-calibrated phylogeny of Rhinocerotidae enamel proteomes. The maximum *a posteriori* (MAP) tree was produced using RevBayes v.1.2.1 [70]; https://revbayes.github.io, with a Fossilized Birth Death (FBD) model. Coloured bars at nodes represent 95% height posterior density (HPD) age interval estimates. Specimen CMNF-59632 represents the early Miocene rhinocerotid from the Haughton Crater.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and Methods

Site and Specimen

Located within the Haughton impact crater (75°N, Nunavut, Canada), the Haughton 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[26,41]. An early Miocene age has also been corroborated by (U-Th)/He thermochronology[42]. While older age estimates between 30-40 Ma have also been suggested[43–45], 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 Haughton formation consists of several vertebrate taxa, including a transitional pinniped[39], a pair of salmoniform fishes, a swan-like anatid, a small artiodactyl, a leporid rabbit, a heterosocid shrew, and a well-preserved rhinocerotid[27,38]. While the megafloral assemblage is not particularly rich, the palynofloral assemblage is well-characterised, allowing for reconstruction of local climatic conditions. In the early Miocene, the Haughton crater lake and its surrounding environs experienced a significantly warmer annual temperature (8-12°C) than the present day[38,46].

The specimen CMNF-59632 is a nearly complete rhinocerotid skeleton, including skull and dentition, uncovered 10.8 m above the base of the formation[27]. The present analysis focuses on a single tooth fragment from a lower left m1 (Figure 1b) that was already separated from the rest of its tooth row due the fragmenting effectings of cryoturbation. The dental specimen's rhinocerotid affinities are further supported by its size and morphology, most notably the presence of vertical Hunter-Schreger bands on its enamel, a defining feature of rhinocerotids and found in few other mammals[47]. A single tusk fragment (left i2) derived from CMNF-59632 was also selected for proteomic extraction. Due to its thin enamel, only limited peptides were recovered from this tusk fragment, and the sample is thus excluded from further analysis and discussion.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Proteomic extraction and LC-MS/MS

The laboratory workflow for the CMNF-59632 teeth and the Fontana Ranuccio *Stephanorhinus* tooth (for comparison) generally follows that of Cappellini et al. (2019) and Taurozzi et al. (2024)[9,48]. Using a sterilised drill, flakes of enamel were removed from the fragmentary teeth, with care taken to avoid sampling the dentine. The CMNF-59632 tooth enamel sample, 154 mg, was then ground to a fine powder, and demineralized overnight using 10% HPLC-grade trifluoroacetic acid (TFA) (Merck, Sigma-Aldrich). The CMNF-59632 tusk enamel sample, weighing 90 mg, was processed in the same way. The Fontana Ranuccio enamel sample was divided into three subsamples - FR2, FR3, and FR4 - weighing 202, 243, and 205 mg, respectively. They were similarly ground to a fine powder, and demineralized using 10% TFA (FR3, FR4) or 10% HCl (hydrochloric acid, FR2). For each sample, the demineralization step was repeated a second time to ensure complete demineralization. As enamel peptides are already hydrolysed *in vivo*, no enzymatic digestion was performed. Subsequently, peptides were collected and desalted on C-18 StageTips[49] 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 (FA) into a 96-well plate. To remove ACN and concentrate the samples, they were vacuum-centrifuged until approximately 3 µL of sample remained. Next, samples were resuspended in 6 µL of 5% ACN 0.1% formic acid (FA).

Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) was used to analyse the samples, based on previously published protocols[9,50]. 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, Germany) on an EASY-nLC 1200 (Proxeon, Odense, Denmark) connected to an Exploris 480 (CMNF-59632) or a Q-Exactive HF-X (Fontana Ranuccio *Stephanorhinus*) mass spectrometer (both from Thermo Scientific, Bremen, Germany), with an integrated column oven. 4 (CMNF-59632) or 5

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(FR sd - 295 - *Stephanorhinus*) µL of sample were injected. 0.1% FA in milliQ water was used as buffer A and the peptides were separated with increasing buffer B (80% ACN 0.1% FA) with a 77 min gradient, increasing 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. An integrated column oven was used to maintain the temperature at 40°C.

The two mass spectrometers were run with the same parameters except where specified, due to changes in running software. Spray voltage was set to 2 kV, the S-lens RF 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 350-1400. The AGC target value was set to 300% / 3e6 (Exploris / HF-X) with a maximum injection time of 25 ms. HCD-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 / 108 ms (Exploris / HF-X), with an AGC target value of 200% / 2e5 (Exploris / HF-X). Normalised collision energy was set at 30 / 28% (Exploris / 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 used in Cappellini *et al*. (2019)[9] was used as a starting point to construct a database for sequence reconstruction. Due to the vast evolutionary distance between CMNF-59632 and any extant taxa (>20 million years), 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 NCBI for each enamel protein, specifying the taxonomic grouping of ‘Theria’ to include all therian mammals. To supplement available sequences, additional sequences were manually extracted from available genomes, following the methodology from[51].

To investigate the relationships at the base of Rhinocerotidae, protein sequences translated from *Elasmotherium sibiricum* genomic data[25] were generated. To obtain the corresponding amino acid sequences, we firstly collapsed the Paired-End (PE) reads and masked the conflict bases as "N"

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

using adapterRemoval[52]. Then, we mapped the collapsed reads against the reference genome of the white rhinoceros (GCF_000283155.1_CerSimSim1) using the BWA MEM function[53] with the shorter split hits being abandoned. After that, we removed duplicates using an in-house Perl script following Liu et al. (2021)[25]. Finally, we extracted the gene sequences according to their locations on the reference genome.

The remaining steps generally follow that outlined by[9]. To translate relevant genes, we used ANGSD[54] to create consensus sequences for only those BAM files representing chromosomes with 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 alignment[55], with the corresponding *Ceratotherium simum* sequences as queries. Finally, we used ProSplign to recover the spliced alignments, and ultimately, the translated protein sequences[56].

Protein identification

Thermo .raw files generated by the mass spectrometer were searched with various softwares using an iterative search strategy to interpret spectra, characterise PTMs, and ultimately, reconstruct protein sequences. For comparison, .raw files from a mediaeval ovicaprine (control) and an early Pleistocene *Stephanorhinus* generated by Cappellini et al. (2019)[9] were also analysed. Among samples from the Fontana Ranuccio *Stephanorhinus*, only FR4 was analysed.

We primarily employed MaxQuant[57] for sequence reconstruction and other downstream aspects of data analysis. We performed two initial runs: 1) a more focussed run using the database we modified from Cappellini et al. (2019)[9], and 2) a broad run using the ‘Theria’-wide database we constructed from publicly-available sequences.

In all runs, an Andromeda 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 (FDR) was used (0.01), while protein FDR was

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

increased to 1 to show possible low-abundance proteins. "Unspecific" digestion was specified. No fixed post-translational modifications were set. Several PTMs were set as variable modifications in our initial runs: glutamine and asparagine deamidation, methionine and proline oxidation, N-terminal pyroglutamic acid from glutamic and aspartic acids, phosphorylation of serine, threonine, and tyrosine, and the conversion of arginine to ornithine.

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 manually removed and not further examined. 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 additional search tools. Peaks v. 7.0 wass used to attempt *de novo* sequencing and homology search was performed using the SPIDER algorithm[58–60]. The open search capabilities of openPFind[61] and MSFragger[62] 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 + filtering

Prior to 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)[63] to determine if peptides match any contaminants, beyond those included in

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MaxQuant by default, such as soil bacteria and fungi. Next, MS/MS spectra are manually inspected for each PSM to examine ion series coverage. At this stage, peptide sequences are accepted for the strictly-filtered dataset only if each amino acid residue is covered (e.g., at least y-, b-, or a- ion designates the mass of that specific amino acid, plus any identified PTMs) by at least two spectra, following the approach outlined by Coutu et al. (2020)[64]. Additionally, for both strict- and minimally-filtered datasets, poorly-supported spectra are also 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.

Protein damage analysis

Characterization of protein degradation and post-translational modifications roughly follows that of Cappellini et al. (2019)[9]. In addition to the primary run, three additional runs were performed on each rhinocerotid sample, alongside the mediaeval control, to assess several different post-translational modifications: 1) oxidative degradation of tryptophan, and included kynurenine (ΔM = +3.994915), oxolactone (ΔM = +13.979265), tryptophandione (ΔM = +29.974178) as variable modifications, 2) oxidative degradation of histidine (His), including oxohistidine (ΔM = +15.9949), dioxohistidine (ΔM = +31.990), His to hydroxyglutamate (ΔM = +7.979), and His -to aspartic acid (ΔM = -22.032) as variable modifications, and 3) aromatic, including oxidation (WY) (ΔM = +15.9949) and dioxidation (WY) (ΔM = +31.990) as variable modifications. Deamidation (NQ) was also included as a variable modification for each run. After removing potential contaminants and reverse hits, we used spectral counting to assess the extent of each PTM. In the case of arginine, histidine, and tryptophan, we summed the total number of each amino acid residues in each sample, after filtering for reverse hits and potential contaminants, and the total number of each modified amino acid residue in each sample (Figure 2C and Ext. Figure 1BC). Deamidation levels were calculated following the approach described in Mackie *et al.* (2018)[50] (Figure 2C) and site specific rates were observed using the DeamiDATE algorithm[65].

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Intra-crystalline protein decomposition analysis

Chiral amino acid analysis was undertaken on CMNF-59632 to evaluate the overall extent of amino acid degradation in the intra-crystalline fraction of the enamel, enabling comparison to previously-analysed specimens[9], and samples that had been experimentally-heated between 60 and 80ºC for up to 17520 hours and to samples heated to 200 - 500 ºC for up to 25 min. Enamel chips were drilled using a Dremel ® 4000 (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 of Dickinson et al. (2019)[66]. To remove excess powders, enamel chips were washed in deionized water and ethanol (Analytical-grade) before being powdered with an agate pestle and mortar. Powdered samples were weighed into a single plastic microcentrifuge tube and bleached (NaOCl-12%, 50 µL mg^{-1} of enamel) for 72 hours 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 replicate analysis of the free amino acids (FAA) and two for replicate analysis for the total hydrolysable amino acids (THAA). The THAA subsamples were dissolved in HCl (7 M, 20 µL mg^{-1}, Analytical grade) in a sterile 2 mL glass vial (Wheaton), purged with N_2 to reduce oxidation and heated at 110℃ for 24 h in an oven (BINDER GmbH series). The acid was then removed by centrifugal evaporation (Christ RVC2-25). Then, THAA and FAA fractions were subjected to a biphasic separation procedure[66,67] to remove inorganic phosphate from the enamel samples. HCl was added to both FAA (1 M, 25 µL mg^{-1}) and THAA (1 M, 20 µL mg^{-1}) fractions in separate 0.5 mL plastic microcentrifuge tubes (Eppendorf), and KOH (1 M, 28 µL mg^{-1}) was added into the acidified solutions, which then formed mono-phasic cloudy suspensions. Samples were agitated and then samples were 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-

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

crystalline amino acids, and their extent of racemisation (D/L value) were then quantified using RP-HPLC (Agilent 1100 series HPLC fitted with HyperSil C18 base deactivated silica column (5 µm, 250 x 3 mm) and fluorescence detector) following a modified method of Kaufman & Manley (1998)[68].

For the RP-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 acetonitrile:methanol:sodium buffer (21 mM sodium acetate trihydrate, sodium azide,1.3 µM EDTA, pH adjusted to 6.00 ± 0.01 with 10% acetic acid and sodium hydroxide)) was used for analysis. D and L peaks of the following amino acids were separated: aspartic acid and asparagine (Asx); glutamic acid and glutamine (Glx); serine (Ser), alanine (Ala), valine (Val), phenylalanine (Phe), isoleucine (Ile), leucine (Leu), threonine (Thr), arginine (Arg), tyrosine (Tyr) and glycine (Gly). During preparation, asparagine and glutamine undergo rapid irreversible deamination to aspartic acid and glutamic acid respectively[69] and hence they are reported together as Asx and Glx respectively.

Phylogenetic Analysis

A time-calibrated phylogenetic tree was inferred with the Bayesian phylogenetic software RevBayes v.1.2.1[70]; https://revbayes.github.io/) under a constant-rate Fossilised Birth Death (FBD) model[71,72]. The dataset consisted of enamel proteome data for 16 perissodactyl species (10 extant and 6 extinct), totalling 7 proteins and 3446 amino acids. Phylogenetic analyses were performed with both the strict-filtered and minimally-filtered sequences for CMNF-59632, to observe any topological differences between the two datasets and assess if filtering is warranted. As no major differences were observed, only the results from the 'strictly-filtered' dataset are discussed. The proteome dataset was partitioned by protein. A GTR + I (General Time Reversible + Invariant sites) amino acid substitution model—where 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**(Gamma)-distributed rate variation only for MMP20, hence **Γ**-distributed rate variation was modelled only for the

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MMP20 partition. A relaxed clock model with uncorrelated lognormal-distributed rates (UCLN) was applied to allow rate variation across branches. The prior on the average clock rate was set as a loguniform distribution (min=10^{-8}, max=10^{-2} substitutions per lineage/million year). 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 fossilisation (sampling of species in the past). The priors on speciation, extinction, and fossilisation parameters were set as uniform distributions bounded between 0 and 10. The sampling probability for extant species was fixed to 0.5882353 (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*, *E. przewalskii*, *E. asinus*) were constrained as outgroup to other perissodactyls (Tapiridae + Rhinocerotidae). Tip ages of fossil taxa were given a uniform prior distribution ranging from minimum to maximum age of the deposit where 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 Shale[73]), and maximum = 100 Ma, corresponding to the beginning of the Late Cretaceous and a very lax upper boundary on the origin of placental mammals[74]. Additional constraints on node ages based on the fossil record of perissodactyls were set to improve 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

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

rhinoceros in the fossil record (*Gaindatherium* cf. *browni* from the Aquitanian upper member of the Chitarwata Formation[75,76]; soft maximum = 44 Ma, corresponding to the minimum age of Rhinocerotidae as supported by fossil and phylogenetic evidence [25]. 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[77,78]; soft maximum = 7.3 Ma, as in Liu et al. (2021)[25]. 4) Node = *Rhinoceros unicornis* + *Rhinoceros sondaicus*; soft minimum = 1.9 Ma, corresponding to the early Pleistocene appearance of *Rhinoceros unicornis* in the fossil record[79,80]; soft maximum = 5.3 Ma, as in Liu et al. (2021)[25]. 5) Node = *Dicerorhinus* + *Stephanorhinus* + *Coelodonta*; soft minimum = 13 Ma, corresponding to middle Miocene remains of *Dicerorhinus* from the middle Siwaliks of Pakistan[25,81]; 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 (MCMC) was set up as 4 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 (ESSs) of parameter estimates on Tracer v.1.7.2[82]. A maximum *a posteriori* (MAP) tree was calculated to summarise the posterior distribution of trees, with 20% burn-in. In the analysis of the minimally-filtered dataset, one of the 4 runs was discarded from the MAP tree calculation, as it converged only in the last 10% of the MCMC.

Methods References

41. Omar, G. *et al.* Fission-track dating of haughton astrobleme and included biota, devon island, Canada. *Science* **237**, 1603–1605 (1987).

42. Young, K. E. *et al.* Impact thermochronology and the age of Haughton impact structure, Canada. *Geophys. Res. Lett.* **40**, 3836–3840 (2013).

43. 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).

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

44.Sherlock, S. C. *et al.* Re-evaluating the age of the Haughton impact event. *Meteorit. Planet. Sci.* **40**, 1777–1787 (2005).

45.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).

46.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).

47.Von Koenigswald, W., Holbrook, L. T. & Rose, K. D. Diversity and Evolution of Hunter-Schreger Band Configuration in Tooth Enamel of Perissodactyl Mammals. *acpp* **56**, 11–32 (2011).

48.Taurozzi, A. J. *et al.* Deep-time phylogenetic inference by paleoproteomic analysis of dental enamel. *Nat. Protoc.* (2024) doi:10.1038/s41596-024-00975-3.

49.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).

50.Mackie, M. *et al.* Palaeoproteomic Profiling of Conservation Layers on a 14th Century Italian Wall Painting. *Angewandte Chemie* vol. 130 7491–7496 Preprint at https://doi.org/10.1002/ange.201713020 (2018).

51.Rüther, P. L. *et al.* SPIN enables high throughput species identification of archaeological bone by proteomics. *Nat. Commun.* **13**, 2458 (2022).

52.Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **5**, 337 (2012).

53.Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

54.Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).

55.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).

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

56.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).

57.Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).

58.Han, Y., Ma, B. & Zhang, K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. *Proc. IEEE Comput. Syst. Bioinform. Conf.* 206–215 (2004).

59.Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A. & Ma, B. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *& cellular proteomics* (2012).

60.Ma, B. & Johnson, R. De novo sequencing and homology searching. *Mol. Cell. Proteomics* **11**, O111.014902 (2012).

61.Chi, H. *et al.* Open-pFind enables precise, comprehensive and rapid peptide identification in shotgun proteomics. *bioRxiv* 285395 (2018) doi:10.1101/285395.

62.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).

63.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).

64.Coutu, A. N. *et al.* Palaeoproteomics confirm earliest domesticated sheep in southern Africa ca. 2000 BP. *Sci. Rep.* **11**, 6631 (2021).

65.Ramsøe, A. *et al.* DeamiDATE 1.0: Site-specific deamidation as a tool to assess authenticity of members of ancient proteomes. *J. Archaeol. Sci.* **115**, 105080 (2020).

66.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).

67.Dickinson, M. R. Enamel amino acid racemisation dating and its application to building Proboscidean geochronologies. (University of York, 2018).

68.Kaufman, D. S. & Manley, W. F. A new procedure for determining dl amino acid ratios in

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fossils using reverse phase liquid chromatography. *Quat. Sci. Rev.* **17**, 987–1000 (1998).

69.Hill, R. L. Hydrolysis of proteins. *Adv. Protein Chem.* **20**, 37–107 (1965).

70.Höhna, S. *et al.* RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Syst. Biol.* **65**, 726–736 (2016).

71.Heath, T. A., Huelsenbeck, J. P. & Stadler, T. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences* **111**, E2957–E2966 (2014).

72.Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A. & Ronquist, F. Total-Evidence Dating under the Fossilized Birth-Death Process. *Syst. Biol.* **65**, 228–249 (2016).

73.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).

74.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).

75.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).

76.Antoine, P.-O. *et al.* Chapter 16. Mammalian Neogene Biostratigraphy of the Sulaiman Province, Pakistan. in *Fossil Mammals of Asia* 400–422 (Columbia University Press, 2013).

77.Pickford, M., Senut, B. & Hadoto, D. Geology and palaeobiology of the Albertine Rift valley, Uganda-Zaire. Volume I: geology. *Publication occasionnelle-Centre international pour la formation et les échanges géologiques* (1993).

78.Brown, F. H. & McDougall, I. Geochronology of the Turkana depression of northern Kenya and southern Ethiopia. *Evol. Anthropol.* **20**, 217–227 (2011).

79.Tong, H. & Moigne, A.-M. Quaternary rhinoceros of China. *Acta anthropologica sinica* **19**, 257–263 (2000).

80.Antoine, P.-O. *et al.* A new rhinoceros clade from the Pleistocene of Asia sheds light on

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mammal dispersals to the Philippines. *Zool. J. Linn. Soc.* **194**, 416–430 (2022).

81.Heissig, K. Palaeontologische und geologische untersuchungen im Tertiaer von Pakistan. V. Rhinocerotidae (Mamm.) aus den unteren und mittleren Siwalik-Schichten. (1972).

82.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).

83.Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).

Acknowledgements: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement (No. 861389 - PUSHH). EC, FZ, IP, KP, FR, JVO, MM and RP are supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101021361 - BACKWARD). MM was also supported by the Danish National Research Foundation grant PROTEIOS (DNRF128), and MD, KP & CB by the UK NERC (NE/S010211/1 & NE/S00713X/1). DF 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-Program (Award HO 6201/1-1 to SH) and by the European Union (ERC, MacDrive, GA 101043187). Fieldwork in Canada's High Arctic was supported by a palaeontology permit 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.

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

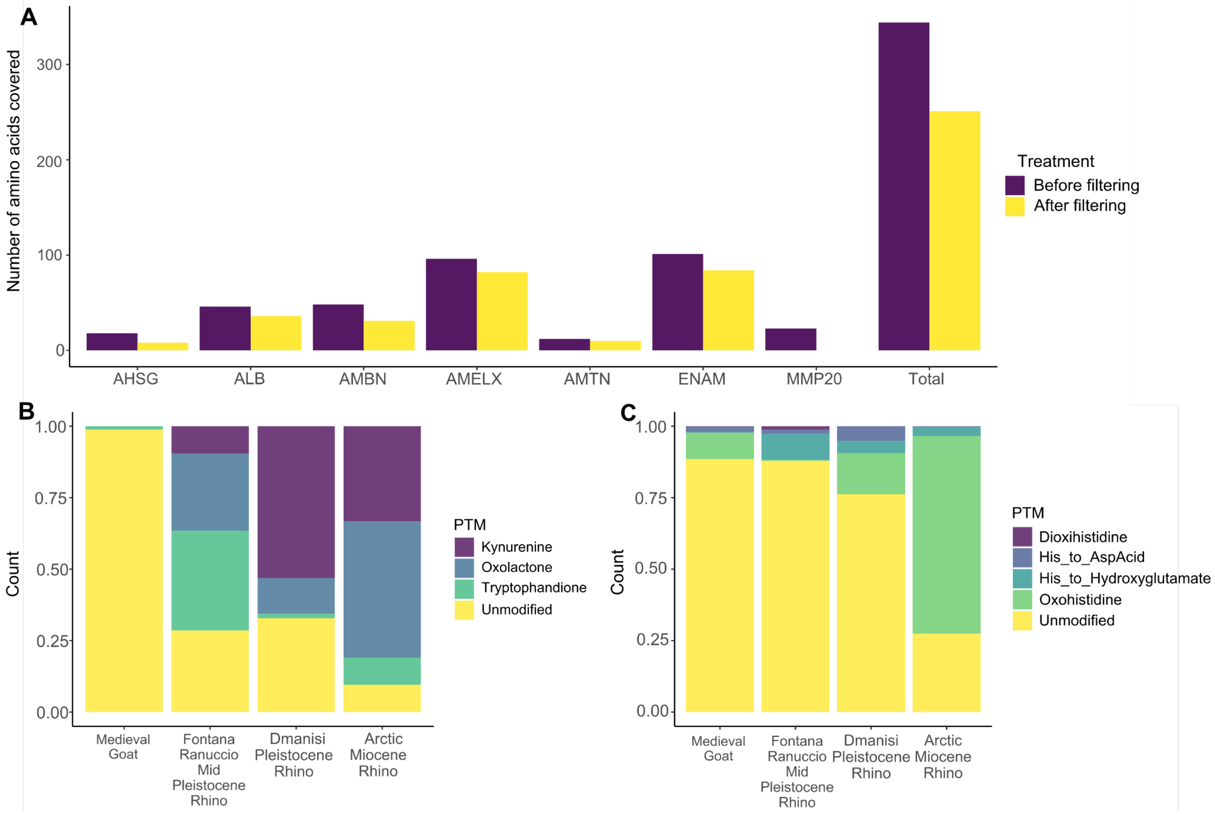
bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Competing interests: The authors declare no competing interests.

Corresponding authors: Correspondence to Ryan Sinclair Paterson, Danielle Fraser, or Enrico Cappellini.

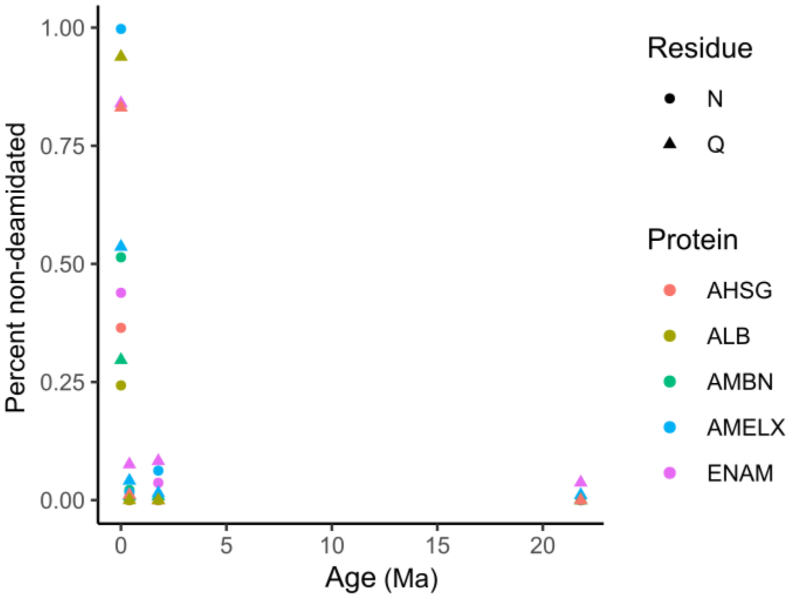
bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Extended Data



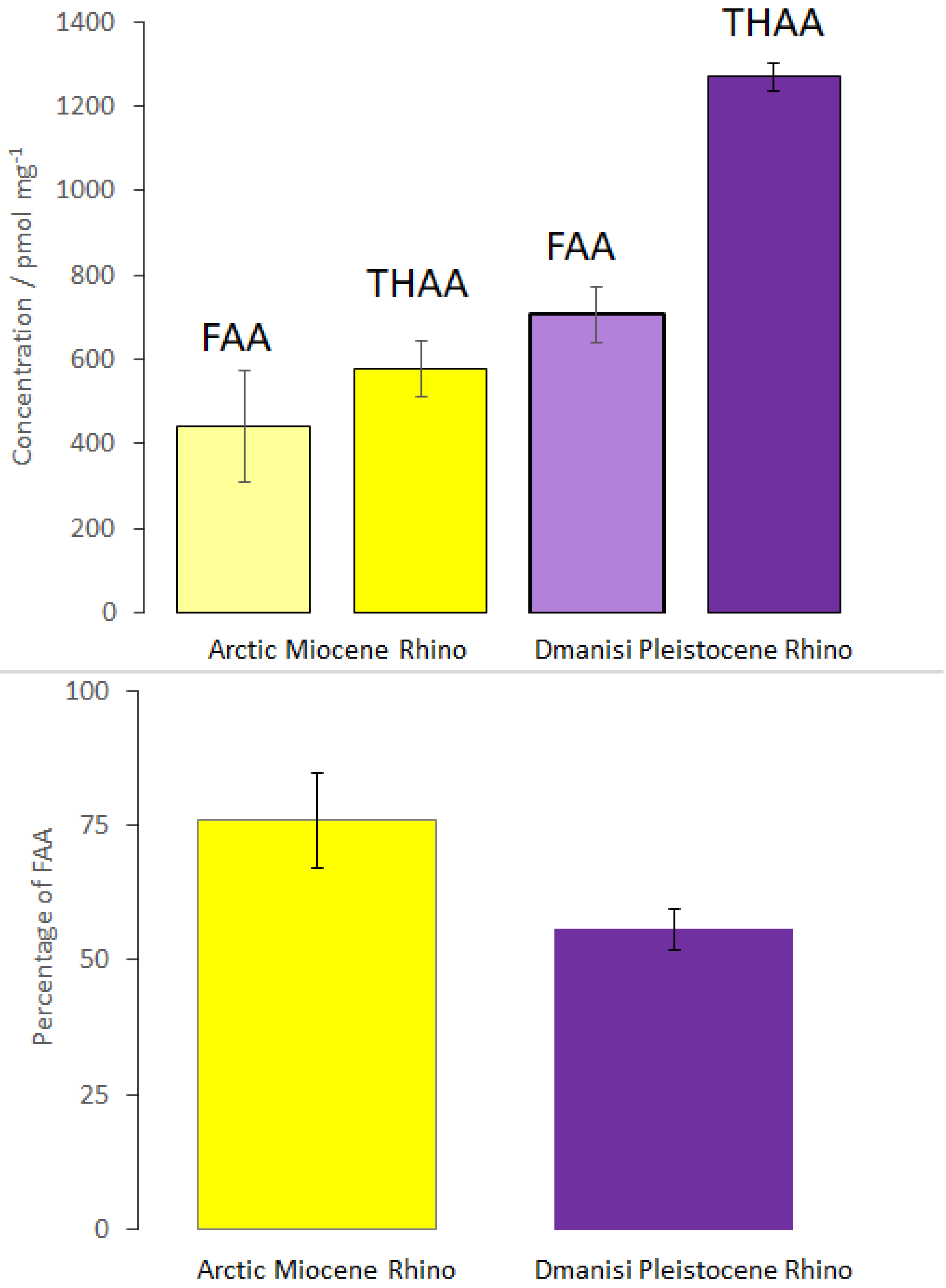
Extended Figure 1) Proteome preservation in CMNF-59632. A) Amino acid count for each identified protein, before and after filtering (see Methods), B) PTMs related to oxidative degradation of tryptophan for CMNF-59632, compared to enamel proteomes from other ancient rhinos and a mediaeval ovicaprine, C) PTMs related to oxidative degradation of histidine for same taxon set. The moderate protein preservation in CMNF-59632, indicated by the lower amino acid coverage compared to other ancient enamel proteomes, is further supported by the high incidence of PTMs related to oxidative degradation, compared to other fossil rhinocerotids.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



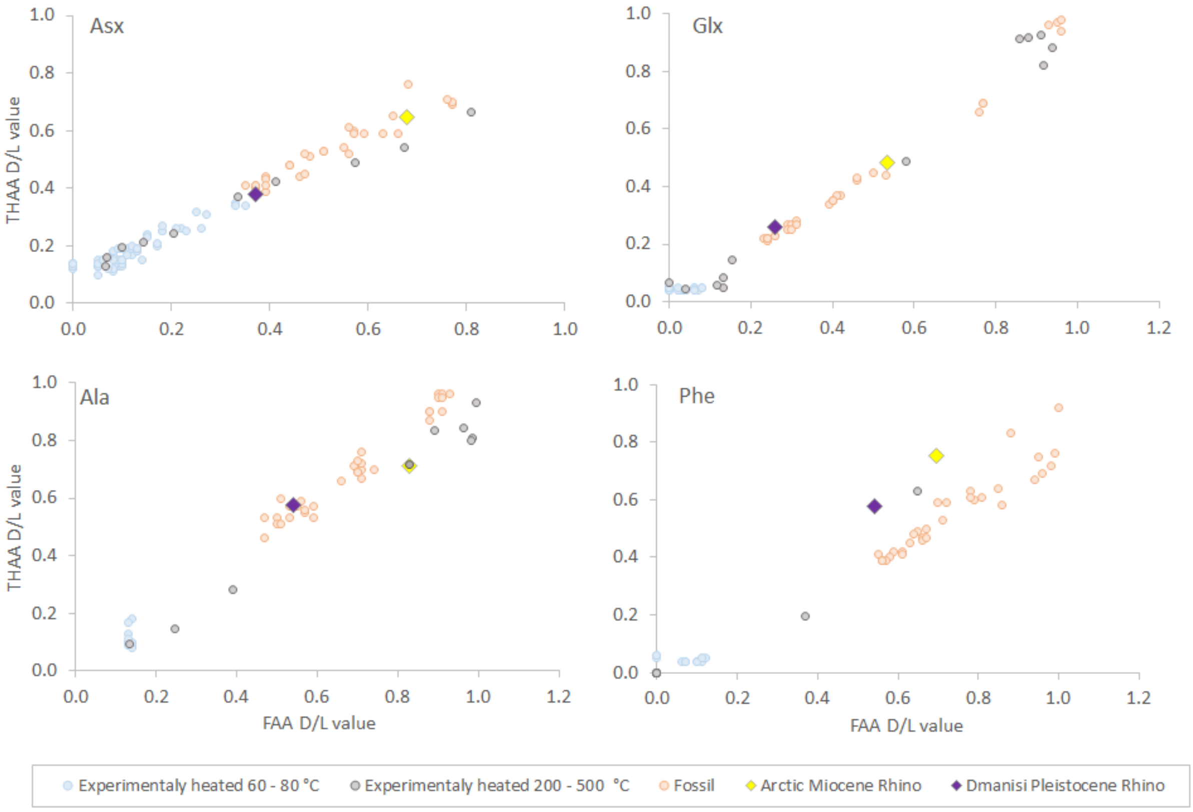
Extended Figure 2) Deamidation rates in fossil rhinocerotid enamel proteomes, plotted against geological age. Data used is CMFN-59632 from Haughton Crater (21.8 Ma), DM.5/157 from Dmanisi (1.77 Ma), CGG 1_023342 from Fontana Ranuccio (0.4 Ma), and a mediaeval control sample (0.005 Ma). While useful for 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.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



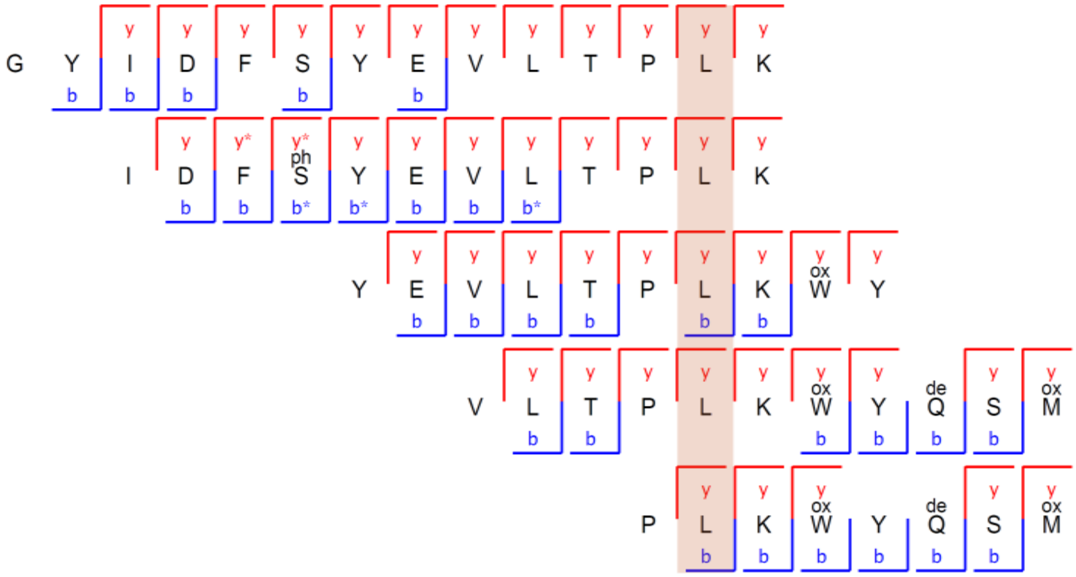
Extended Figure 3) Comparison of FAA & THAA concentration (top) and %FAA (bottom) of the Arctic Miocene rhinocerotid with the Dmanisi Pleistocene *Stephanorhinus*[11]. Error bars represent 1 standard deviation about the mean for preparative replicates. The lower overall concentration, higher %FAA and yet incomplete hydrolysis in the Arctic Miocene rhino is consistent with endogenous peptides in the tooth enamel.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Figure 4) 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. The good correlation between FAA & THAA for the Arctic Miocene rhino (CMNF-59632) sample supports the presence of closed system original peptides and their constituent amino acids in this Miocene sample.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.07.597871; this version posted June 9, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Figure 5) Subset of overlapping peptides supporting a SAP at AMELX position 53.

While all other later-diverging rhinocerotids, including *Elasmotherium*, display a valine (V) at position 53 (highlighted in light red; following numbering of reference sequence A0A5F5PLN8), CMNF-59632 displays a leucine (L) (or isoleucine (I)), representing the ancestral condition in Perissodactyla. The peptide sequences depicted here represent just a small portion of the peptide-spectrum matches covering this position. Together, these peptides display high Andromeda scores, extended, in some cases complete, ion series, and the presence of several PTMs supporting their endogeneity (phosphorylation) and ancientness (tryptophan oxidation, glutamine deamidation).