nature portfolio

https://doi.org/10.1038/s41586-025-09040-9

Supplementary information

Eighteen million years of diverse enamel proteomes from the East African Rift

In the format provided by the authors and unedited

Supplementary Information

18 million years of diverse enamel proteomes from the East African Rift

Authors

1. Daniel R. Green[1,2*]
2. Kevin T. Uno[1,2]
3. Ellen R. Miller[3]
4. Craig S. Feibel[4]
5. Eipa Emmanuel Aoron[1,5]
6. Cat Beck[6]
7. Aryeh Grossman[7]
8. Francis Kirera[8]
9. Martin M. Kirinya[9]
10. Louise Leakey[9,10]
11. Cynthia Liutkus-Pierce[11]
12. Fredrick K. Manthi[12]
13. Emmanuel K. Ndiema[13]
14. Isaiah Nengo[†9,10]
15. Cyprian Nyete[9]
16. John Rowan[14]
17. Gabrielle Russo[10]
18. William Sanders[15]
19. Tara Smiley[17]
20. Patricia Princehouse[16]
21. Natasha Vitek[17]
22. Timothy P. Cleland[18*]

Affiliations

1. Department of Human Evolutionary Biology, Harvard University, Cambridge MA 02138, USA

2. Division of Biology and Paleoenvironment, Lamont-Doherty Earth, Columbia University, Palisades NY 10964, USA3. Department of Anthropology, Wake Forest University, Winston-Salem, NC 27109, USA
4. Department of Earth and Planetary Sciences, Rutgers University, New Brunswick, NJ 08901, USA
5. Turkana University College, Lodwar, 3J88+5WP, Kenya

6. Department of Geosciences, Hamilton College, Clinton, NY 13323, USA
7. Department of Anatomy, Midwestern University, Glendale, AZ 85308, USA
8. Mercer School of Medicine, Mercer University, Macon, GA 31207, USA
9. Turkana Basin Institute, Nairobi 24467, Kenya
10. Department of Anthropology, Stony Brook University, Stony Brook, NY 11794, USA
11. Department of Geological and Environmental Sciences, Appalachian State University, Boone, NC, USA
12. Department of Earth Sciences, National Museums of Kenya, Nairobi, 40658, Kenya
13. Sites and Monuments, National Museums of Kenya, Nairobi, 40658, Kenya
14. Department of Archaeology, University of Cambridge, Cambridge CB2 3DZ, UK
15. Museum of Paleontology, University of Michigan, Ann Arbor, MI 48109, USA
16. Institute for the Science of Origins, Case Western Reserve University, Cleveland, OH 44106, USA
17. Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794, USA
18. Museum Conservation Institute, Smithsonian Institute, Suitland, MD 20746, USA

†Deceased

* Authors to whom correspondence should be addressed:

Timothy P. Cleland
Museum Conservation Institute, Office MSC E2110
Smithsonian Institution
4210 Silver Hill Rd. Suitland, MD 20746, USA
301-238-1208; clelandtp@si.edu

Daniel R. Green
Human Evolutionary Biology
Harvard University
11 Divinity Avenue, Cambridge, MA, 02138, USA
248-250-4495; drgreen@fas.harvard.edu

Kevin T. Uno
Human Evolutionary Biology
Harvard University
11 Divinity Avenue, Cambridge, MA, 02138, USA
801-503-5028; kevinuno@g.harvard.edu

SI Text

Enamel sample collection

We primarily sampled enamel from all specimens reported here. Potentially altered or exogenous material was avoided by removing the exterior 0.3–0.5 mm of tissue from all enamel surfaces prior to sampling the well-preserved enamel interior. While we made efforts to avoid sampling dentin, small quantities of dentin in our sample cannot be ruled out. Collagen fragments reported from enamel samples here may have derived from enamel or dentin, since collagens have been found in modern and fossil enamel using traditional and LC-MS/MS methods (Felszeghy 2000; Acil 2005; Jagr 2019; Cappellini 2019).

Koobi Fora

The Koobi Fora elaphantid specimen was collected from the surface of the A7 arenaceaous bioclastic carbonate layer capping western exposures of the Koobi Fora Tuff complex in Area 103, on the eastern shores of Lake Turkana in northern Kenya. The name Koobi Fora derives from the Gabbra people, and refers to “small hills near Darate” (koopi) and “cattle shade” (fora). Sediments in the complex include tuffaceous sand, silt, and mud deposits of lacustrine, deltaic, and fluvial origin, and preserve fossils and footprints of a variety of taxa, including hominins[1-3]. Immediately below A7, fossil diversity indicates a shift over time from a primarily terrestrial to an aquatic community, whereas collections on the A7 surface are mixed.

The complex is bounded by the KBS Tuff below (1.86 Ma) and the Chari Tuff (1.38 Ma) above but includes a series of smaller tuffs and pumice-rich tuffaceous deposits that can be more precisely dated[2,3]. These include the Lower Koobi Fora and the Koobi Fora tuffs, dated to 1.48 and 1.49 Ma, respectively[4,5]. The A7 layer is derived from algal stromatolites and is overlain elsewhere in 103 by the Akait Tuff, estimated to be approximately 1.40-1.45 Ma[4]. We therefore constrain the elephantid specimen as likely derived from sediments deposited between 1.48 and 1.40 Ma.

Lothagam

Sample KNM-LT 26247 was collected from sediments in the lower member of the Nawata Formation at the Lothagam fossil site in Turkana Basin, Kenya. Located approximately 55 km east-southeast of the town of Lodwar, Lothagam is an uplifted footwall block that exposes sediments deposited in the Kerio Basin from the Middle Miocene to the Pliocene[6,7]. The site name derives from the Turkana words Lo, which describes a place, and athagamam, which means rough and therefore describes the undulating topography of uplifted blocks that form two north-south trending ridges that define the area. Fossil-bearing sediments are exposed between the ridges and on their flanks. The eastern ridge is a horst called Lothagam Hill and is the more prominent of the two. From ~1989 to 1993, Meave Leakey's team recovered fossils and studied the geology at Lothagam[8]. The team determined precise radiometric ages throughout the ~900 m section[9].

The Nawata Formation is bounded at the lower, older end of its stratigraphic sequence by the volcaniclastic Nabwal Arangan beds. Whole rock basalt and phonolite K-Ar dates from Nabwal Arangan range from 14.2 ± 0.2 Ma at the base, to 9.1 ± 0.2 at the top[9]. The Nawata Formation is informally split into Lower and Upper Members. The base of the 137 m thick Lower Nawata consists of alluvial plain sediments and gravel conglomerates interpreted as ephemeral stream deposits[6]. The upper part of the Lower Nawata contains fluvial deposits characteristic of a perennial river system. In the member's lower portion, pumices derived from a series of altered tuffs called the Lower Markers, have an arithmetic mean $^{40}Ar/^{39}Ar$ age on alkali feldspar crystals of 7.44 ± 0.05 Ma, and represents the age of Lower Nawata fauna[9]. Most Lower Nawata fauna come from above this dated tuff. An informally named "Marker Tuff" signifies the base of the Upper Nawata, providing the only datable material from the Upper Nawata. Multiple single crystal alkali feldspars analyzed from the Marker Tuff yield an arithmetic mean $^{40}Ar/^{39}Ar$ age of 6.54 ± 0.04 Ma[9]. The 125 m thick Upper Nawata contains abundant thick, multi-storied sandstones, suggesting increased reworking of sediments resulting from reduced subsidence rates in the basin[6,10]. The Upper Nawata contains fewer paleosols than the Lower Nawata, despite similar overall thicknesses. The age of the hippopotamid specimen KNM-LT 26247 is therefore estimated to be ~ 7.4 to 6.5 Ma.

Napudet Red Series

Hippopotamid specimen KNM-NP 64551 was collected from the "Red Series" depositional unit at Napudet, meaning "river bend" in Turkana, on the southern and western margin of the Turkana Basin in northern Kenya. The age of Red Series sediments is poorly constrained, likely containing late Miocene and possibly early Pliocene sedimentary sequence composed of primarily oxidized red sandstone,

conglomerates and rare paleosols. Measured section, description, and stratigraphic contexts are provided in Nengo et al.[10]. Compared to the lower pyroclastic- and tuff-dominated Emunyan Beds of the Middle Miocene, the Red Series provide less evidence of volcanic input, with pumice beds found in the upper portion of the Red Series only, indicating a transition in depositional environment from a volcanic to a fluvial setting. The Red Series unconformably overlies the undated “Upper Basalt” at Napudet. A lower columnar basalt dated by $^{40}Ar/^{39}Ar$ to 13.31 +/- 0.04 Ma provides an absolute maximum age constraint[10]; however, based on stratigraphic context, the Red Series beds are likely much younger. Lonyumun Lake deposits, dated to 4.1 Ma, unconformably overlie the Red Series, providing the minimum age of the deposits. Based on the totality of available context we estimate KNM-NP 64551 to be 11–5 Ma.

Buluk

Proboscidean specimens KNM-WS 65217 (*Prodeinotherium hobleyi*), KNM-WS 65653 (*Zygolophodon* sp.), and KNM-WS 101824 (Gomphotheriidae indet.), were collected from the Early Miocene site of Buluk. Buluk refers to ‘monsters’ in the Daasanach language. The site is located between the northeastern end of Lake Turkana, northern Kenya, and the southwestern margin of lake Chew Bahir, Ethiopia. Fossiliferous deposits at Buluk outcrop as part of the uplift of the Suregai-Asille Plateau[11,12]. During the early Miocene, Buluk was dominated by a large, mature river system, and the majority of fossils are recovered from coarse sandstones and pebble conglomerate channel fills associated with this fluviatile complex[13,14]. A basalt outcropping stratigraphically above most of the fossil localities has been dated and helps provide a minimum age for the Buluk material.[15] The presence of some large, silicified tree trunks suggests at least a partially wooded environment, perhaps as gallery forest[13,16,17]. Paleoclimate proxies suggest a strongly seasonal subhumid to subarid climate[14]. Paleosol chemistry, bulk organic matter, and enamel isotope data indicate the presence of a C_3-dominated woodland to woody-grassland mosaic[14,18]. A recent study further suggests that C_4 grasses were locally abundant, possibly contributing to a heterogeneous wooded grassland environment[18]. The totality of available faunal and geological context provide an approximate age for proboscidean specimens analyzed here of ~16 Ma.

Locherangan

Anthracotheriid specimen KNM-LC 30762 (*Brachyodus aequatorialis*) and rhinocerotid specimen KNM-LC 17680 (Rhinocerotidae indet.) were collected from Locherangan, a

site on the western side of Lake Turkana, in the Turkana Basin of northern Kenya[19]. Locherangan means “red water” in the Turkana language. The stratigraphy at Locherangan includes 50.3 m of mudstones, shales, and siltstones, interspersed with sandstones, representing deposition in lacustrine and fluvial environments[20]. Silt, clay, and shale deposits include coprolites, ostracod, fish, turtle, and crocodile fossil remains, but mammalian fauna are found throughout the section in paleosols and fluvial sandstones. The faunal assemblage resembles those found at Kalodirr, Moruerot, and Buluk, also in Turkana and dated to ~17-16 Ma. Locherangan’s age is constrained by anorthoclase crystals from a bentonite in the middle of the section which provide a maximum $^{40}K/^{40}Ar$ age of 17.5 Ma for the overlying fossils[19].

Loperot

The Rhinocerotid specimen KNM-LP 30 was collected from the site of Loperot in the Turkana Basin, west and south of Lake Turkana. Loperot means “one road” in Turkana, and the site falls within Turkana's Lokichar Basin, which overlies Precambrian basement rock and is comprised of two geological members: the older Lokone Member deposited from the Paleogene to the Early Miocene, and the younger Auwerwer Member dated to the Middle Miocene[21]. Loperot exposes sediments from the upper portion of Lokone Member, which was formed by a mixture of fluvial, deltaic, and lacustrine environments. Basalts overlying fossiliferous strata were collected providing a mean $^{40}Ar/^{39}Ar$ age of 17.0 Ma[21]. Paleontological research at Loperot has been ongoing since the 1940s, revealing a rich fauna including both aquatic and terrestrial fauna[21]. Similarities between the Loperot fauna and those of nearby sites in eastern Africa help further constrain the likely age of Early Miocene fossiliferous deposits. Loperot fauna resemble those of the Hiwegi Formation of Rusinga Island in southwest Kenya, dated to 18 Ma or older[22]. Furthermore, Loperot likely shares an unusual cf. *Archaeobelodon* specimen with the 20–19 Ma site of Songhor in western Kenya. For these reasons, we estimate that the specimen KNM-LP 30 is~18 Ma.

Topernawi

The *Arsinoitherium* specimen KNM-TP 102134 was collected at Topernawi, a site on the western side of Lake Turkana, meaning “home sleep” in Turkana. Fossil-bearing sediments are pyroclastic in origin and occur within the Topernawi Formation, which contains 92 m of deposition overall. The formation is exposed in the smaller Ekitale Basin[23]. Paleontological research by the Topernawi Research Project has revealed a diverse array of hyracoids, with five species representing four genera ranging from ~8 to

150 kg. Vertebrate fossils recovered span a taxonomic and body size range from a small frog (Anura) to multiple proboscideans and the embrithopod *Arsinotherium perhaps* weighing 3000 kg. The fauna also include Rodentia (including anomaluroids[24]), Primates, Anthracotheriidae, Ptolemaiidae, and Hyaenodonta, among others. At least 7 species of plants using the C_3 photosynthetic pathway have been identified, including fossilized fruit from two species, suggesting a mosaic of closed canopy and seasonal environments. The fossils are overlain by ignimbrites dated by $^{40}Ar/^{39}Ar$ to 29.24 ± 0.08 Ma, while basalts below the deposits have been given an $^{40}Ar/^{39}Ar$ age of 29.5 ± 0.5 Ma. Therefore, the age of KNM-TP 102134 is estimated to be 30–29.2 Ma[25].

Data dependent acquisition

In addition to the DIA data presented, data dependent acquisition (DDA) was also attempted with the following parameters: MS1: 120000 resolving power, 1000000 AGC target, 10 ms max inject time, MS2: 15000 resolving power, top 10 most intense ions fragmented with 30% NCE HCD with a 2.0 m/z isolation width, 500000 AGC target, 100 ms max inject time, 1+ charges were included. An initial MetaMorpheus search was applied against the Uniprot enamel genes database and no matches were found that met the strict score of 9 cutoff. We repeated the TFA extraction of the Koobi Fora *Palaeoloxodon recki* sample and applied the above DDA approach (with the exception of increased max inject times of 100 ms (MS1) and 250 ms (MS2) and a modified instrument method. In the modified method, the nanoLC was run at 100 nL/min instead of the original 300 nL/min with the following gradient: 0−6 min 2% B, 6−68 min 2−31% B, 68−70 min 31−90% B, 70−73 min 90% B, 73−74 min 90−2% B, 74−120 min 2%B. Solvent A is 0.1% formic acid in water and solvent B is 0.1% formic acid in acetonitrile.. The max inject time for MS1 was increased to 100 ms and the MS2 was increased to 250 ms. 1+ was excluded, but all other parameters were kept identical. Using the original gradient and flow rate, no DDA peptides were identified for searches against the Afrotheria enamel genes database or Uniprot enamel genes databases, but with the lower flow rate and modified gradient, one diagenetiform (EVLTPLK; 9 PSMs with 3 above the strict 9 score cutoff; AMELX) was identified with the Afrotheria enamel genes database (Figure S11). The identical diagenetiform was detected using the DIA data. The limited repeat supports the use of DIA to maximize the number of identified diagenetiforms from these samples from low concentration samples.

Database generation

Protein databases were downloaded from NCBI or uniprot with the following genes for enamel genes (ambn, amelx, amely, amtn, dmp1, dspp, enam, fam20c, klk4, mmp20) or collagen genes (col1a1, col1a2, col3a1). Databases were restricted by taxonomy to Afrotheria, Perissodactyla, and Whippomorpha. The rhinoceratid enamel database was provided by Enrico Cappellini and Jazmin Madrigal. Individual PSM databases were created in Notepad++ and are available in the GPTMD folders on MassIVE ftp://massive.ucsd.edu/MSV000091838/ and ftp://MSV000093266@massive.ucsd.edu.

Combining results from multiple databases

A total of 406 peptide spectral matches (PSMs) were characterized. Of those 406, a total of 82 PSMs had multiple sequences for individual scans. Five of the 82 scans had the same sequences except for ambiguity between leucine and isoleucine. 43 out of the 82 had results from separate but overlapping databases (e.g., proboscidean enamel genes and mammal enamel genes). Other PSMs for single scans came from non-overlapping databases (e.g., FAM20C database and collagen genes). 282 identifications overlapped regardless of database supporting their combining across databases.

Results filtering and peptide confidence assessment

All PSM files were concatenated in RStudio and filtered at 1% false discovery rate (FDR), a strict score cutoff of 9, and filtered for any diagenetiforms/peptides detected in the buffer control such as keratins, resulting in a total of 406 PSMs and resulting sequences. Buffer control peptides included 110 PSMs and resulting sequences. All diagenetiform and peptide parameters and figures were calculated/generated with a custom R script in RStudio 2023.3.0 Build 386.

The 406 sample-derived peptides passing the FDR threshold were categorized using several criteria. Those sequences whose top match in a MetaMorpheus search was to a clade broadly including the fossil were classified as “clade match.” In this category, proboscideans and arsinoithere-derived sequences matched to Afrotheria, hippopotamid and anthracothere-derived sequences to Whippomorpha, and rhinocerotid sequences to Perissodactyla. Clade match PSMs totaled 154 peptides. Those clade match sequences that also matched to the same broadly inclusive clades in a second search, conducted manually with NCBI’s Blastp, were categorized as “putative” diagenetiforms. A total of 133 PSMs met this threshold. Putative sequences were further classified as “high confidence” if they were not collagens and met at least

one of two additional criteria: sequence specificity within Mammalia, and few or no matches to microbial products. Altogether, 19 PSMs met this threshold (Table S3; Table S4; SI Spectra).

Of those sequences classified as high confidence, classifications sometimes depend more strongly on certain filtering criteria and not on others.

For instance QNTESKSSE sampled from the enamel of three proboscideans (one from Koobi Fora and four from two fossils at Buluk) match to DMP1 among extant proboscideans *Loxodonta africana* and *Elephas maximus*, but not to DMP1 in other afrotherians. The sequence does match DMP1 in many extant mammals outside Afrotheria, but matches no known microbial products. Its uniqueness among afrotherians and dissimilarity to known microbial peptides lead us to assess that it is ancient and thus a fossil-derived diagenetiform despite its similarity to other mammalian DMP1 sequences.

The sequence PAGHSRGPP discovered from the 18 Ma fossil rhinocerotid sample matches to AMBN in extant rhinos and tapirs but is only a c. 90 % match to AMBN in equids and other mammals, making its sequence relatively diagnostic among mammals. It furthermore does not match to any known microbial products, leading us to classify it as a diagenetiform with high confidence. The sequences EVVSEPR and SLQSKNAP derived from the same 18 Ma fossil rhinocerotid sample match to DMP1 and ENAM in extant rhinos, but not to other perissodactyls or mammals. The sequence EVVSEPR matches a single turtle collagen, and both EVVSEPR and SLQSKNAP match many bacterial products. Despite its similarity to microbial products, given their specificity to rhinocerotid DMP1 and ENAM compared to all available mammalian sequences, we assess these as diagenetiforms with high confidence.

While we note that some of the results presented here pass the traditional 2-peptide rule (Fig. 2, Table S4, and Fig. S2), not all do, and thus we rely on the filtration and confidence assessment described above. To illustrate the relatively low probability of preserving two peptides from a single protein into the middle or early Miocene, we fit a simple decay constant model to our putative and high confidence PSMs to better understand peptide fragment loss over time (Fig. S13). Note that in the model, the preservation of two peptides can only be inferred; the model shows single (blue) or two-plus (green) PSM counts and their relative frequencies, for a given single, generalized peptide, over time.

Phylogenetic analysis

A phylogenetic analysis was performed with Geneious Prime 2023.1.1 Build 2023-04-03. In an initial analysis conducted using high confidence diagenetiforms, undigested diagenetiforms from AMBN, ENAM, DMP1, and MMP20 were aligned to sequences from NCBI using MAFFT v. 7.490 with the BLOSUM62 scoring matrix and the gap opening penalty set to 3. Erroneous alignments (i.e., N-terminal diagenetiform lysine being aligned to the N-terminus of the protein instead of adjacent to the rest of the diagenetiform) were manually corrected. Concatenated alignments of all were then analyzed using PhyML 3.3.20180621 with the BLOSUM62 substitution model. The resultant tree was rooted against *Canis lupus*. In a second analysis this procedure was repeated using both high confidence and putative diagenetiforms matched to AMBN, AMELX, AMTN, COL1A1, COL3A1, DMP1, ENAM, and MMP20. Final trees were exported to FigTree v. 1.4.4.

Linear Mixed Models (LMM)

Restricting our analysis to high confidence diagenetiforms and excluding SP3 results that derive only from one specimen, we test the effect of age on counts per specimen using an LMM (Python3 statsmodels mixedlm), treating age as a fixed effect, and fossil specimens as a random effect, in test 1. We evaluate solutions using a restricted maximum likelihood approach (REML). We find a decline in counts over time (slope of -0.081 counts / Ma) that is not significant ($p = 0.509$), and no effect of mass-adjusted counts over time (test result not shown). In addition to the results presented in the text, we report information from the test below:

Test 1: Impact of age on protein counts, age = fixed effect, individual = random effect
Sample set: High confidence diagenetiforms (n = 14, SP3 excluded)

Intercept: Coeff. = 3.779 (SE 1.175), $z = 3.215$, $p = 0.001$
Age: Coeff. = -0.081 (SE 0.122), $z = -0.661$, $p = 0.509$
Group Var: Coeff. = 1.892

Analyzing putative diagenetiforms and again excluding SP3 results, we test the effect of age on counts per specimen, treating age as a fixed effect, and fossil specimens as a random effect, in tests 2–3. We find a decline in counts over time (slope of -0.22 counts / Ma) and in mass-adjusted counts over time (-0.043 counts / Mg enamel / Ma), both of which are highly significant ($p < 0.001$). We report information from these tests below:

Test 2: Impact of age on protein counts, age = fixed effect, individual = random effect
Sample set: Putative diagenetiforms (n = 108, SP3 excluded)

Intercept: Coeff. = 50.895 (SE 3.400), z = 14.967, p < 0.001
Age: Coeff. = -2.331 (SE 0.584), z = -3.989, p < 0.001
Group Var: Coeff. = 184.610

Test 3: Impact of age on mass-adjusted protein counts, age = fixed effect, individual = random effect
Sample set: Putative diagenetiforms (n = 108, SP3 excluded)

Intercept: Coeff. = 1.025 (SE 0.002), z = 440.919, p < 0.001
Age: Coeff. = -0.043 (SE 0.005), z = -9.343, p < 0.001
Group Var: Coeff. = 0.021

Using the entire dataset and restricting only to peptides meeting the FDR threshold, but excluding SP3 results that derive only from one specimen, we test the effect of age on counts per protein, counts per fossil specimen, and on diagenetiform length across all sites in tests 4–6. We treat age as a fixed effect, and sites as a random effect. In further tests 7–8, we find that sample mass influences PSM counts, and that mass-adjusted PSM counts decline with age. We report information from these tests below:

Test 4: Impact of age on protein counts, age = fixed effect, site = random effect
Sample set: All FDR threshold peptides (n = 222, SP3 excluded)

Intercept: Coeff. = 5.933 (SE 1.155), z = 5.135, p = 0.000
Age: Coeff. = -0.228 (SE 0.083), z = -2.738, p = 0.006
Group Var: Coeff. = 0.000

Test 5: Impact of age on fossil specimen counts, age = fixed effect, site = random effect
Sample set: All FDR threshold peptides (n = 222, SP3 excluded)

Intercept: Coeff. = 62.952 (SE 18.790), z = 3.350, p = 0.001
Age: Coeff. = -2.688 (SE 1.152), z = -2.333, p = 0.020
Group Var: Coeff. = 585.638 (SE 79.214)

Test 6: Impact of age on diagenetiform lengths, age = fixed effect, site = random effect
Sample set: All FDR threshold peptides (n = 222, SP3 excluded)

Intercept: Coeff. = 9.218 (SE 0.825), z = 11.173, p = 0.000

Age: Coeff. = -0.008 (SE 0.054), z = -0.149, p = 0.881
Group Var: Coeff. = 1.077 (SE 0.480)

Test 7: Impact of sample mass on fossil specimen counts, mass = fixed effect, individual = random effect
Sample set: All FDR threshold peptides (n = 222, SP3 excluded)

Intercept: Coeff. = -14.531 (SE 7.448), z = -1.951, p = 0.051
mass: Coeff. = 1.211 (SE 0.008), z = 145.694, p < 0.001
Group Var: Coeff. = 258.959

Test 8: Impact of sample age on mass-adjusted specimen counts, age = fixed effect, site = random effect
Sample set: All FDR threshold peptides (n = 222, SP3 excluded)

Intercept: Coeff. = 1.584 (SE 0.471), z = 3.360, p = 0.001
age: Coeff. = -0.059 (SE 0.029), z = -2.059, p = 0.039
Group Var: Coeff. = 0.312 (SE 1.431)

References

1 Behrensmeyer, A. K. & Laporte, L. F. Footprints of a Pleistocene Hominid in Northern Kenya. *Nature* **289**, 167-169 (1981). https://doi.org:DOI 10.1038/289167a0

2 Brown, F. H. & Feibel, C. S. Stratigraphical Notes on the Okote-Tuff-Complex at Koobi-Fora, Kenya. *Nature* **316**, 794-797 (1985). https://doi.org:DOI 10.1038/316794a0

3 Mcdougall, I., Davies, T., Maier, R. & Rudowski, R. Age of the Okote Tuff Complex at Koobi-Fora, Kenya. *Nature* **316**, 792-794 (1985). https://doi.org:DOI 10.1038/316792a0

4 Brown, F. H., Haileab, B. & McDougall, I. Sequence of tuffs between the KBS Tuff and the Chari Tuff in the Turkana Basin, Kenya and Ethiopia. *J Geol Soc London* **163**, 185-204 (2006). https://doi.org:Doi 10.1144/0016-764904-165

5 Brown, F. H. & McDougall, I. Geochronology of the Turkana depression of northern Kenya and southern Ethiopia. *Evol Anthropol* **20**, 217-227 (2011). https://doi.org:10.1002/evan.20318

6 Feibel, C. S. in *Lothagam* (eds G. Leakey Meave & M. Harris John) 17-30 (Columbia University Press, 2003).
7 Schofield, N. *et al.* Linking surface and subsurface volcanic stratigraphy in the Turkana Depression of the East African Rift system. *J Geol Soc London* **178** (2021). https://doi.org:ARTN jgs2020-110 10.1144/jgs2020-110
8 Leakey, M. G. *et al.* Lothagam: A record of faunal change in the late Miocene of east Africa. *Journal of Vertebrate Paleontology* **16**, 556-570 (1996). https://doi.org:Doi 10.1080/02724634.1996.10011339
9 McDougall, I. & Feibel, C. S. Numerical age control for the Miocene-Pliocene succession at Lothagam, a hominoid-bearing sequence in the northern Kenya Rift. *J Geol Soc London* **156**, 731-745 (1999). https://doi.org:DOI 10.1144/gsjgs.156.4.0731
10 Nengo, I. *et al.* New infant cranium from the African Miocene sheds light on ape evolution. *Nature* **548**, 169-174 (2017). https://doi.org:10.1038/nature23456
11 Harris, J. M. & Watkins, R. New Early Miocene Vertebrate Locality near Lake Rudolf, Kenya. *Nature* **252**, 576-577 (1974). https://doi.org:DOI 10.1038/252576a0
12 Watkins, R. T. *The geology of the Suregei-Asille district and the upper Bakate Valley, northern Kenya*, Birkbeck (University of London), (1983).
13 Watkins, R. The Buluk Member, a fossil hominoid-bearing sedimentarysequence of Miocene age from northern Kenya. *Journal of African Earth Sciences (and the Middle East)* **8**, 107-112 (1989).
14 Lukens, W. E. *et al.* Paleoenvironments and mammalian fauna of the early Miocene fossil site at Buluk, Kenya. *Am J Phys Anthropol* **162**, 268-269 (2017).
15 Peppe, D. *et al.*, Geochronology and terrestrial paleoenvironments of the Middle Miocene hominoid-bearing strata at Buluk, Turkana Basin, Kenya, Geological Society of America Abstracts with Programs. Vol. 56, No. 5, 10.1130/abs/2024AM-404038.16 (2024). Wheeler, E. A., Wiemann, M. C. & Fleagle, J. G. Woods from the Miocene Bakate Formation, Ethiopia Anatomical characteristics, estimates of original specific gravity and ecological inferences. *Rev Palaeobot Palyno* **146**, 193-207 (2007). https://doi.org:10.1016/j.revpalbo.2007.04.002
17 Leakey, M., Grossman, A., Gutierrez, M. & Fleagle, J. G. Faunal change in the Turkana Basin during the late Oligocene and Miocene. *Evol Anthropol* **20**, 238-253 (2011). https://doi.org:10.1002/evan.20338
18 Peppe, D. J. *et al.* Oldest evidence of abundant C(4) grasses and habitat heterogeneity in eastern Africa. *Science* **380**, 173-177 (2023). https://doi.org:10.1126/science.abq2834
19 Anyonge, W. Fauna from a New Lower Miocene Locality Wets of Lake Turkana, Kenya. *Journal of Vertebrate Paleontology* **11**, 378-390 (1991).
20 Beck, C., Langworthy, M., Feibel, C., Beverly, E. & Horgan, J. *LOWER MIOCENE PALEOENVIRONMENT AND PALEOCLIMATE AS RECORDED BY THE LOCHERANGAN FORMATION, TURKANA BASIN, KENYA*. (2017).
21 Liutkus-Pierce, C. M. *et al.* Reconstruction of the Early Miocene Critical Zone at Loperot, Southwestern Turkana, Kenya. *Front Ecol Evol* **7** (2019). https://doi.org:ARTN 44

10.3389/fevo.2019.00044
22 Grossman, A., Liutkus-Pierce, C., Kyongo, B. & M'Kirera, F. New Fauna from Loperot Contributes to the Understanding of Early Miocene Catarrhine Communities. *Int J Primatol* **35**, 1253-1274 (2014). https://doi.org:10.1007/s10764-014-9799-8
23 Ragon, T. *et al.* Evolution of the northern Turkana Depression (East African Rift System, Kenya) during the Cenozoic rifting: New insights from the Ekitale Basin (28-25.5 Ma). *Geol J* **54**, 3468-3488 (2019). https://doi.org:10.1002/gj.3339
24 Sousa, F. J. *et al.* New Discovery of Oligocene Strata in the Topernawi Formation, Turkana County, Kenya. *Front Earth Sc-Switz* **10** (2022). https://doi.org:ARTN 799097
10.3389/feart.2022.799097
25 Seiffert, E. R. *et al.* Oldest record of a crown anomaluroid rodent from sub-Saharan Africa: a new genus and species from the early Oligocene Topernawi formation of northern Kenya. *Hist. Biol.* 1-11 (2024).

SI Tables and Figures

Table S1:

Enamel samples and masses from Kenyan sites.

Prefix	Number	Family or Order; species	Mass (mg)	Site	Age	Epoch
fdrg103	13a	Elephantidae; *Palaeoloxodon recki*	127.2	Area 103	1.5	Pleistocene
KNM-LT	26247	Hippopotamidae; *Archaeopotamus sp.*	16.7	Lothagam	c. 7	Miocene
KNM-NP	64551	Hippopotamidae; Hippopotamidae indet.	37.4	Napudet	c. 5-11	Miocene
KNM-WS	65217	Deinotheriidae; *Prodeinotherium hobleyi*	24.9	Buluk	c. 16	Miocene
KNM-WS	65653	Mammutidae; *Zygolophodon sp.*	43.9	Buluk	c. 16	Miocene
KNM-WS	101824	Gomphotheriidae; Gomphotheriidae indet.	34.5	Buluk	c. 16	Miocene
KNM-LC	17680	Rhinocerotidae; Rhinocerotidae indet.	9.4	Locherengan	c. 17	Miocene
KNM-LC	30762	Anthracotheriidae; *Brachyodus aequatorialis*	12.3	Locherengan	c. 17	Miocene
KNM-LP	30	Rhinocerotidae; cf. *Chilotheridium pattersoni*	23.2	Loperot	c. 18	Miocene
KNM-TP	102134	Arsinoitheriidae; *Arsinoitherium sp.*	37.5	Topernawi	c. 29	Oligocene

Figure S1:

The process of buffer control, DIA, and DDA analyses for our samples.

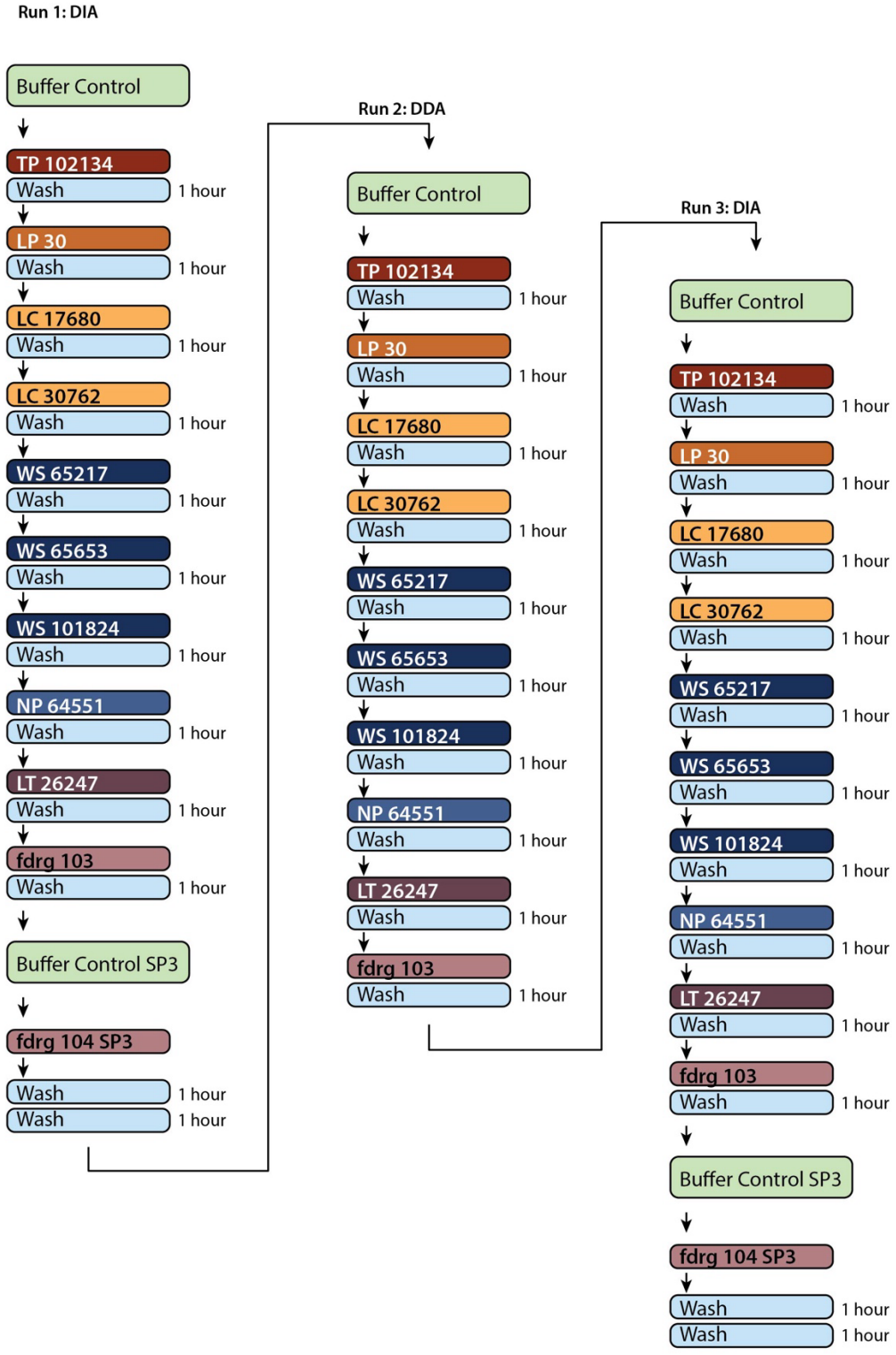

Table S2:
Databases used for MetaMorpheus searching. DIA represents the databases used for the first extraction and DIA2 represents the databases used for the second extraction of the enamel. All Sample PSM (e.g., 1.5Ma PSM) databases are targeted sequences derived from broader searches.

Database	DIA	DIA2
1.5Ma PSM	+	
13Ma PSM	+	
16Ma 101824	+	
16Ma 101824 -17Ma PSM -13Ma PSM	+	
16Ma 101824 synthetic	+	
16Ma 65217 PSM	+	
16Ma 65653 PSM	+	
17Ma 30762 PSM	+	
17Ma PSM	+	
17Ma PSM -13Ma PSM	+	
18Ma PSM	+	
29Ma PSM	+	
7Ma PSM	+	
Ameloblastin 16Ma 101824 aligned	+	
MetaMorpheus Contaminants	+	+
ncbi_hippopotamidae	+	+
ncbi_hippopotamidae_13Ma subset-uniprot-enamel_13Ma subset	+	
ncbi_hippopotamidae_17Ma subset-uniprot-enamel_17Ma hippo subset	+	
probosidea_enamel	+	+
probosidea_enamel_16Ma 101824 subset-uniprot-enamel_16Ma 101824 subset	+	
probosidea_enamel_16Ma 65217 subset-uniprot-enamel_16Ma 65217 subset	+	
probosidea_enamel_16Ma 65653 subset-uniprot-enamel_16Ma 65653 subset	+	
probosidea_enamel_29Ma subset-uniprot-enamel_29Ma subset	+	
probosidea_enamel_7Ma subset-uniprot-enamel_7Ma subset	+	
probosidea_enamel_subsetallsamples-uniprot-enamel_subsetallprobsamples	+	
probosidea_enamel-uniprot-enamel	+	

probosidea_enamel-uniprot-enamel-Rhinocerotidae_16635-ncbi_hippopotamidae	+	
Rhinocerotidae_16635		+
Rhinocerotidae_16635_17Ma 17680 subset-uniprot-enamel_17Ma 17680 subset	+	
Rhinocerotidae_16635_18Ma subset subset-uniprot-enamel_18Ma subset subset	+	
Rhinocerotidae_16635-uniprot-enamel	+	
test	+	
uniprot-Afrotheria	+	
uniprot-Afrotheria coll	+	+
uniprot-Afrotheria coll-uniprot-periscoll-uniprot-whippocoll		+
uniprot-afrotheria enam	+	+
uniprot-colgenes	+	+
uniprot-colgenes-uniprot-enamgenes	+	
uniprot-enamel		+
uniprot-enamel_1.5Masubset	+	
uniprot-enamel_1.5Masubset-probosidea_enamel_1.5Masubset	+	
uniprot-enamel-ncbi_hippopotamidae	+	
uniprot-enamel-probosidea_enamel	+	
uniprot-enamel-Rhinocerotidae_16635	+	
uniprot-enamel-uniprot-enamel-Rhinocerotidae_16635	+	
uniprot-enamel-uniprot-enamgenes	+	
uniprot-enamgenes	+	+
uniprot-fam	+	+
uniprot-mammalenam		+
uniprot-periscoll	+	+
uniprot-perissenam	+	+
uniprot-whippocoll	+	+
uniprot-whippoenam	+	+
UP000248480_127582manatee	+	

Uniprot-enamgenes derives from gene:amtn OR gene:amelx OR gene:amely OR gene:ambn OR gene:enam OR gene:dspp OR gene:mmp20 OR gene:klk4 OR gene:dmp1
Uniprot-colgenes derives from gene:col1a1 OR gene:col1a2 OR gene:col3a1

Table S3:
All Clade Match Peptides. Peptides all meet the clade match classification, but some are additionally considered to be putative diagenetiforms (yellow), meaning that their ancient provenance cannot be easily ruled out, or high confidence diagenetiforms, where some (green) are considered slightly higher confidence than others (blue).

Base Sequence	Protein	Accession	Age (Ma)	Classification
DGCSRKTNE	Collagen I alpha 2	KNM-TP102134	29	Putative
EVVSEPR	DMP1	KNM-LP30	18	High confidence
IDVAPLDI	Collagen I alpha 1	KNM-LP30	18	Putative
IDVAPLDI	Collagen I alpha 1	KNM-LP30	18	Putative
PAGHSRGPP	Ameloblastin	KNM-LP30	18	High confidence
SLQSKNAP	Enamelin	KNM-LP30	18	High confidence
EPGPAGLPGAQGL	Collagen III alpha 1	KNM-WS101824	17	Putative
EPGPAGLPGAQGL	Collagen III alpha 1	KNM-WS101824	17	Putative
KQAARPG	Enamelin	KNM-WS101824	17	Putative
KQAARPG	Enamelin	KNM-WS101824	17	Putative
QNTESKSSE	DMP1	KNM-WS101824	17	High confidence
QNTESKSSE	DMP1	KNM-WS101824	17	High confidence
SPPTVNVSN	Enamelin	KNM-WS101824	17	High confidence
GPSGPPGAPGAP	Collagen I alpha 1	KNM-WS65217	17	Putative
GSPGSNGSP	Collagen III alpha 1	KNM-WS65217	17	Putative
KGDAGPAGP	Collagen I alpha 1	KNM-WS65217	17	Clade match
KGHNGIDGI	Collagen I alpha 2	KNM-WS65217	17	Clade match
KGPYFSR	Enamelin	KNM-WS65217	17	Putative
QNTESKSSE	DMP1	KNM-WS65217	17	High confidence
QNTESKSSE	DMP1	KNM-WS65217	17	High confidence
SSKLATHLFSGT	Enamelin	KNM-WS65217	17	High confidence
TGPAGPAGPAGP	Collagen I alpha 1	KNM-WS65217	17	Putative
TVPHGPPQ	MMP20	KNM-WS65217	17	High confidence
SQEGADSP	DMP1	KNM-WS65653	17	Putative
SESRDFASE	DMP1	KNM-NP64551	5-11	Clade match
AAAADPLMTPG	Ameloblastin	KNM-LT26247	7	Putative
AAAADPLMTPG	Ameloblastin	KNM-LT26247	7	Putative
AAAADPLMTPG	Ameloblastin	KNM-LT26247	7	Putative
ARGAAGPQ	Collagen I alpha 1	KNM-LT26247	7	Putative
ARGAAGPQ	Collagen I alpha 1	KNM-LT26247	7	Putative
ARGAAGPQ	Collagen I alpha 1	KNM-LT26247	7	Putative
ARGAAGPQ	Collagen I alpha 1	KNM-LT26247	7	Putative
ENKAQQP	Ameloblastin	KNM-LT26247	7	Clade match

LSAGQAGANP	Amelotin	KNM-LT26247	7	Clade match
LSAGQAGANPA	Amelotin	KNM-LT26247	7	Putative
LSAGQAGANPA	Amelotin	KNM-LT26247	7	Putative
LSAGQAGANPA	Amelotin	KNM-LT26247	7	Putative
TQPNAPNPRG	Enamelin	KNM-LT26247	7	High confidence
AGNGPGPNP	Enamelin	fdrg103	1.5	High confidence
ATNAGSFWGP	Other	fdrg103	1.5	Clade match
EEQANEEP	DMP1	fdrg103	1.5	Putative
EGPFEED	Enamelin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
EVLTPLK	Amelogenin	fdrg103	1.5	Putative
KGTNLQD	Enamelin	fdrg103	1.5	High confidence
LPPHPGHPG	Amelogenin	fdrg103	1.5	Putative
LPPQPPL	Amelogenin	fdrg103	1.5	Clade match
LPPQPPL	Amelogenin	fdrg103	1.5	Clade match
LPQPPLP	Amelogenin	fdrg103	1.5	Clade match
LPQPPLP	Amelogenin	fdrg103	1.5	Clade match
QNTESKSSE	DMP1	fdrg103	1.5	High confidence
QPARAFPP	Enamelin	fdrg103	1.5	High confidence
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative

QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQP	Amelogenin	fdrg103	1.5	Putative
QPLLPQPP	Amelogenin	fdrg103	1.5	Clade match
QPLLPQPP	Amelogenin	fdrg103	1.5	Clade match
QPLLPQPP	Amelogenin	fdrg103	1.5	Clade match
QPLLPQPP	Amelogenin	fdrg103	1.5	Clade match
QPLLPQPP	Amelogenin	fdrg103	1.5	Clade match
QPLLPQPP	Amelogenin	fdrg103	1.5	Clade match
QPLLPQPP	Amelogenin	fdrg103	1.5	Clade match
QPLLPQPP	Amelogenin	fdrg103	1.5	Clade match
QPLLPQPP	Amelogenin	fdrg103	1.5	Clade match
QPLLPQPP	Amelogenin	fdrg103	1.5	Clade match
QPQPPVHP	Amelogenin	fdrg103	1.5	Putative
QPQPPVHP	Amelogenin	fdrg103	1.5	Putative
QPQPPVHP	Amelogenin	fdrg103	1.5	Putative
QPVQPQPP	Amelogenin	fdrg103	1.5	Putative
QPVQPQPP	Amelogenin	fdrg103	1.5	Putative
QPVQPQPP	Amelogenin	fdrg103	1.5	Putative
QPVQPQPP	Amelogenin	fdrg103	1.5	Putative
QPVQPQPP	Amelogenin	fdrg103	1.5	Putative

QPVQPQPP	Amelogenin	fdrg103	1.5	Putative
QPVQPQPP	Amelogenin	fdrg103	1.5	Putative
QPVQPQPP	Amelogenin	fdrg103	1.5	Putative
TKREEVD	Amelogenin	fdrg103	1.5	Putative
TKREEVD	Amelogenin	fdrg103	1.5	Putative
TKREEVD	Amelogenin	fdrg103	1.5	Putative
TKREEVD	Amelogenin	fdrg103	1.5	Putative
TPPYDSPH	FAM20C	fdrg103	1.5	Putative
YEVLTPL	Amelogenin	fdrg103	1.5	Putative
YEVLTPL	Amelogenin	fdrg103	1.5	Putative
YEVLTPL	Amelogenin	fdrg103	1.5	Putative
YEVLTPL	Amelogenin	fdrg103	1.5	Putative
YEVLTPL	Amelogenin	fdrg103	1.5	Putative
YEVLTPL	Amelogenin	fdrg103	1.5	Putative
YEVLTPLK	Amelogenin	fdrg103	1.5	Putative
YEVLTPLK	Amelogenin	fdrg103	1.5	Putative
YEVLTPLK	Amelogenin	fdrg103	1.5	Putative
YEVLTPLK	Amelogenin	fdrg103	1.5	Putative
ARLGIMSSEEIAVSNVF	Ameloblastin	fdrg103_SP3	1.5	Clade match
ARLGIMSSEEIAVSNVF	Ameloblastin	fdrg103_SP3	1.5	Clade match
KGAPGPPGPPGTSGSPGLQGMPGER	Collagen III alpha 1	fdrg103_SP3	1.5	Putative
KHNTYHTLYPEEIPSPAR	Enamelin	fdrg103_SP3	1.5	High confidence
LTGNSHGPNLGSNPTAQ	Enamelin	fdrg103_SP3	1.5	Putative
LTGNSHGPNLGSNPTAQ	Enamelin	fdrg103_SP3	1.5	Putative
LTGNSHGPNLGSNPTAQ	Enamelin	fdrg103_SP3	1.5	Putative
LTGNSHGPNLGSNPTAQ	Enamelin	fdrg103_SP3	1.5	Putative
LTGNSHGPNLGSNPTAQ	Enamelin	fdrg103_SP3	1.5	Putative
LTGNSHGPNLGSNPTAQ	Enamelin	fdrg103_SP3	1.5	Putative
LTGNSHGPNLGSNPTAQ	Enamelin	fdrg103_SP3	1.5	Putative
LTGNSHGPNLGSNPTAQ	Enamelin	fdrg103_SP3	1.5	Putative
LTGNSHGPNLGSNPTAQ	Enamelin	fdrg103_SP3	1.5	Putative
LTGNSHGPNLGSNPTAQ	Enamelin	fdrg103_SP3	1.5	Putative
QAHLMAGIRPSTITSSFPQI	MMP20	fdrg103_SP3	1.5	High confidence
QAHLMAGIRPSTITSSFPQI	MMP20	fdrg103_SP3	1.5	High confidence
QYIYRPAGGLSRSGGKEGDAK	DMP1	fdrg103_SP3	1.5	Putative
QYIYRPAGGLSRSGGKEGDAK	DMP1	fdrg103_SP3	1.5	Putative
QYIYRPAGGLSRSGGKEGDAK	DMP1	fdrg103_SP3	1.5	Putative
QYIYRPAGGLSRSGGKEGDAK	DMP1	fdrg103_SP3	1.5	Putative
WNAFTLEGKQIARQGYPAFHKAYA	Enamelin	fdrg103_SP3	1.5	High confidence

YIYRPAGGLSRSGGKEEDAK	DMP1	fdrg103_SP3	1.5	Putative
YIYRPAGGLSRSGGKEEDAK	DMP1	fdrg103_SP3	1.5	Putative
YIYRPAGGLSRSGGKEEDAK	DMP1	fdrg103_SP3	1.5	Putative
YIYRPAGGLSRSGGKEEDAK	DMP1	fdrg103_SP3	1.5	Putative

Table S4:
High confidence diagenetiforms. Protein blast searches have been conducted using NCBI for all clade match peptides, checking for peptide sequence similarity with living relatives from the same clade, within vertebrata, within humans, and among microbiota. Lighter, yellow colors indicate either specificity of the sequence to close relatives within the clade, or poor matches to other vertebrates, or no matches to known microbial products. The number of high-confidence PSMs for each sequence as shown in the column “n=”. Cases where a protein is represented by two peptides (including both high confidence and putative diagenetiforms, the latter not shown in this table) are indicated by purple dots. Note that the 1.5 Ma Koobi Fora elephantid and the 16 Ma Buluk deinotheriid include proteins making the 2-peptide cutoff but not shown here: AMELX and COL1A1, respectively. Animal silhouettes were created by artists at PhyloPic, as follows: Elephantidae and Rhinocerotidae cf. *Chilotheridium pattersoni*: 'An Ignorant Atheist’ under a Creative Commons licence CC0 1.0 ; Deinotheriidae: D. Bogdanov under a Creative Commons licence CC BY 3.0 ; Hippopotamidae: J. Bayona under a Creative Commons licence CC BY 3.0 ; Mammutidae: S. Traver under a Creative Commons licence CC0 1.0 ; Gomphotheriidae: T. M. Keesey under a Creative Commons licence CC0 1.0.

	Sequence	Protein	Specimen	Age	Clade	Vertebrata	Microbial	n =	2 peptide
	QNTESKSSE	DMP1	KF	1.5				1	●
	AGNGPGPNP	ENAM	KF	1.5				1	●
	KGTNLQD	ENAM	KF	1.5				1	●
	QPARAFPP	ENAM	KF	1.5				1	●
	KHNTYHTLYPEEIPSPAR	ENAM	KF-SP3	1.5				1	●
	WNAFTLEGKQIARQGYPAFHKAYA	ENAM	KF-SP3	1.5				1	●
	QAHLMAGIRPSTITSSFPQI	MMP20	KF-SP3	1.5				2	
	TQPNAPNPRG	ENAM	KNM-LT26247	7				1	
	QNTESKSSE	DMP1	KNM-WS101824	17				2	
	SPPTVNVSN	ENAM	KNM-WS101824	17				1	●
	QNTESKSSE	DMP1	KNM-WS65217	17				2	
	SSKLATHLFSGT	ENAM	KNM-WS65217	17				1	●
	TVPHGPPQ	MMP20	KNM-WS65217	17				1	
	PAGHSRGPP	AMBN	KNM-LP30	18				1	
	EVVSEPR	DMP1	KNM-LP30	18				1	
	SLQSKNAP	ENAM	KNM-LP30	18				1	

Legend

Clade

- 100% match to extant closely related taxa; but to few or no others within clade
- 100% match to extant closely related taxa; close match to others within clade
- 100% match to some taxa within clade

Vertebrata

- Does not match vertebrate sequences outside closely related taxa
- c. 90 % match to enamel protein in other mammals
- 100 % match to enamel protein in other mammals

Microbial

- Does not match any known microbial products
- Matches three or fewer unrelated microbial sequences
- 100 % match to several or more microbial products

Table S5:
Control peptides. Results frequently matched human sequences, and included almost no enamel-derived proteins. In the few instances that sequences matched enamel proteins, these were to taxa unrelated to fossil samples within mammalia.

Base Sequence	Protein	Accession
TSNVPSQR	Enamelin	Control
TSNVPSQR	Enamelin	Control
TSNVPSQR	Enamelin	Control
SSQVPQSR	Alpha-lactalbumin	Control
SSQVPQSR	Alpha-lactalbumin	Control
SSQVPQSR	Alpha-lactalbumin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VATVSLPR	Trypsin	Control
SSQVPQSR	Alpha-lactalbumin	Control
SSQVPQSR	Alpha-lactalbumin	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
FQLFGSPSGQK	Lactotransferrin	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
SSQVWKH	Filaggrin-2	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
SSQVPQSR	Alpha-lactalbumin	Control
SSQVPQSR	Alpha-lactalbumin	Control
SSQVPQSR	Alpha-lactalbumin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control

VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VATVSLPR	Trypsin	Control
SSQVPQSR	Alpha-lactalbumin	Control
SSQVPQSR	Alpha-lactalbumin	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
FQLFGSPSGQK	Lactotransferrin	Control
VATVSLPR	Trypsin	Control
VGAPGPAGSR	Collagen I alpha 2	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
SSQVWKH	Filaggrin-2	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VGPSGPAGAR	Collagen I alpha 2	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VAGSRGAPGP	Collagen I alpha 2	Control
TLGSIVMM	Complement factor H	Control
SSQVPQSR	Alpha-lactalbumin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
SSRLSGGGAP	FAM20C	Control
SSQVPQSR	Alpha-lactalbumin	Control
SSQVPQSR	Alpha-lactalbumin	Control
SSQVPQSR	Alpha-lactalbumin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control

VATVSLPR	Trypsin	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VATVSLPR	Trypsin	Control
SSQVPQSR	Alpha-lactalbumin	Control
SSQVPQSR	Alpha-lactalbumin	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
FQLFGSPSGQK	Lactotransferrin	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
VATVSLPR	Trypsin	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
SSQVWKH	Filaggrin-2	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
FAAFIDK	Keratin, type II cuticular Hb6	Control
DGAKGEPGVVGA	Collagen I alpha 2	Control
ENGIEIHAQGTK	Collagen I alpha 2	Control
VTAVNLVLLATKET	Collagen I alpha 2	Control
QGPKGEIGPMGPI	Collagen I alpha 2	Control
LSSDITAS	Lysozyme C	Control
DGAKGEPGVVGA	Collagen I alpha 2	Control
ENGIEIHAQGTK	Collagen I alpha 2	Control
IVDALAVFLRRL	Aminoglycoside 3'-phosphotransferase	Control
VTAVNLVLLATKET	Collagen I alpha 1	Control
QGPKGEIGPMGPI	Collagen I alpha 2	Control
VTAVNLVLLATKET	Collagen I alpha 1	Control
VGPQGIKG	Collagen III alpha 1	Control
HSLSEYFSLLT	Beta-1,3-N-acetylglucosaminyltransferase lunatic fringe	Control
HSLSEYFSLLT	Beta-1,3-N-acetylglucosaminyltransferase lunatic fringe	Control

Figure S2:
All proteins found in our fossil specimens that are classified as either putative or high confidence diagenetiforms. This figure shows two results from FAM20C and COL1A2, not shown in the main text Fig. 2B because they are represented only by a single PSM. Putative diagenetiforms are shown in faded green, and high confidence diagenetiforms in bright green; proteins that pass the two-peptide threshold are indicated by purple circles.

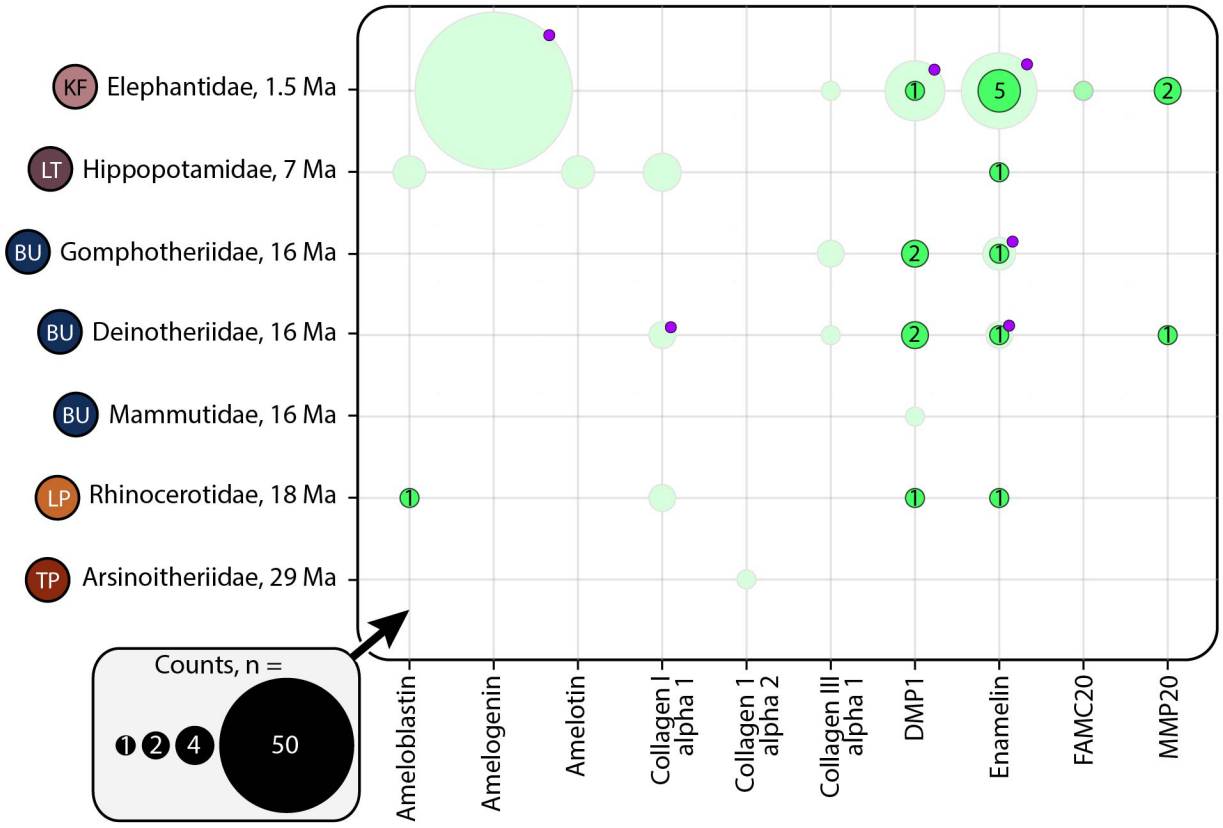

Figure S3:
Probability density function showing that the lengths of peptides derived from fossil samples (blue) are shorter than those derived from buffer controls (gray). We omit sequences derived from the Koobi Fora specimen processed using the SP3 approach, and buffer control sequences matching either trypsin or lysozyme c, from this analysis.

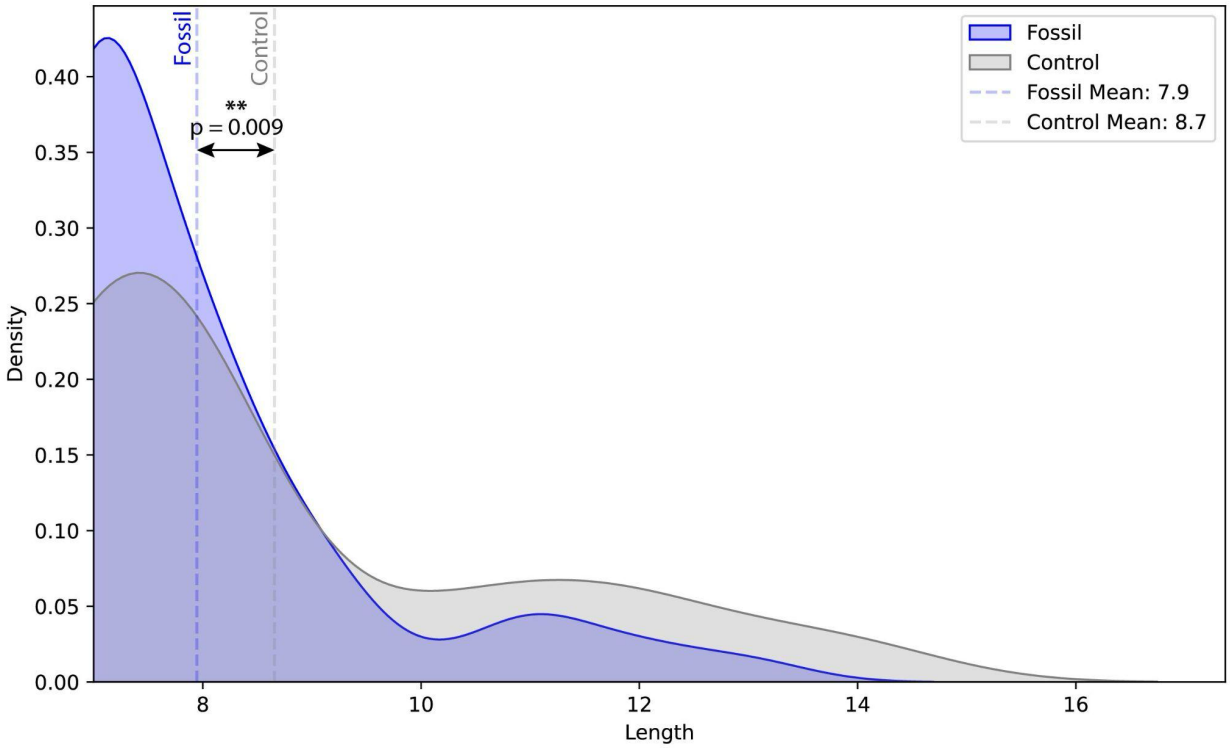

Figure S4:
Peptides derived from fossil samples are more acidic than buffer control samples; this result is statistically significant for fossil sample-derived peptides belonging to either FDR threshold, clade match, or putative diagenetiforms.

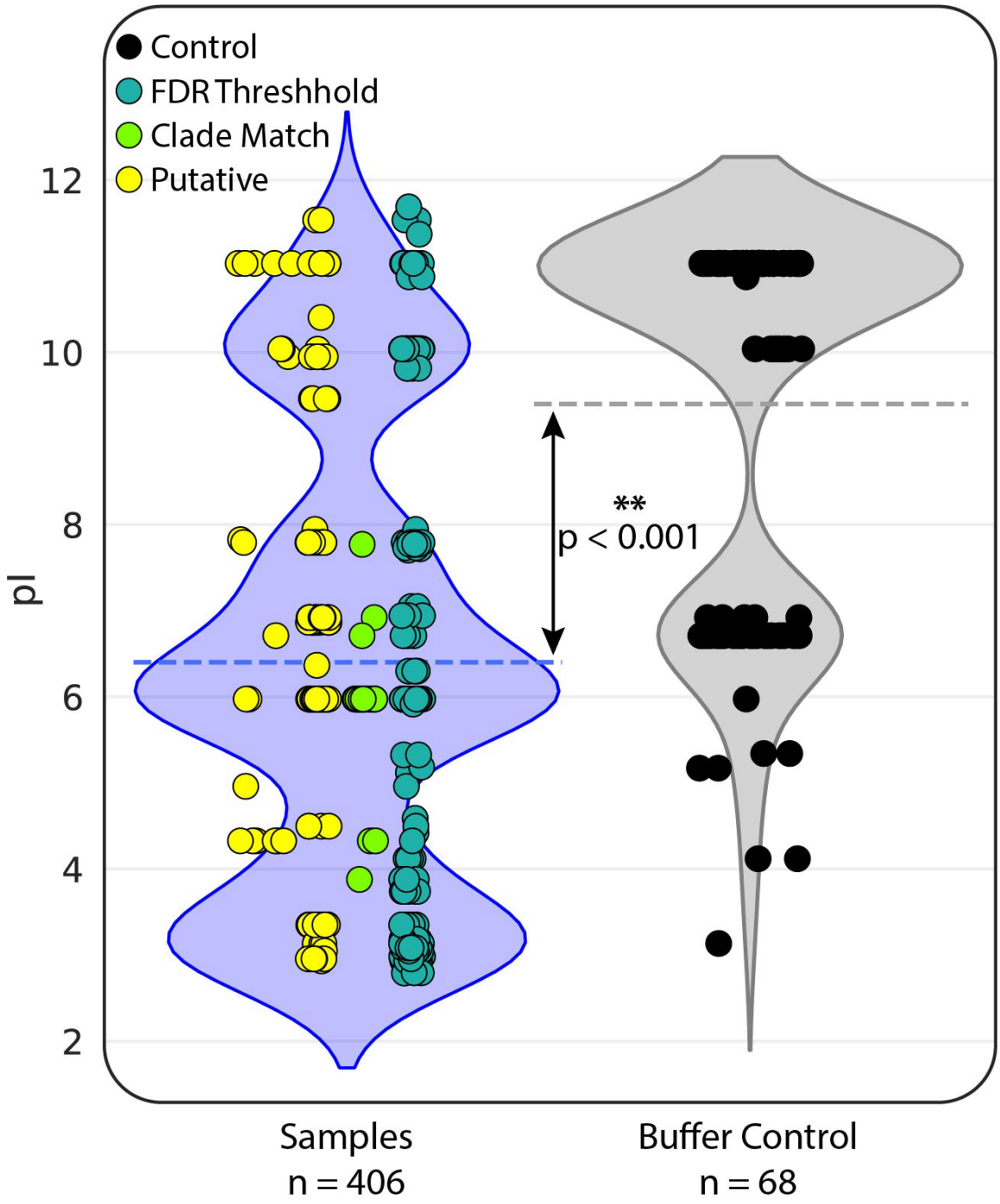

Figure S5:
Advanced glycation end products from peptides and diagenetiforms, all samples.

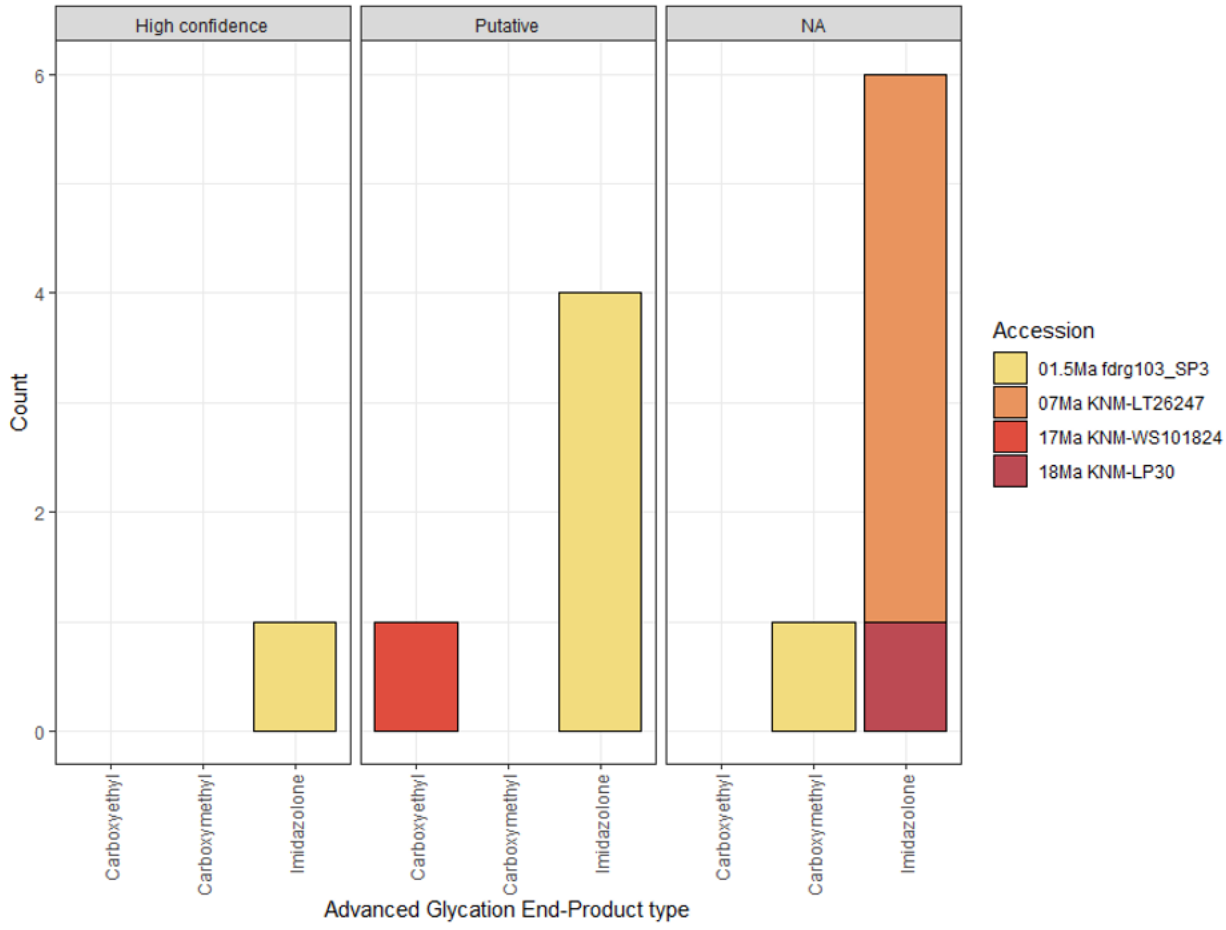

Figure S6:
Diagenetic carbamylation from peptides and diagenetiforms, all samples

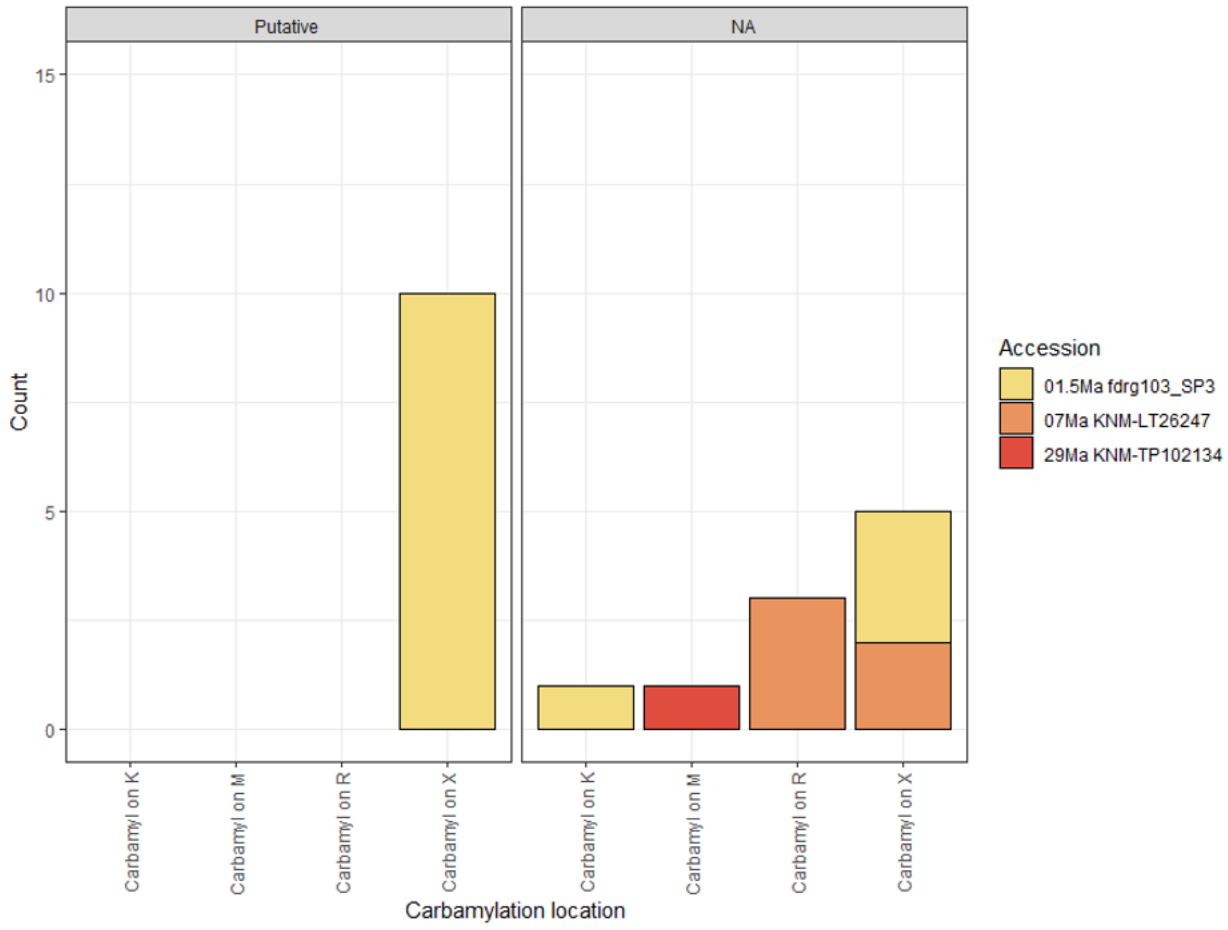

Figure S7:
Examples of hydrolysis in diagenetiforms

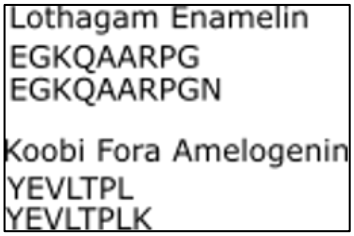

Figure S8:
Oxidation levels from methionine and histidine, all samples.

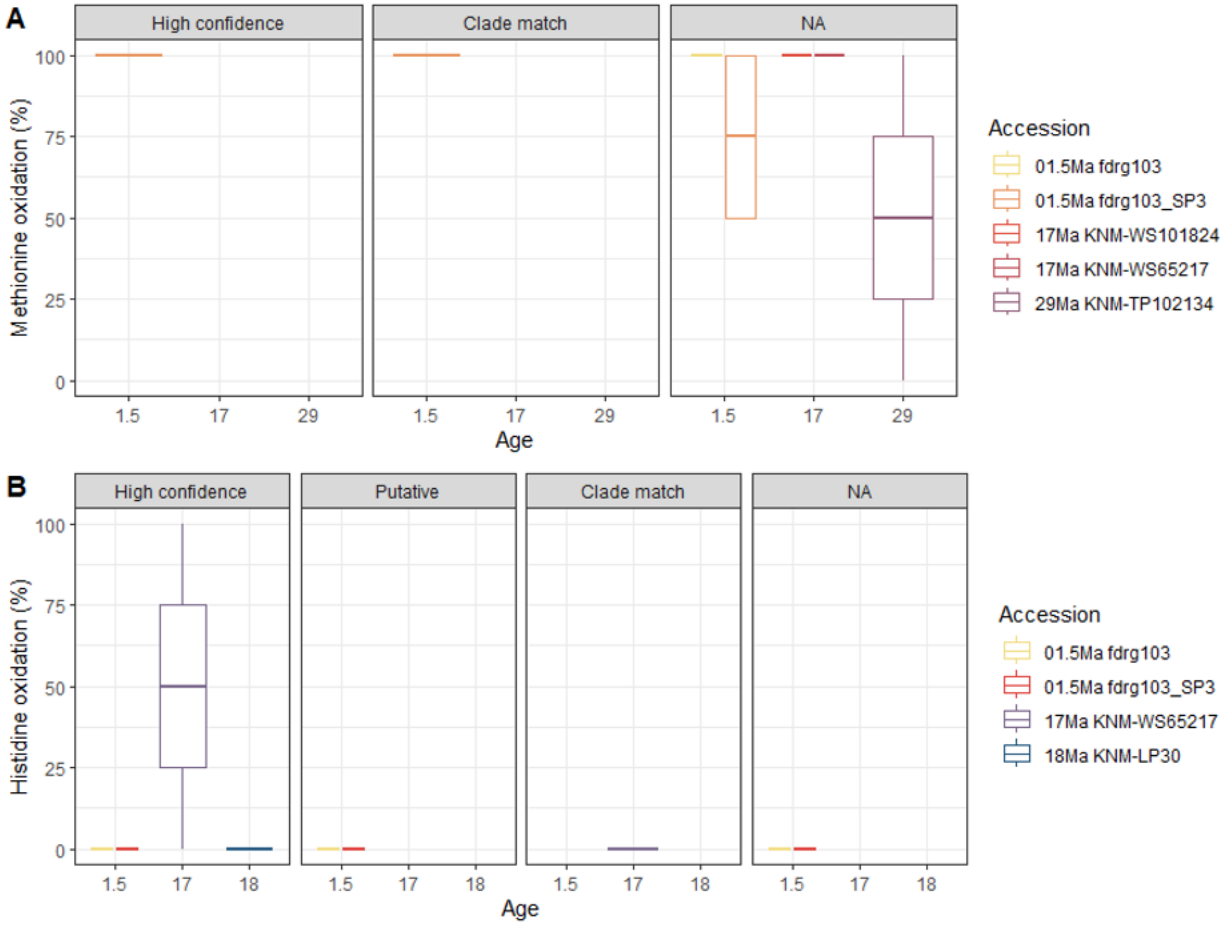

Figure S9:
Deamidation levels of asparagine and glutamine, all samples.

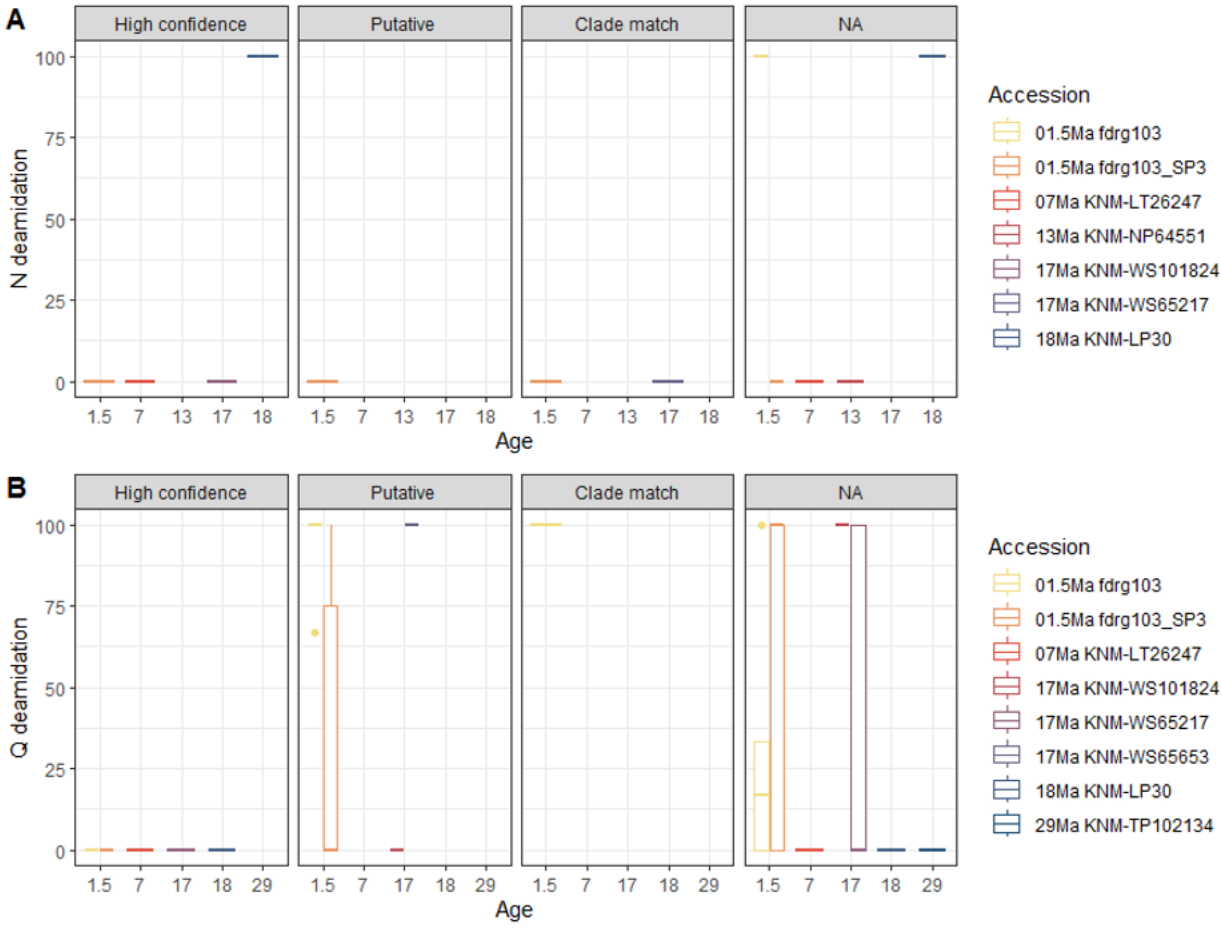

Figure S10:
Enamelin and Matrix Metalloprotease-20 fragment alignments for Buluk specimens. Enamelin and matrix metalloprotease-20 fragments are shown aligned against other known or predicted Afrotherian sequences derived from Uniprot and NCBI. Putative phylogenetic trees are shown for reference at the left of sequences. Pale green markers indicate putative diagenetiforms and bright green high confidence diagenetiforms.

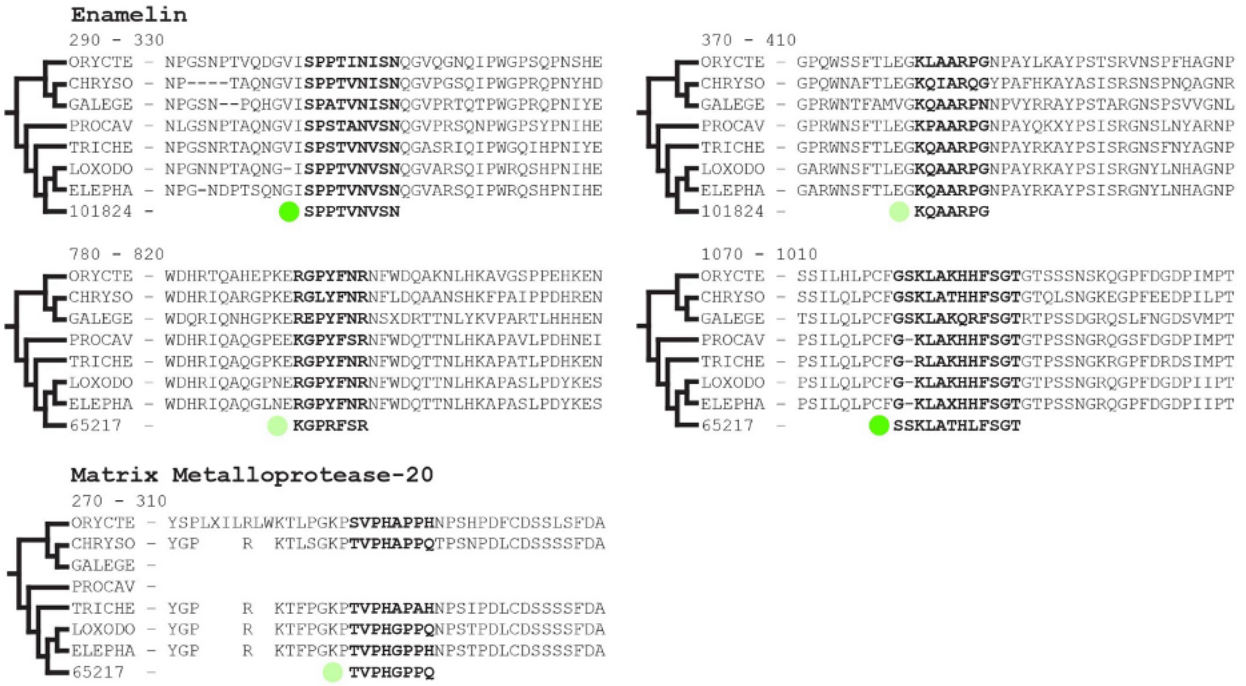

Figure S11:
A maximum likelihood phylogenetic tree built using all putative and high confidence diagenetiforms in our study, compared to other afrotherians, xenarthrans, ungulates, and a canid outgroup. The tree places all proboscideans within Afrotheria, though two are placed alongside modern proboscideans and two in a sister clade. The fossil hippo is placed alongside extant hippopotamids and the fossil rhino alongside extant rhinocerotids. Animal silhouettes were created by artists at PhyloPic, as follows: Elephantidae and Rhinocerotidae cf. *Chilotheridium pattersoni*: 'An Ignorant Atheist' under a Creative Commons licence CC0 1.0 ; Deinotheriidae: D. Bogdanov under a Creative Commons licence CC BY 3.0 ; Hippopotamidae: J. Bayona under a Creative Commons licence CC BY 3.0 ; Mammutidae: S. Traver under a Creative Commons licence CC0 1.0 ; Gomphotheriidae: T. M. Keesey under a Creative Commons licence CC0 1.0 . The Arsinoitheriidae image was created by author D.R.G. The silhouette used for *Stephanorhinus kirchbergensis* is *Rhinoceros unicornis* by H. F. O. March (vectorized by T. Michael Keesey) Public Domain Mark 1.0.

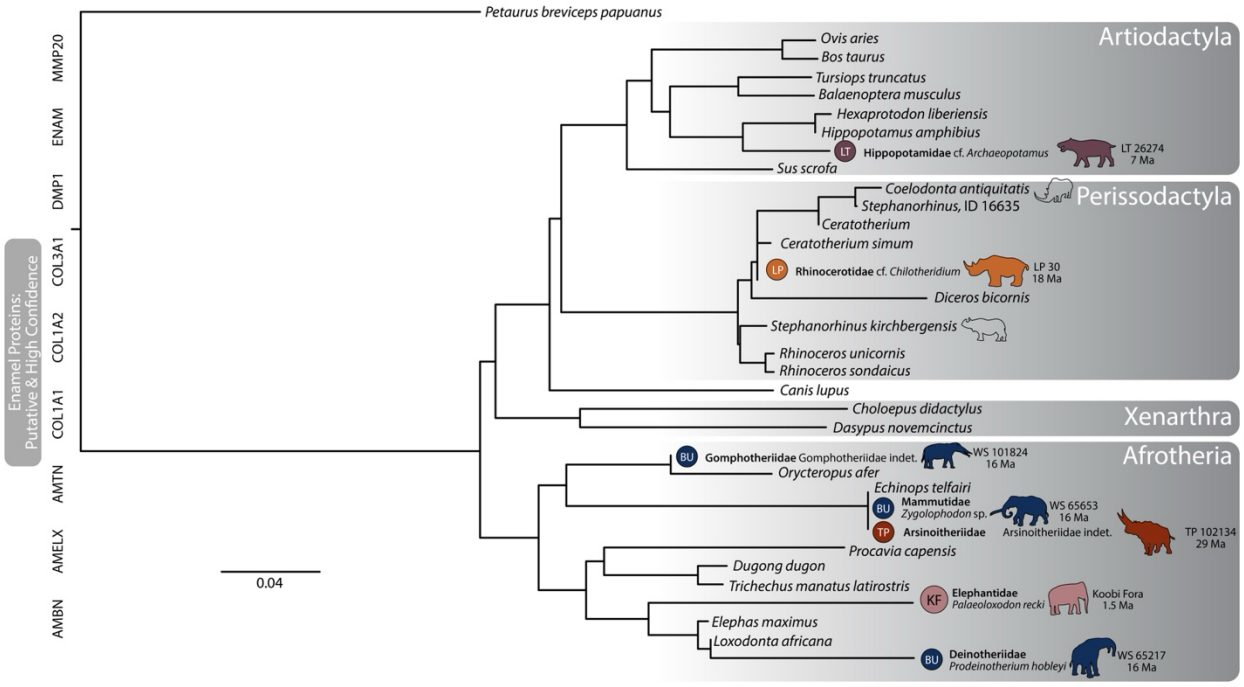

Figure S12:
AMELX peptide from the Koobi Fora *Palaeoloxodon* enamel sample detected with DDA (top) and DIA (bottom).