The dental proteome of *Homo antecessor*

Frido Welker[1,21,*], Jazmín Ramos-Madrigal[1,21], Petra Gutenbrunner[2,21], Meaghan Mackie[1,3], Shivani Tiwary[2], Rosa Rakownikow Jersie-Christensen[3], Cristina Chiva[4,5], Marc R. Dickinson[6], Martin Kuhlwilm[7], Marc de Manuel[7], Pere Gelabert[7], María Martinón-Torres[8,9], Ann Margvelashvili[10], Juan Luis Arsuaga[11,12], Eudald Carbonell[13,14], Tomas Marques-Bonet[4,7,15,16], Kirsty Penkman[6], Eduard Sabidó[4,5], Jürgen Cox[2], Jesper V. Olsen[3], David Lordkipanidze[10], Fernando Racimo[17], Carles Lalueza-Fox[7], José María Bermúdez de Castro[8,9,*], Eske Willerslev[17,18,19,20,*], Enrico Cappellini[1,*]

[1]Evolutionary Genomics Section, Globe Institute, University of Copenhagen, Copenhagen, Denmark.

[2]Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany.

[3]The Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark.

[4]Centre for Genomic Regulation (CNAG-CRG), Barcelona Institute of Science and Technology, Barcelona, Spain.

[5]Proteomics Unit, Universitat Pompeu Fabra, Barcelona, Spain.

[6]Department of Chemistry, University of York, York, United Kingdom.

[7]Institute of Evolutionary Biology (UPF-CSIC), University Pompeu Fabra, Barcelona, Spain.

[8]Centro Nacional de Investigación sobre la Evolución Humana (CENIEH), Burgos, Spain.

[9]Anthropology Department, University College London, London, United Kingdom.

[10]Georgian National Museum, Tbilisi, Georgia.

[11]Centro Mixto UCM-ISCIII de Evolución y Comportamiento Humanos, Madrid, Spain.

[12]Departamento de Paleontología, Facultad Ciencias Geológicas, Universidad Complutense de Madrid, Madrid, Spain.

*Corresponding authors: E. Cappellini (ecappellini@bio.ku.dk), E. Willerslev (ewillerslev@bio.ku.dk), J.-M. Bermudez de Castro (josemaria.bermudezdecastro@cenieh.es) and F. Welker (frido.welker@bio.ku.dk).

Supplementary Information
Supplementary information is available in the online version of this article.

[13]Departamento d'Història i Història de l'Art, Universidad Rovira i Virgili, Tarragona, Spain.

[14]Institut Català de Paleoecologia Humana i Evolució Social (IPHES), Tarragona, Spain.

[15]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain.

[16]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain.

[17]Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark.

[18]Department of Zoology, University of Cambridge, Cambridge, United Kingdom.

[19]Wellcome Sanger Institute, Hinxton, United Kingdom.

[20]Danish Institute for Advanced Study, University of Southern Denmark, Odense, Denmark.

[21]These authors contributed equally.

The phylogenetic relationships between Early Pleistocene Eurasian hominins, like *Homo antecessor*, and hominins that appear in the fossil record during the late Middle Pleistocene, like *Homo sapiens*, are highly debated[1–5]. For the most ancient remains, the molecular study of these relationships is hindered by ancient DNA degradation. However, recent research has demonstrated that ancient protein analysis can address this challenge[6–8]. Here, we obtain dental enamel proteomes from *Homo antecessor* (Atapuerca, Spain)[9,10] and *Homo erectus* (Dmanisi, Georgia)[1], two key fossil assemblages that have a central role in models of Pleistocene hominin morphology, dispersal, and divergence. We demonstrate that *Homo antecessor* is a close sister lineage to subsequent Middle and Late Pleistocene hominins such as modern humans, Neanderthals, and Denisovans. This placement implies that the modern-like face of *Homo antecessor* may have a considerably deep ancestry in the genus *Homo*, and that the Neanderthal cranial morphology represents a derived form. By recovering AMELY-specific peptide sequences we also conclude that the Atapuerca molar fragment we analysed belonged to a male individual. Finally, we observe *in vivo* enamel proteome phosphorylation and proteolytic digestion that occurred during tooth formation. Our results thereby provide important insights into the evolutionary relationships of *Homo antecessor* to other hominin groups, and pave the way for further insights into hominin biology across the existence of the genus *Homo* through the study of their enamel proteomes.

Since 1994, over one hundred and seventy human fossil remains have been recovered from level TD6 of the Gran Dolina site of the Sierra de Atapuerca[10] (Burgos, Spain, Extended Data Fig. 1; Supplementary Information). These fossils have been dated to the late Early Pleistocene and exhibit a unique combination of cranial, mandibular and dental features[9,11]. To accommodate the variation observed in the TD6 human fossils, a new species of the genus *Homo*, *H. antecessor*, was proposed in 1997[9]. The relationships of this species to earlier hominins in Eurasia (such as the *Homo erectus* specimens from Dmanisi), and to later hominins (such as Neanderthals, Denisovans, and modern humans), have been the subject of considerable debate[3,4,12,13]. These issues remain unresolved due to the fragmentary nature of hominin fossils at other sites, and the failure to recover ancient DNA in Eurasia from the Early and most of the Middle Pleistocene. On the contrary, recent developments in the

extraction and tandem mass-spectrometric analysis of ancient proteins have made it possible to retrieve phylogenetically informative protein sequences from Early Pleistocene contexts[6,8]. We therefore applied ancient protein analysis to a *Homo antecessor* molar from Atapuerca, Gran Dolina TD6.2 (Specimen ATD6–92; Extended Data Fig. 2a). This specimen, identified as an enamel fragment of a permanent lower left first or second molar, has been directly dated to 772–949 thousand years ago (ka) using a combination of electron spin resonance (ESR) and U-series dating[11]. In addition, we sampled dentine and enamel from an isolated *Homo erectus* upper first molar (D4163; Extended Data Fig. 2b) from Dmanisi, Georgia, dated to 1.77 million years ago (Ma)[1,14,15], as amino acid racemization analysis of this specimen indicated the presence of an endogenous protein component in the intra-crystalline enamel fraction of the tooth (Extended Data Fig. 3; Supplementary Information). On both specimens, we performed digestion-free peptide extraction optimised for the recovery of short, degraded, protein remains[6]. Nano liquid chromatography tandem mass spectrometry (nanoLC-MS/MS) acquisition was replicated in two independent proteomic laboratories (Extended Data Tab. 1), implementing common precautions and analytical workflows to minimize protein contamination (Methods). We compared the proteomic datasets retrieved from the Pleistocene hominin tooth specimens with those generated from a positive control, a recent human premolar (Ø1952, male, approximately three centuries old), and previously published Holocene teeth[16] (Methods, Supplementary Information). Finally, to validate our enamel peptide spectrum matches (PSMs), we performed machine learning-based MS/MS spectrum intensity prediction using the wiNNer algorithm[17]. Results show that the wiNNer model, re-trained for randomly cleaved and heavily modified peptides, provides similar predictive performance compared to the wiNNer model trained on modern, trypsin-digested samples, assuring accurate sequence identification for the phylogenetically informative peptides (median Pearson correlation coefficients of   0.76; Fig. S6; see Methods and Supplementary Information).

Protein recovery from the Dmanisi dentine sample was limited to sporadic collagen type I fragments. Therefore, in-depth analysis of this material was not further pursued. In contrast, we recover ancient proteomes from both hominin enamel samples and observe that their composition is similar to those from the recent human specimen we processed as a positive control and ancient enamel proteomes previously published[6,16,18,19] (Extended Data Tab. 2; Tab. S6). The enamel-specific proteins include amelogenin (AMELX and AMELY), enamelin (ENAM), ameloblastin (AMBN), amelotin (AMTN), and the enamel-specific protease matrix metalloproteinase-20 (MMP20). Serum albumin (ALB), and collagens (COL1α1, COL1α2, COL17α1) are also present. For the enamel-specific proteins, the peptide sequences retrieved cover approximately the same protein regions in all the specimens analysed (Extended Data Fig. 4). Although destructive, our sampling of Pleistocene hominin teeth resulted in higher protein sequence coverage than acid-etching of Holocene enamel surfaces[16,20] (Fig. S7). The AMTN-specific peptides largely derive from a single sequence region involved in hydroxyapatite precipitation through the presence of phosphorylated serines[21]. Finally, the observation of AMELY-specific peptides, the amelogenin isoform coded on the non-recombinant portion of the Y-chromosome, demonstrates that the studied *Homo antecessor* molar belonged to a male individual[16] (Extended Data Fig. 5).

Besides proteome composition and sequence coverage, several further lines of evidence independently support the endogenous origin of the hominin enamel proteomes. Unlike exogenous trypsin, keratins and other human skin contaminants identified, the enamel proteins have high deamidation rates (Extended Data Fig. 6), and above that observed for the recent human specimens (Fig. S8). Both Pleistocene hominins have average peptide lengths shorter than observed for our recent human controls (Extended Data Fig. 6d). The average peptide length is shorter in the Dmanisi hominin, but longer in the younger Atapuerca hominin (Extended Data Fig. 6d). In contrast, we observe that the Dmanisi hominin peptide lengths are indistinguishable from those of the faunal remains from the same site. Together, our protein data is therefore in agreement with theoretical and experimental[6,22] expectations for samples of their relative age. In addition to diagenetic modifications, we observe two kinds of *in vivo* modifications in our recent and ancient enamel proteomes. First, we detect serine phosphorylation within the S-x-E/phS motif (Fig. 1a, b). This motif is recognized by the FAM20C secreted kinase, which is active in the phosphorylation of extracellular proteins[23,24]. The presence of phosphoserine in fossil enamel and its location in the S-x-E/phS motif has previously also been observed in other Pleistocene enamel proteomes[6,25]. Phosphorylation occupancy can be computed successfully for ancient and recent samples, and reveals differences in phosphorylated peptide ratios between samples (Fig. 1c; Tab. S5). Second, the peptide populations we retrieve primarily cover the ameloblastin, enamelin, and amelogenin sequence regions representing cleavage products deriving from *in vivo* activity of the proteases MMP20 and, subsequently, kallikrein-4 (KLK4; Extended Data Fig. 4; Methods). The peptide populations are also enriched in N- and C-termini corresponding to known MMP20 and KLK4 cleavage sites (Extended Data Fig. 7, Fig. S9). FAM20C phosphorylation and MMP20 and KLK4 proteolysis are the two main processes occurring *in vivo* during enamel biomineralization. Our observation of products deriving from both processes opens up the possibility to study *in vivo* processes of hominin tooth formation across the Pleistocene.

Homo antecessor is only known from the Gran Dolina TD6.2 assemblage in Atapuerca[9]. Its relationship with other European Middle Pleistocene fossils is heavily debated[3–5,26,27]. It is still contentious whether *Homo antecessor* could represent the last common ancestor of *Homo sapiens*, Neanderthals, and Denisovans[9], or whether it represents a sister lineage to the last common ancestor of these species (here collectively called HNDs)[28,29]. We address this issue by conducting a set of phylogenetic analyses based on our ancient protein sequences from *Homo antecessor* (ATD6–92), a panel of present-day great ape genomes, and protein sequences translated from archaic hominin genomes (Methods).

We built several phylogenetic trees using maximum likelihood and Bayesian methods (Figs. 2a, Figs. S13–16). In these trees, the *Homo antecessor* sequence represents a sister taxon closely related to, but not part of, the group composed of Late Pleistocene hominins for which molecular data is available (Fig. 2a, S13, S15, S16). The enamel protein sequences do not resolve the relationships between HNDs due to the low number of informative single amino acid polymorphisms (SAPs). However, pairwise amino acid sequence divergence between *Homo antecessor* and HNDs is larger than between HNDs (Fig 2b, S12; Supplementary Information). The concatenated gene tree may suffer from incomplete lineage sorting, and we have too little sequence data to discard this possibility at the

moment. If we were, however, to use the concatenation of available gene trees as a best guess for the population tree, and assuming such a population tree is a good descriptor of the relationships among ancient hominins, then our results support the placement of *Homo antecessor* as a closely related sister taxon of the last common ancestor of HNDs. The phylogenetic position of *Homo antecessor* agrees with a divergence of the *Homo sapiens* and Neanderthal+Denisovan lineages between 550 ka and 765 ka[30,31], while ATD6–92 has been dated to 772–949 ka[11]. This is further supported by recent reconsiderations of the morphology of *Homo antecessor* in relation to Middle and Late Pleistocene hominins[29].

Homo antecessor was tentatively proposed as the last common ancestor of Neanderthals and modern humans[9]. The modern-like face of some immature individuals, and particularly that of the more complete specimen ATD6–69, as well as the zygomaxillary fragment ATD6–58 of one adult individual, were key in this proposition[9,32]. Additional studies of the face of ATD6–69 have confirmed that *Homo antecessor* exhibits the oldest known modern-like face of the fossil record[12,13]. The phylogenetic placement of *Homo antecessor* implies that the modern-like face as represented by *Homo antecessor* must have a considerably deep ancestry in the genus *Homo*. New findings made between 2003 and 2005 have shown that the *Homo antecessor* hypodigm includes some previously considered autapomorphic Neanderthal features[28]. Our results suggest that these features appeared during the Early Pleistocene and were retained by Neanderthals and lost by modern humans.

In contrast, the phylogenetic tree built with the *Homo erectus* specimen from Dmanisi has only moderate resolution (Extended Data Fig. 8; Fig. S11), despite deeper shotgun protein sequencing for this specimen (Extended Data Tab. 1). This partly inconclusive result might be due to the shorter average peptide lengths compared to the Atapuerca specimen (Extended Data Fig. 6d; Methods) and an absence of uniquely segregating SAPs (Tab. S9). Although our *Homo erectus* (Dmanisi) data demonstrate that ancient hominin proteins can be reliably obtained from the Early Pleistocene, it also highlights the current limits of ancient protein analysis when applied to attempt phylogenetic placement of Early Pleistocene hominin remains. Our dataset provides a unique molecular resource of hominin biomolecular sequences from Early and Middle Pleistocene hominins, and is older than the oldest ancient hominin genomes presented to date. Comparison of hominin and fauna proteomes from different skeletal tissues reveals that the dental enamel proteome outlasts dentine and bone proteome preservation (Fig. 3). Here, the prolonged survival of hominin enamel proteomes is exploited to show that *Homo antecessor* represents a hominin taxon closely related to the last common ancestor of *Homo sapiens*, Neanderthals, and Denisovans. In addition, our datasets demonstrate that *in vivo* proteome modifications, like serine phosphorylation, survive over the same timescales. Current research therefore suggests that dental enamel, the hardest tissue in the mammalian skeleton, is the material of choice for deep-time analysis of hominin evolution.

METHODS

Site Location & Specimen Selection

Recent human control specimens.—We analysed one human premolar recovered in an archaeological excavation in Copenhagen (Almindeligt Hospital Kirkegård,, excavated in

1952, from "kisse '2' ", hereafter Ø1952). The tooth is approximately three centuries old as the cemetery was in use from approximately 1600–1800 AD and originates from a male individual. We also re-analyzed published data from Stewart *et al.*[16]. Specimens presented therein are between approximately 5,700 and 200 years old. We took SK339 as a recent example in our comparative figures. SK339 represents a male individual from Fewston (United Kingdom, 19th century AD).

Atapuerca.—One fragmentary permanent lower left first or second molar (ATD6–92, field number and museum accession number at CENIEH) was used for ancient protein analysis (Extended Data Fig. 2a; Supplementary Information). ATD6–92 originates from layer TD6.2 from the Gran Dolina, Atapuerca, Spain. Layer TD6.2 contains a large number of faunal remains, about one hundred and seventy hominin fossils, and about 830 archaeological artefacts. All hominin specimens from layer TD6.2, including specimen ATD6–92, are attributed to *Homo antecessor*[9]. Specimen ATD6–92 has recently been directly dated through Electron Spin Resonance, LA-ICP-MS U-series and bulk U-series dating[11]. Together with previous chronological research at the site, this constraints the age of specimen ATD6–92 to 772–949 ka[11].

Dmanisi.—One fragmentary permanent upper first molar (D4163, field number and museum accession number at the Georgian National Museum) was used for ancient protein analysis (Extended Data Fig. 2b; Supplementary Information). D4163 derives from layer B1 in excavation block M6, Dmanisi, Georgia. Layer B1 at Dmanisi contains one of the richest paleontological assemblages attributed to the Eurasian Early Pleistocene, including several hominin crania. Here, we simply refer to these specimens as *Homo erectus* (Dmanisi). They represent the earliest hominin fossils outside Africa, and are dated to 1.76–1.78 Ma[14]. Faunal material from the site previously demonstrated ancient protein survival for most specimens, but a total absence of ancient DNA[6] (Fig. 3).

Amino Acid Racemization

Chiral amino acid analysis was undertaken on one Pleistocene sample from the hominin tooth (D4163) to test the endogeneity of the enamel protein through its degradation patterns. The tooth chip was separated into the enamel and dentine portions, and each was powdered with an agate pestle and mortar. All samples were prepared using modified procedures of Penkman *et al.*[36], but optimized for enamel, using a bleach time of 72 hours to isolate the intra-crystalline protein, demineralization in HCl, KOH neutralization, and formation of a biphasic solution through centrifugation[37]. Two subsamples were analyzed from each portion: one fraction was directly demineralized and the free amino acids analyzed (referred to as the 'free' amino acids, FAA, F), and the second was treated to release the peptide-bound amino acids, thus yielding the 'total hydrolysable' amino acid fraction (THAA, H*). Samples were analyzed in duplicate by RP-HPLC, with standards and blanks analysed alongside samples. During preparative hydrolysis, both asparagine (Asn) and glutamine (Gln) undergo rapid irreversible deamidation to aspartic acid (Asp) and glutamic acid (Glu) respectively[38]. It is therefore not possible to distinguish between the acidic amino acids and their derivatives and they are reported together as Asx and Glx, respectively. See Supplementary Information for additional methods description and results.

Proteomic Extraction and nanoLC-MS/MS

Protein extraction.—Protein extraction was conducted on enamel samples (Atapuerca, Dmanisi, Ø1952) and a dentine sample (Dmanisi) using one of three protocols. In short, the first extraction method employed HCl for demineralization, but included no subsequent alkylation or digestion. The second extraction method employed a more standard approach, in which the pellet left from the demineralization in extraction one was reduced, alkylated, and digested with LysC and trypsin. The third extraction method employed TFA for demineralization, and no subsequent alkylation or digestion. The first and third extraction approaches provided more extensive peptide recovery in ancient enamel proteomes[6] compared to the second extraction approach[39]. Further details can be found in the Supplementary Information and Cappellini *et al.*[6]. Ø1952 was processed using extraction methods one and three. No proteinase and phosphatase inhibitors were used during extraction as we assumed that catalytically active enzymes were not present in our specimens, while the high acidic conditions during our extraction would have irreversibly denatured any proteases possibly present as contaminants in our reagents. See Extended Data Table 1 for a breakdown of the employment of specific extraction methods, hominin samples, and hominin tissues.

NanoLC-MS/MS analysis.—Shotgun proteomic data was obtained on peptide extracts of both hominins at separate facilities at the Novo Nordisk Centre for Protein Research, University of Copenhagen (Denmark), and the Proteomics Unit, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (Spain). Full peptide elutions were injected, in some cases across replicate runs in both Copenhagen and Barcelona. Briefly, samples processed in Copenhagen were suspended in 0.1% trifluoroacetic acid, 5% acetonitrile, and analyzed on a Q-Exactive HF or HF-X mass spectrometer (Thermo Fisher Scientific) coupled to an EASY-nLC 1200 (Thermo Fisher Scientific). The HF/HF-X was operated in positive ion mode with a nanospray voltage of 2 kV and a source temperature of 275°C. Data-dependent acquisition (DDA) mode was used for all mass spectrometric measurements. Full MS scans were done at a resolution of 120,000 with a mass range of m/z 300–1750/350–1400 (HF/HF-X) with detection in the Orbitrap mass analyzer. Fragment ion spectra were produced at a resolution of 60,000 via high-energy collision dissociation (HCD) at a normalized collision energy of 28% and acquired in the Orbitrap mass analyzer. In addition, test runs for the Dmanisi sample were performed at a shorter gradient (see Supplementary Information). In Barcelona, samples were dissolved in 0.1% formic acid and analyzed on a LTQ-Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) coupled to an EASY-nLC 1000. The mass spectrometer was operated similarly to the parameters stated for the HF/HF-X in Copenhagen, except the nanospray voltage was 2.4 kV and full MS scans with 1 micro scans were used over a mass range of m/z 350–1500. Further details on LC-MS/MS analysis can be found in the Supplementary Information.

Proteomic Data Analysis

Protein Sequence Database construction.—We constructed an initial Hominidae sequence database containing protein sequences of all major and minor enamel proteins derived from all extant great apes, a hylobatid (*Nomascus leucogenys*), and a macaque (*Macaca mulatta*). Additionally, we added protein sequences translated from extinct Late

Pleistocene hominins[30,40], and sequences from *Gorilla beringei*, *Pongo pygmaeus*, and *Pongo tapanuliensis*[41–43]. For each protein, we reconstructed the protein sequence of ancestral nodes in the Hominidae family through PhyloBot[44] to minimize cross-species proteomic effects[45], and added missing isoform variation based on the isoforms present for each protein in the human proteome as given by UniProt (Supplementary Information). Furthermore, we downloaded the entire human reference proteome from UniProt (downloaded 04.09.2018) for a single separate search to allow matches to proteins previously not encountered in enamel proteomes. To each constructed database we added a set of known or possible laboratory contaminants, to allow for the identification of possible protein contaminants[46].

Proteomic software, settings, FDR.—Raw mass spectrometry data was searched for each specimen and tissue separately in either PEAKS[47] (v. 7.5) or MaxQuant[48] (v. 1.5.3.30). No fixed modifications were specified in any search. For PEAKS, variable modifications were set to include proline hydroxylation, glutamine and asparagine deamidation, oxidation (M), phosphorylation (STY), carbamidomethylation (C), and pyroglutamic acid (from Q and E). For MaxQuant, the following variable PTMs were additionally included: ornithine formation (R), oxidation (W) dioxidation (MW), histidine to aspartic acid (H>D), and histidine to hydroxyglutamate. Searches were conducted with "unspecific" digestion. For PEAKS, precursor mass tolerance was set to 10 ppm and fragment mass tolerance to 0.05 Da, and the FDR of peptide spectrum matches was set to equal 1.0%. For MaxQuant, default settings of 20 ppm for the first search and 4.5 ppm for the final search were used, a fragment mass tolerance of 20 ppm, and PSM and protein FDR was set to 1.0%, with a minimum required Andromeda score of 40 for all peptides. Protein matches were accepted with a minimum of two unique peptide matches in either the PEAKS or MaxQuant search. Proteins that conform these criteria are detailed in Extended Data Table 2. Example MS/MS spectra from the MaxQuant search and overlapping sites of phylogenetic interest (SAPs) are included in the "Key MSMS file.pdf".

Data search iterations.—For both Dmanisi and Atapuerca, we conducted two separate, initial searches. First, we conducted a search in PEAKS against the entire human proteome. Only standard enamel proteins were identified in these searches, allowing us to continue with more specific searches. For the Dmanisi dentine sample, this first search resulted in a small number of peptides matching to collagen type I only. Based on the limited amount of sequence data, no further analysis of the Dmanisi dentine data was therefore conducted. Second, for the enamel data, we conducted a search in PEAKS and MaxQuant against the entire enamel proteome database of all extant and extinct Hominidae. This search was used to observe single amino acid polymorphisms (SAPs) outside the known sequence variation in PEAKS and MaxQuant through the *de novo*, error-tolerant, and/or dependent peptide approaches implemented in each of these search engines. These initial searches indicate overall good protein preservation in both samples and the presence of peptide matches to *Pan*- and *Homo*-derived proteins only.

Based on these two initial searches, a novel protein sequence database was used that only includes sequences from the genus *Pan*, the genus *Homo*, their predicted ancestral

sequences, and novel protein sequences observed for either Dmanisi and Atapuerca. Final searches and subsequent data analysis were conducted against this database using the above search and PTM settings. Positions supported by insufficient spectral data were replaced by "X" in resulting peptide alignments prior to phylogenetic analysis.

Data analysis of Ø1952 and the Stewart *et al.*[16] dataset was only conducted in MaxQuant against a database restricted to *Homo sapiens*. All other search settings and database restrictions were similar between these two recent human controls and the ancient hominin proteomes.

Peptide sequence and SAP validation.—To validate the PSMs covering SAPs of interest, we performed peptide spectrum intensity prediction and validation on our dataset through wiNNer[17]. Data from the ancient samples (Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*) was divided into phylogenetically informative peptide sequences, and the larger subset not containing such phylogenetically informative peptides. A training dataset was prepared by taking a subset of the latter peptides, and adding a previously published dataset of enamel proteomes from Dmanisi fauna[6]. We build two models, one for HCD +2 spectra and one for HCD +3 spectra. We took into account the large number of variable modifications observed in our ancient enamel proteomes, and split the retained data for each model into subsets for training, validation, and testing (80:10:10). We then obtained Pearson correlation coefficients (PCCs) for the predicted and true fragment intensities in the test dataset and the phylogenetically informative spectra. The architecture of wiNNer was build using Keras (version 2.0.8; https://keras.io) and Tensorflow (version 1.3.0). wiNNer analysis indicated close correspondence between predicted and true fragment ion intensities (PCC medians between 0.85 and 0.76 for different subsets of the data), indicating adequate peptide sequence identification for all our peptides, including phylogenetically informative positions and (localization of) variable post-translational modifications. The wiNNer model can be accessed on GitHub (https://github.com/cox-labs/wiNNer.git). See the Supplementary Information for additional methodological details on wiNNer architecture.

Protein damage analysis.—Ancient proteins can be modified diagenetically in a variety of ways compared to their modern counterparts. We quantify glutamine and asparagine deamidation following Mackie *et al.*[39] for MaxQuant output, based on MS1 spectral intensities and protein-based bootstrapping (1000 bootstraps). Further details can be found in Mackie *et al.*[39]. We observe that both glutamines and asparagines are almost all deamidated to glutamic acid and aspartic acid, respectively (Extended Data Fig. 6a–c). In addition, peptide length distributions were obtained for datasets presented here and elsewhere[6,8], demonstrating a shortening of average peptide length and overall peptide length distributions for older samples (Extended Data Fig. 6d).

Protein *in vivo* modification analysis.—The existing literature on enamel and enamel proteome biomineralization describes three processes that are key to the maturation of the enamel proteome: protein hydrolysis by MMP20 and KLK4[49–52], *in vivo* phosphorylation of serine residues[6,8,23], and expression of different isoforms of AMELX, AMBN, and AMTN[49,52,53]. We sought to explore the presence of both *in vivo* protein hydrolysis and serine phosphorylation modifications in our Pleistocene hominin proteomes.

For protein hydrolysis by MMP20 and KLK4, we made use of the Atapuerca digestion-free dataset and the described locations of AMBN, AMEL(X/Y), and ENAM cleavage by MMP20 and KLK4[49-52]. We compared the experimentally observed cleavage sites to a random cleavage model of each protein separately and tested if the cleavage sites are present in a larger portion of PSMs in the ancient sample. Here, we can indeed show an increased presence of PSMs with termini at, or close to, known MMP20 and KLK4 cleavage locations (Extended Data Fig. 7). This corresponds with our observation that protein regions with continuous sequence coverage correspond to known proteolytic fragments after MMP20 and KLK4 activity (Extended Data Fig. 4).

Phosphorylation of serines (S), threonines (T), and tyrosines (Y) was assessed using Icelogo[54] sequence motif analysis. This analysis was based on the MaxQuant results, where only identified phosphorylation sites with a localization probability of 0.95 were selected. STY sites with no phosphorylation or localization probabilities 0.95 were taken as the non-phosphorylated background, and a sequence motif window of 7 amino acids on either side of the STY were selected. Sequence motif analysis indicates a strong preference for the phosphorylation of serines (S) with a glutamic acid (E) on the +2 position (S-x-E/phS motif; Fig. 1a, b) in both hominin enamel proteomes. This substrate motif is characteristic for the phosphorylation kinase FAM20C, which is known to be active *in vivo* on proteins involved in biomineralization[23], and has previously been reported for ancient, non-hominin, enamel proteomes as well[6,8].

To compare phosphorylation occupancy between the Dmanisi and Atapuerca enamel proteomes, we performed a separate MaxQuant database search (Supplementary Information) and restricted our analyses to amino acid positions covered by phosphorylated and non-phosphorylated peptides, observed in both hominins, and quantified through label-free quantification.

Phylogenetic Analysis

Comparison between the ancient protein sequences and modern reference proteins.—We compared the reconstructed ancient protein sequences from the Dmanisi *Homo erectus* and Atapuerca *Homo antecessor* hominins with protein sequences from great apes[41,43], three Neanderthals[31,40,55], a Denisovan[56] and a panel of present-day humans, including 256 samples from the Simons Genome Diversity Panel (SGDP)[57] and 41 high-coverage individuals from the 1000 Genomes Project[58]. Altogether, our reference data represents worldwide human and great ape variation data (Tabs. S7, S8). Additionally, we included protein sequences from macaque (*Macaca mulatta*) and gibbon (*Nomascus leucogenys*) to root phylogenetic trees. The protein sequences were retrieved from the UniProt database or reconstructed from the reference whole-genome sequences as described in the supplementary methods.

The ancient and reference protein sequences were aligned using *mafft*[59]. We aligned the sequences of each protein separately and obtained an alignment for each of the ancient individuals independently (Tab. S9). The isobaric amino acids leucine (L) and isoleucine (I) cannot be distinguished with the experimental procedure used for this study. Therefore, we have to take the following precautions to avoid unintentional sequence differences. If, at a

specific amino acid position, either I or L were present in the reference protein sequences, we replace all corresponding amino acids in the ancient protein sequences to the amino acid that is present. Alternatively, if both amino acids are present in the reference protein sequence, we replace all I to L for all sequences. We used sequence information for seven proteins (ALB, AMBN, AMELX, AMELY, COL17α1, ENAM and MMP20) for the *Homo antecessor* individual and six proteins for the *Homo erectus* individual (ALB, AMBN, AMELX, COL17α1, ENAM and MMP20) with a total of 22.08% and 22.14% non-missing sites, respectively (Tab. S9). We were able to recover a unique SAP for *Homo antecessor*, however, for *Homo erectus* no unique SAP was detected (Tabs. S9–11; Figs. S10–12).

Phylogenetic reconstruction.—We sought to build phylogenetic trees using the aforementioned protein sequence alignments following three different approaches: a maximum likelihood (ML) approach, using PhyML v3[60], and two Bayesian approaches, using mrBayes[61] and BEAST[62].

Maximum-likelihood approach.: We built ML trees for each protein independently and for a concatenated alignment consisting of all of the available protein sequences for each of the ancient samples (Figs. S13, S14). We used PhyML v3 and the parameters described in the supplementary section 2.3.5a to build and optimize the tree topologies, branch length and substitutions rates for each of the alignments. Support for each bipartition was obtained based on 100 non-parametric bootstrap replicates. We evaluated the effect of significant missingness in the ancient samples on the inferred topology. Finally, we looked at the effect of varying which of the subset of present-day human samples was included in the tree (Supplementary section 2.3.5b, c).

Bayesian approach using mrBayes.: To assess the robustness of the ML inference results, we performed Bayesian phylogenetic inference based on the concatenated alignments using *mrBayes* 3.2 and the parameters described in the supplementary section 2.3.5d (Fig. S16; Extended Data Fig. 8). Bayesian inference was performed using the CIPRES Science Gateway[63].

Bayesian approach using BEAST.: We used BEAST 2.5 to obtain a time calibrated tree for the seven proteins used for *Homo antecessor*. For this analysis, we used a concatenated alignments including the Neanderthals, the Denisovan, seven randomly chosen *Homo sapiens* individuals, and a single individual per great ape species. The alignment was partitioned by gene and a coalescent constant population model was used for the tree prior. The ages of the ancient samples included in the analysis (Vindija Neanderthal: 52 ka[55], Altai Neanderthal: 112 ka[31], Denisovan: 72 ka[56] and *Homo antecessor* 860.5 ka[11]) were used as tip dates for calibration. For each partition, we used the JTT substitution model with four categories for the gamma parameter, for which we allowed the MCMC chain to sample the shape of the gamma distribution (with an exponentially distributed prior) and assigned independent clock models. Additionally, we set a prior for the divergence time of great apes to 23.85 ± 2.5 Ma (normally distributed)[64], and rooted the tree using the macaque (*Macaca mulatta*). The overall topology of the tree was estimated for the seven partitions jointly. The convergence of the algorithm was assessed using Tracer v1.7.0[65]. Finally, we repeated this

analysis with 100 alignments, each of them consisting of seven different present-day humans chosen randomly. While the topology within the clade consisting of present-day humans, Neanderthals and Denisovan (HND) was not consistent across the replicates, 99 of the replicates consistently place the *Homo antecessor* sequence as an outgroup to the HND clade (Fig. 2a).

Further details on phylogenetic analysis and results can be found in the Supplementary Information. Example MS/MS spectra from the MaxQuant search and overlapping sites of phylogenetic interest (SAPs) are included in the file "Key MS-MS Spectra.pdf" for both hominins.

Reporting summary

Further information on research design is available in the Nature

Research Reporting Summary linked to this paper.

Data Deposition

Mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the data set identifier PXD014342. Generated ancient protein consensus sequences used for phylogenetic analysis for *Homo antecessor* (Atapuerca) and *Homo erectus* (Dmanisi) hominins can be found in the supplementary file "Hominin SI File2.txt"), which is formatted as a .fasta file. Full protein sequence alignments used during phylogenetic analysis can be accessed via Figshare (https://doi/10.6084/m9.figshare.9927074). Amino acid racemization data is available online through the NOAA database. The wiNNer model can be accessed on GitHub (https://github.com/cox-labs/wiNNer.git).
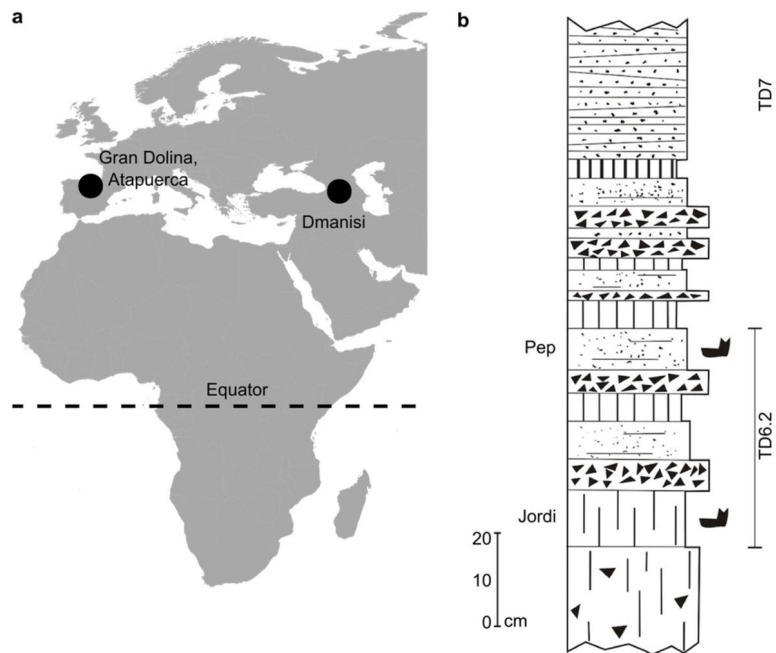
Extended Data

Extended Data Figure 1. Location and stratigraphy of the hominin fossils studied.
a, Geographic location of Gran Dolina, Sierra de Atapuerca (Spain) and Dmanisi (Georgia). Base map was generated using public domain data from www.naturalearthdata.com. **b,** Summarized stratigraphic profile of Gran Dolina, Sierra de Atapuerca, including the location of hominin fossils in sublayers "Pep" and "Jordi" of unit TD6.2.
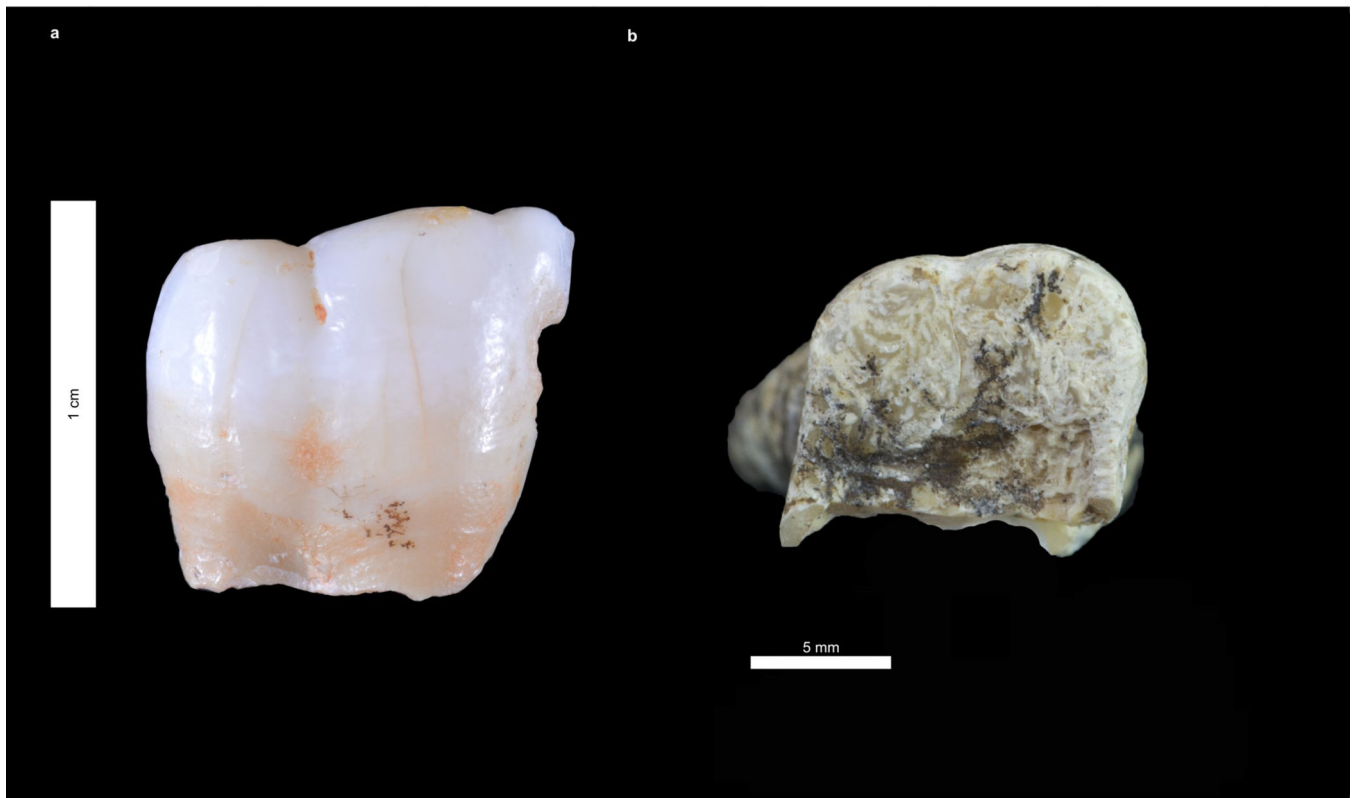
Extended Data Figure 2. Hominin specimens studied.
a, Specimen ATD6–92 from Gran Dolina, Atapuerca (Spain), in buccal view. The fragment represents a portion of a permanent lower left first or second molar. **b,** Specimen D4163 from Dmanisi (Georgia), in occlusal view. The specimen is a fragmented right upper first molar. Note differences in scale bar between **a** and **b**.
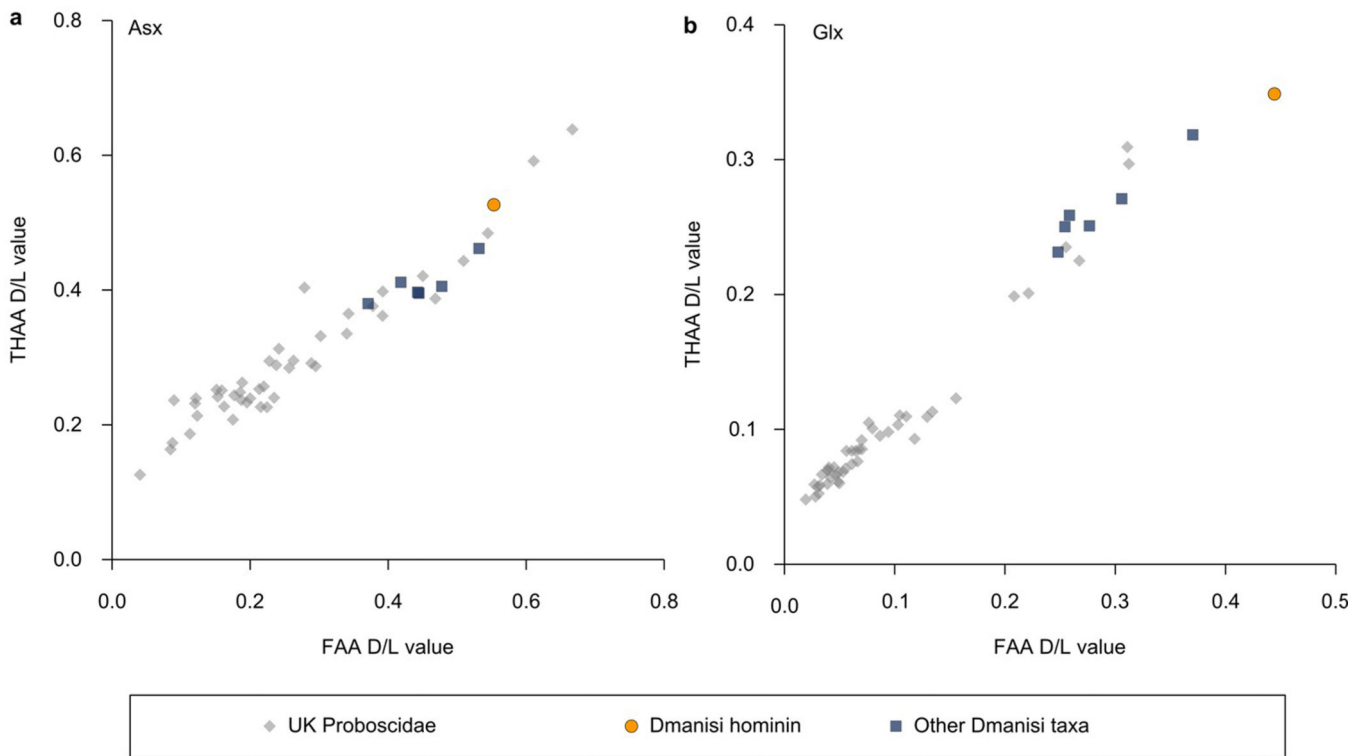
Extended Data Figure 3. Amino acid racemization of D4163 (*Homo erectus* from Dmanisi).
The extent of intra-crystalline racemization in enamel for the free amino acid (FAA, x-axis)
fraction and the total hydrolysable amino acids (THAA, y-axis) fraction for aspartic acid
plus asparagine (here denoted Asx, **a**), and glutamic acid plus glutamine (here denoted Glx,
b), demonstrates endogenous amino acids breaking down within a closed system. The
hominin value is displayed in relation to values for enamel samples from other fauna from
Dmanisi[6] (blue squares) and a range of UK Pleistocene and Pliocene Proboscidea obtained
previously[37] (grey diamonds). Fauna species are shown for comparison, but different rates in
their protein breakdown mean that they will show different extents of racemization. Note
differences in x- and y-axis scales.
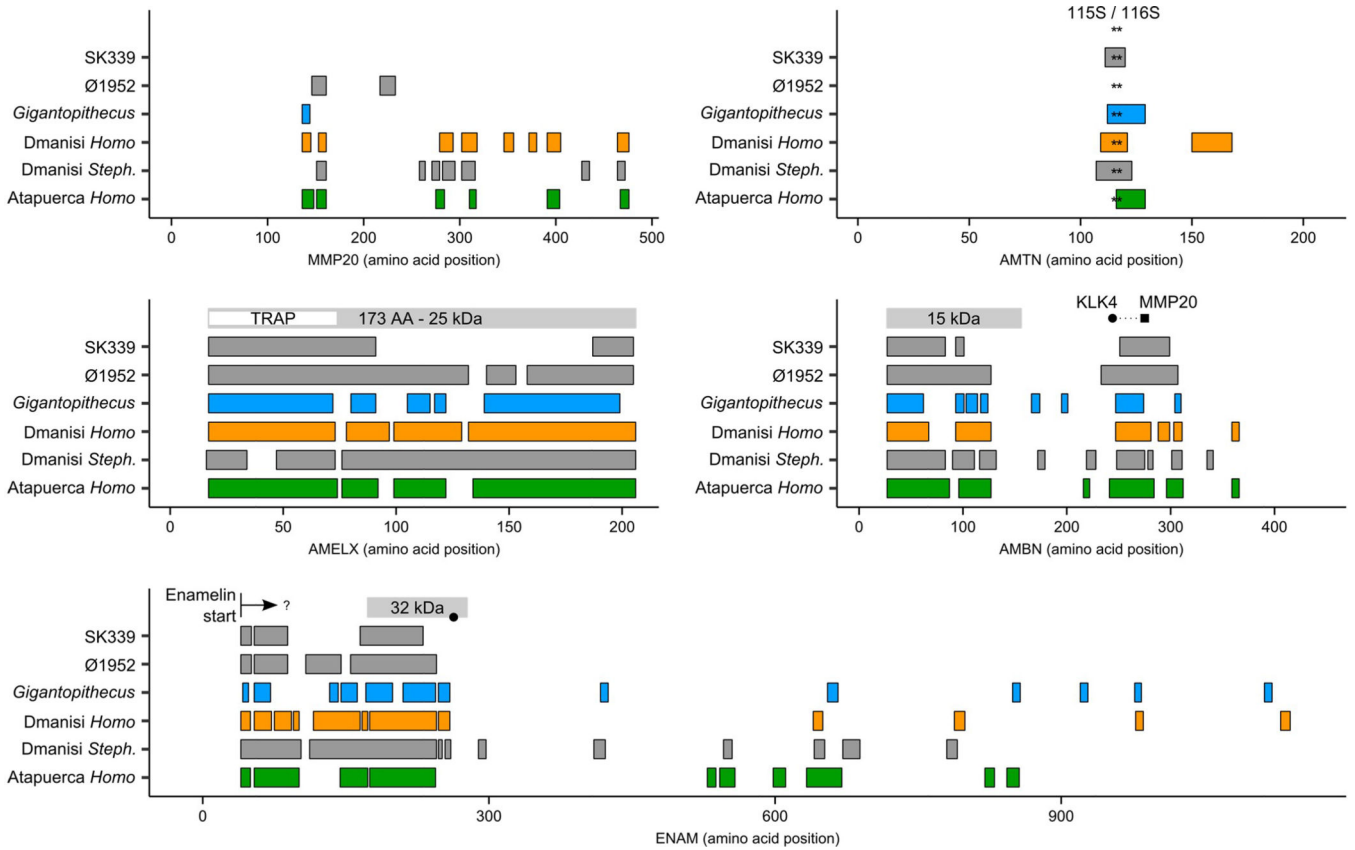
Extended Data Figure 4. Sequence coverage for five enamel-specific proteins across Pleistocene samples and recent human controls.

For each protein, the bars span protein positions covered, with positions remapped to the human reference proteome. The top row indicates the position of a selection of known MMP20 and KLK4 cleavage products of the enamel-specific proteins AMELX[52], AMBN[49], and ENAM[53]. Several *in vivo* proteolytic degradation fragments of ENAM share the same N-terminus, but have unknown C-termini[50]. Dotted line for AMBN indicates a putative cleavage product based on known MMP20 (squares) and KLK4 (circles) *in vivo* cleavage positions. For AMTN, serines (S) at positions 115 and 116 (indicated by asterisks, *) are conserved amongst vertebrates and involved in mineral-binding,[21]. Additional cleavage products and MMP20/KLK4 cleavage sites are known in all enamel-specific proteins. SK339[16] and Ø1952 represent two recent human control samples (see Methods). *Steph.* = *Stephanorhinus*[6]. TRAP = tyrosine-rich amelogenin polypeptide. AA = amino acids. kDa = kilodalton.
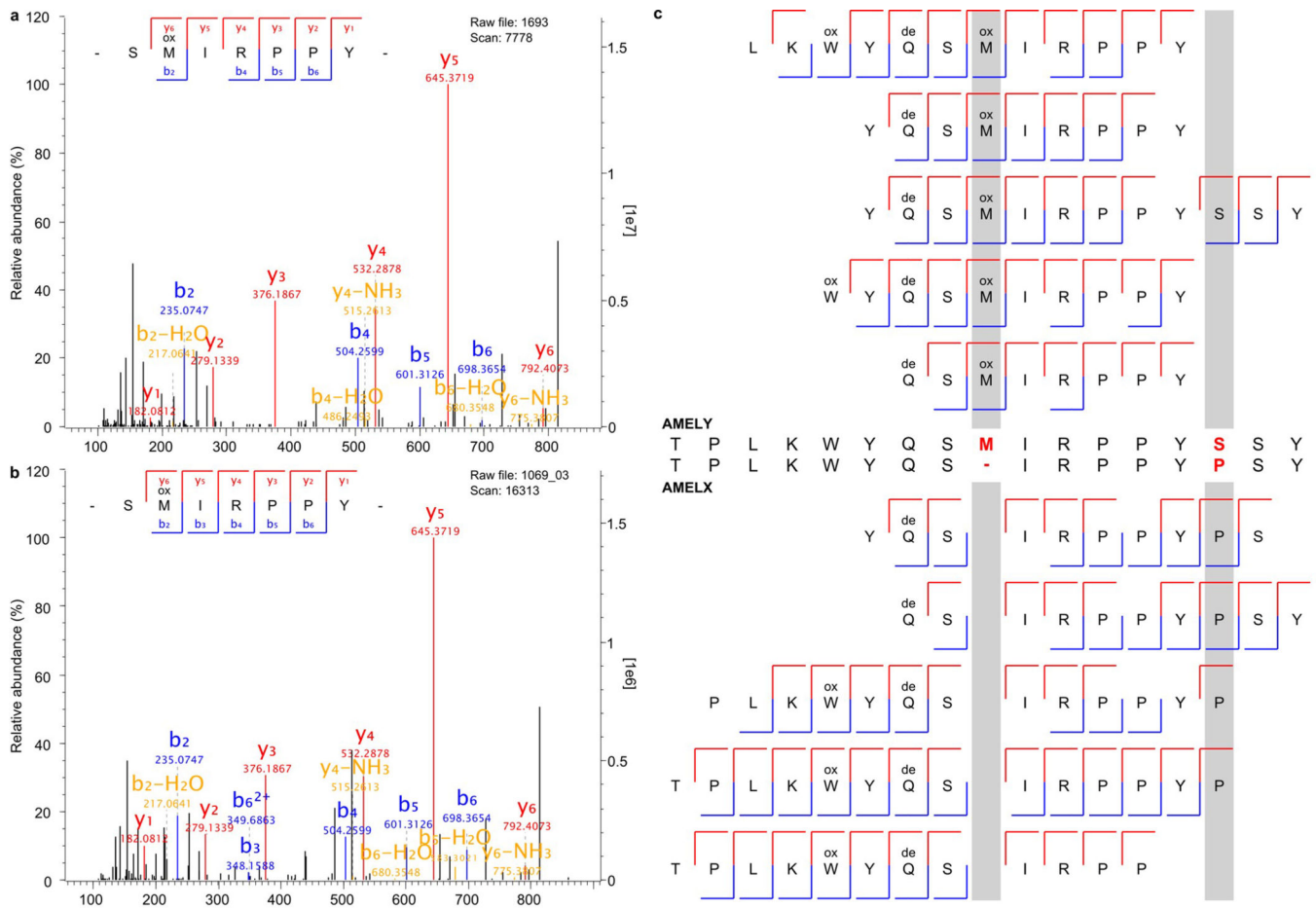
Extended Data Figure 5. *Homo antecessor* **specimen ATD6–92 represents a male hominin.**
a, AMELY-specific peptide from the recent human control Ø1952. **b**, The same AMELY-specific peptide from *Homo antecessor*. **c**, Alignment of a selection of AMELY- and AMELX-specific peptide fragment ion series deriving from *Homo antecessor*. The alignment stretches along AMELX_HUMAN isoform 1, positions 37 to 52 only (AMELX: Uniprot accession Q99217; AMELY: Uniprot accession Q99218). See Figure S5 for another example of an AMELY-specific MS2 spectrum.
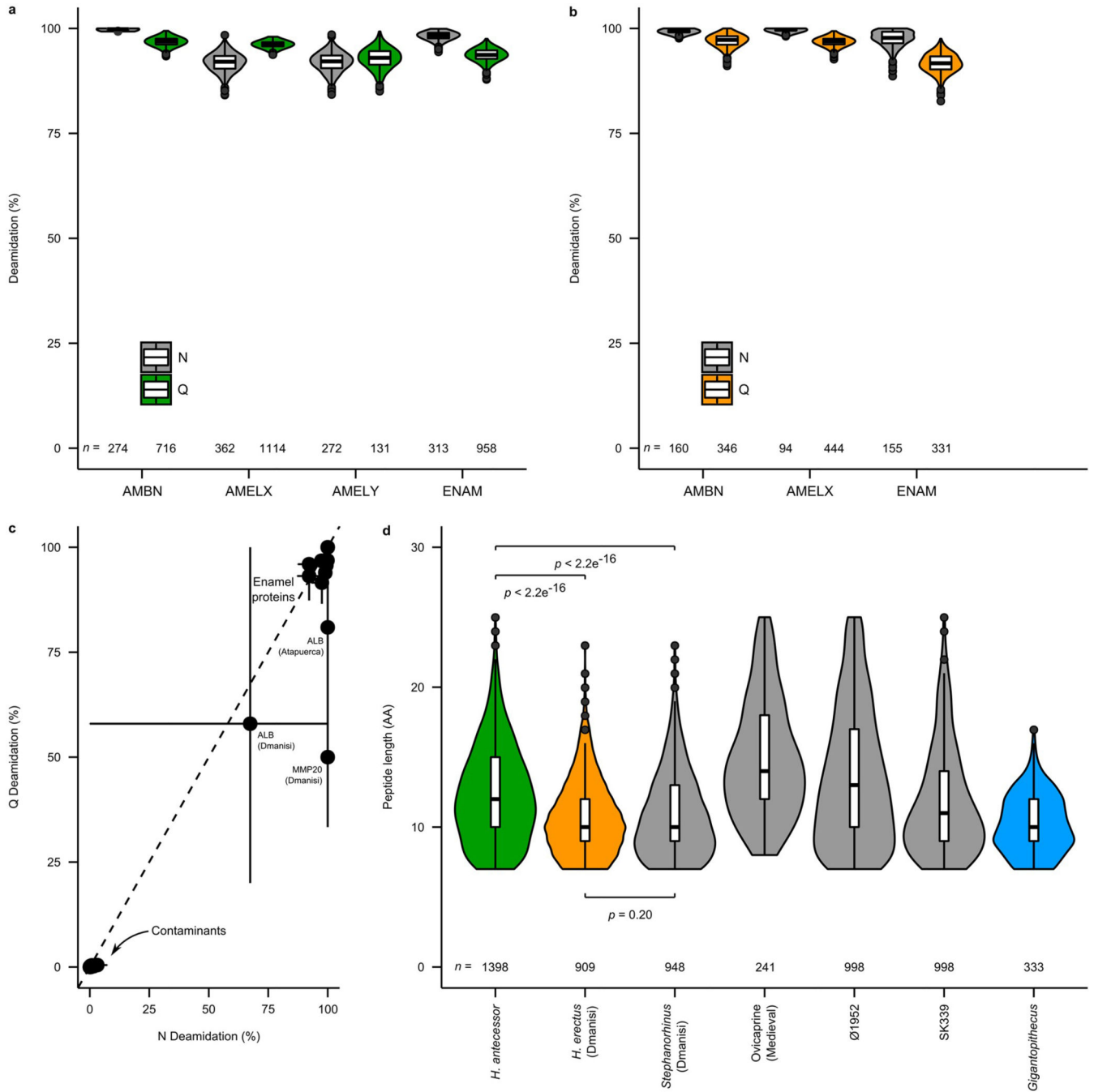
Extended Data Figure 6. Enamel proteome damage.
Glutamine (Q) and asparagine (N) deamidation of enamel-specific proteins from *Homo antecessor* (Atapuerca, **a**), and *Homo erectus* (Dmanisi, **b**). Values based on 1,000 bootstrap replications of protein deamidation. **c,** Relation between mean asparagine (N) and glutamine (Q) deamidation for all proteins in both the Atapuerca and Dmanisi hominin datasets. Error bars represent 95% CI interval window of 1,000 bootstrap replications of protein deamidation. Dashed line is x=y. **d,** Peptide length distribution of *Homo antecessor* (Atapuerca), *Homo erectus* (Dmanisi), four previously published enamel proteomes[6,8,16],

and one additional human Medieval control sample (Ø1952). For **a**, **b**, and **d**, the number of peptides (*n*) is given for each vioplot. The boxplots within define the range of the data (whiskers extending to 1.5x the interquartile range), outliers (black dots, beyond 1.5x the interquartile range), 25th and 75th percentiles (boxes), and medians (centre lines). *P*-values of two-sided *t*-tests conducted between sample pairs are indicated. No independent replication of these experiments was performed.
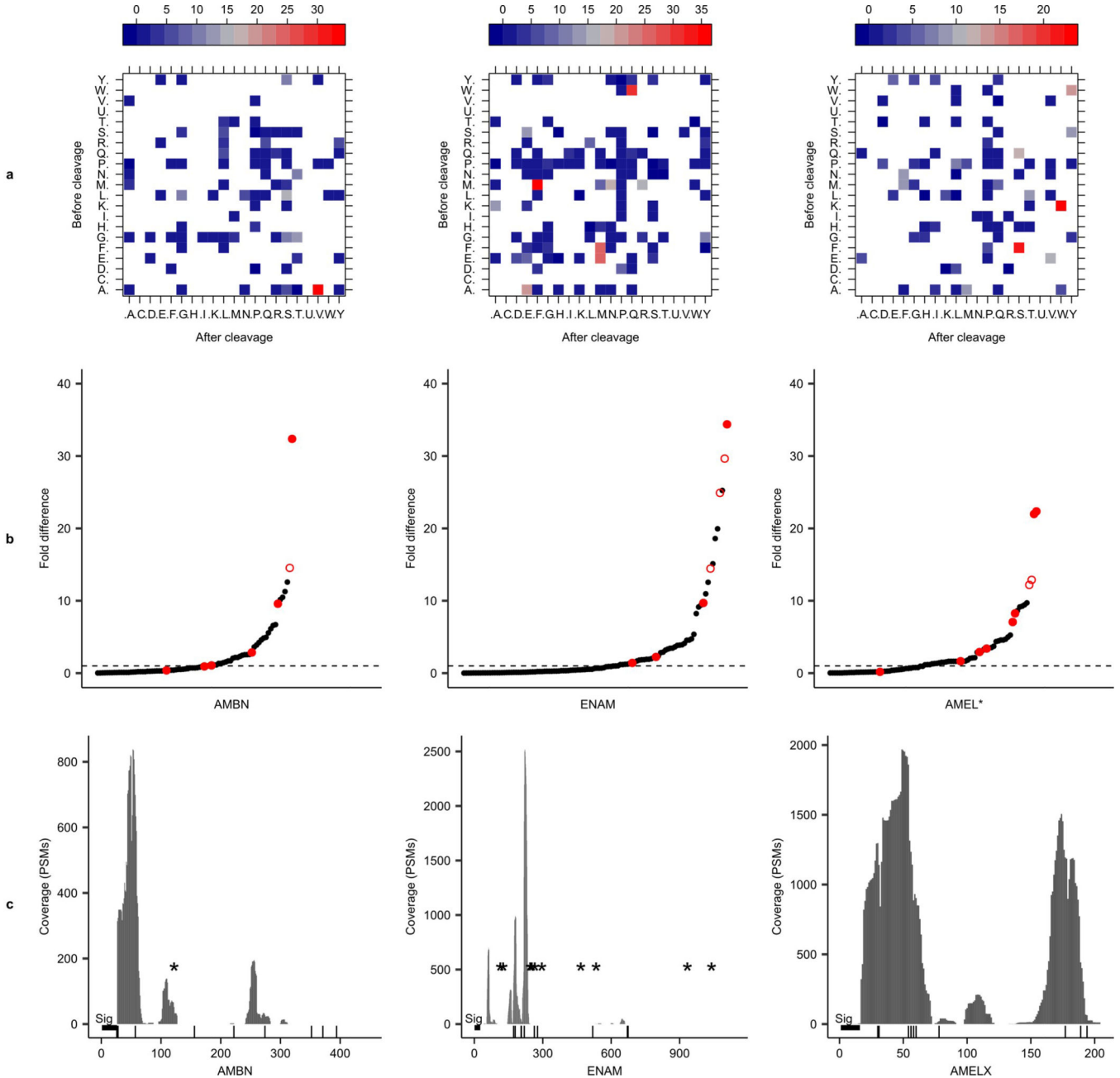
Extended Data Figure 7. Survival of *in vivo* MMP20 and KLK4 cleavage sites in the Atapuerca enamel proteome.

a, Experimentally observed cleavage matrices for ameloblastin (AMBN), enamelin (ENAM), and amelogenin (AMELX+AMELY; see Methods). Fold differences are color-coded by comparing observed PSM cleavage frequencies to a random cleavage matrix for each protein separately[7]. **b,** Fold differences for all observed cleavage pairs per protein. Red filled circles represent MMP20, KLK4 and signal peptide cleavage sites mentioned in the literature[50–53]. Red open circles indicate cleavage sites located up to two amino acid positions away from such sites. **c,** Peptide-spectrum-matches (PSM) coverage for each protein. The signal peptide (thick horizontal bar labelled "Sig."), known MMP20 and KLK4 cleavage sites (vertical bars), and O- and N-linked glycosylation sites (asterisks) are also indicated. For AMELX, peptide positions for all three known isoforms where remapped to the coordinates of isoform 3, which represents the longest isoform (UniProt accession Q99217–3). Note differences in x- and y-axis between the three panels of **c.**
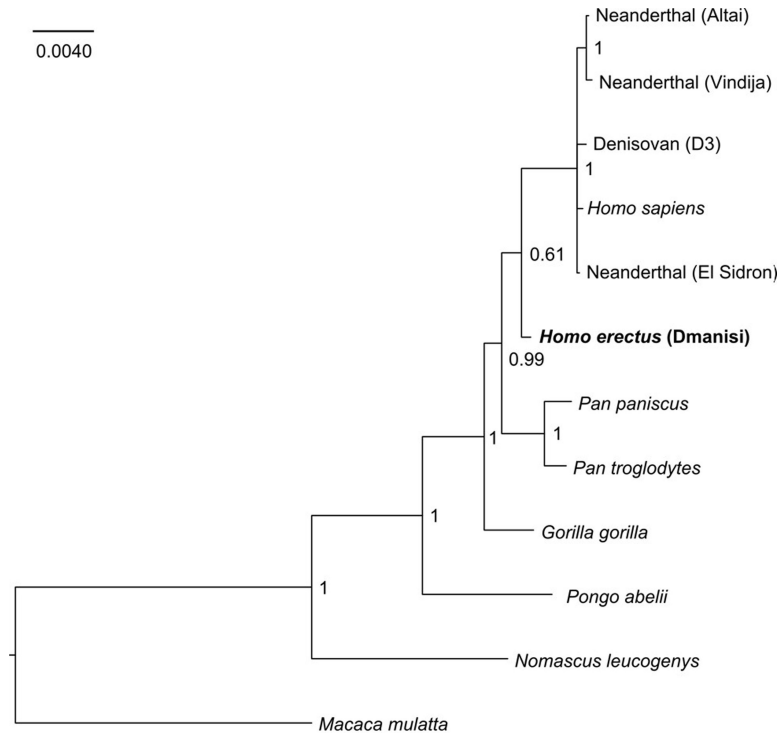
Extended Data Figure 8. Phylogenetic position of *Homo erectus* (D4163, Dmanisi) through Bayesian analysis.

Nomascus leucogenys and *Macaca mulatta* were used as outgroups.

Extended Data Table 1.

Extraction and mass spectrometry details of analyses conducted on both ancient hominin specimens.

Stage Tip number	Tissue	Protein extraction method*	Mass Spectrometer	Mass Spectrometer location	Replicates
Homo antecessor, **specimen ATD6–92, Atapuerca**					

Stage Tip number	Tissue	Protein extraction method[*]	Mass Spectrometer	Mass Spectrometer location	Replicates
1069	Enamel	1	QE-HF	Copenhagen	4
1069	Enamel	1	Fusion Lumos	Barcelona	1
Homo erectus, **specimen D4163, Dmanisi**					
1138	Enamel	1	QE-HF	Copenhagen	2
1141	Enamel	2	QE-HF	Copenhagen	2
1138	Enamel	1	Fusion Lumos	Barcelona	1
1141	Enamel	2	Fusion Lumos	Barcelona	1
1139	Dentine	1	QE-HF	Copenhagen	2
1142	Dentine	2	QE-HF	Copenhagen	2
1139	Dentine	1	Fusion Lumos	Barcelona	1
1142	Dentine	2	Fusion Lumos	Barcelona	1
1386	Enamel	1	QE-HF	Copenhagen	1
1387	Enamel	3	QE-HF	Copenhagen	1
1388	Enamel	1	QE-HF	Copenhagen	1

QE-HF = Q Exactive™ HF hybrid quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific). Fusion Lumos = LTQ-Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific).

[*] Extraction method 1: demineralization in HCl, no subsequent proteolytic digestion. Extraction method 2: demineralization in HCl, alkylation, and digestion with LysC+Trypsin. Extraction method 3: demineralization in TFA, no subsequent proteolytic digestion. See SI for further details.

Extended Data Table 2.

Ancient hominin enamel proteome composition and coverage. Proteins are included only if two or more unique peptides were observed in either the PEAKS or MaxQuant (MQ) searches. Primary accession refers to the Homo sapiens entry in UniProt. Protein sequence coverage in the final column indicates the coverage obtained after combining PEAKS and MaxQuant peptide recovery. For coverage (AA) columns, numbers in brackets refer to the number of amino acid positions uniquely identified in PEAKS or MaxQuant searches. For AMELX and AMELY, coverage statistics combine counts for all isoforms present, while peptide counts only refer to the highest-ranking isoform or database entry. Direct comparisons between PEAKS and MaxQuant are uninformative due to fundamental differences in spectral identification, protein/peptide assignment, and peptide counting approaches.

Protein	Primary accession	MaxQuant				PEAKS				Combined Coverage (%)
		Peptides	Unique peptides	Coverage (AA)	Coverage (%)	Peptides	Unique peptides	Coverage (AA)	Coverage (%)	
Homo antecessor, **specimen ATD6-92, Atapuerca**										
AMELX	Q99217[*]	527	527	170 (0)	82.9	737	12	171 (1)	83.4	83.4
AM ELY	Q99218[*]	220	86	131 (0)	63.6	341	6	141 (10)	68.4	68.4
AMBN	Q9NP70[*]	289	289	160 (3)	35.8	351	350	166 (9)	37.1	37.8
AMTN	Q6UX39	4	4	14	6.7	5	5	14	6.7	6.7

Protein	Primary accession	MaxQuant				PEAKS				Combined Coverage (%)
		Peptides	Unique peptides	Coverage (AA)	Coverage (%)	Peptides	Unique peptides	Coverage (AA)	Coverage (%)	
ENAM	Q9NRM1	424	424	233 (18)	20.4	586	586	245 (32)	21.5	23.0
MMP20	O60882	12	12	65 (0)	13.5	14	14	66 (1)	13.7	13.7
ALB	P02768	11	11	69 (17)	11.3	12	7	76 (24)	12.5	15.3
COL1a1	P02452	17	17	34 (21)	2.3	15	15	29 (16)	2.0	3.4
COL1a2	P08123	1	1	23	1.7	2	2	23	1.7	1.7
COL17a1	Q9UMD9	27	27	96 (24)	6.4	42	42	88 (16)	5.9	7.5
***Homo erectus,* specimen D4163, Dmanisi**										
AMELX	Q99217 *	357	357	182 (9)	88.8	297	297	173 (0)	84.4	88.8
AMBN	Q9NP70 *	219	219	123 (1)	27.5	182	182	139 (17)	31.1	31.3
AMTN	Q6UX39	6	6	31 (13)	15.3	1	1	18 (0)	9.1	14.8
ENAM	Q9NRM1	306	306	224 (78)	19.6	293	293	160 (14)	14.0	20.8
MMP20	O60882	13	13	90 (15)	18.6	16	16	84 (9)	17.4	20.5
ALB	P02768	33	33	216 (12)	35.5	41	28	233 (29)	38.3	40.2
COL1a1	P02452	10	10	202 (44)	13.8	17	17	414 (256)	28.3	31.3
COL1a2	P08123	9	9	130 (3)	9.5	11	11	197 (66)	14.6	14.6
COL17a1	Q9UMD9	10	10	67 (45)	4.5	1	1	22 (0)	1.5	4.5

AA = amino acids.

*
Combined coverage calculated against the longest isoforms for each protein.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

REFERENCES

1. Gabunia L et al. Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: taxonomy, geological setting, and age. Science 288, 1019–1025, doi:10.1126/science.288.5468.1019 (2000). [PubMed: 10807567]

2. Zhu Z et al. Hominin occupation of the Chinese Loess Plateau since about 2.1 million years ago. Nature 559, 608–612, doi:10.1038/s41586-018-0299-4 (2018). [PubMed: 29995848]

3. Stringer C The origin and evolution of Homo sapiens. Philosophical Transactions of the Royal Society B: Biological Sciences 371, doi:10.1098/rstb.2015.0237 (2016).

4. Hublin JJ The origin of Neandertals. Proceedings of the National Academy of Sciences 106, 16022, doi:10.1073/pnas.0904119106 (2009).

5. Rightmire G Human evolution in the Middle Pleistocene: the role of Homo heidelbergensis. Evolutionary anthropology 6, 218–227, doi:10.1002/(SICI)1520-6505(1998)6:6<218::AID-EVAN4>3.0.CO;2-6 (1998).

6. Cappellini E et al. Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny. Nature, doi:10.1038/s41586-019-1555-y (2019).

7. Chen F et al. A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. Nature 569, 409–412, doi:10.1038/s41586-019-1139-x (2019). [PubMed: 31043746]

8. Welker F et al. Enamel proteome shows that Gigantopithecus was an early diverging pongine. Nature 576, 262–265, doi:10.1038/s41586-019-1728-8 (2019). [PubMed: 31723270]

9. Bermúdez de Castro JM et al. A Hominid from the Lower Pleistocene of Atapuerca, Spain: Possible Ancestor to Neandertals and Modern Humans. Science 276, 1392–1395, doi:10.1126/science.276.5317.1392 (1997). [PubMed: 9162001]

10. Carbonell E et al. Lower Pleistocene hominids and artifacts from Atapuerca-TD6 (Spain). Science 269, 826–830, doi:10.1126/science.7638598 (1995). [PubMed: 7638598]

11. Duval M et al. The first direct ESR dating of a hominin tooth from Atapuerca Gran Dolina TD-6 (Spain) supports the antiquity of Homo antecessor. Quaternary Geochronology 47, 120–137, doi:10.1016/j.quageo.2018.05.001 (2018).

12. Freidline SE, Gunz P, Harvati K & Hublin J-J Evaluating developmental shape changes in Homo antecessor subadult facial morphology. Journal of Human Evolution 65, 404–423, doi:10.1016/j.jhevol.2013.07.012 (2013). [PubMed: 23998458]

13. Lacruz RS et al. Facial Morphogenesis of the Earliest Europeans. PLOS ONE 8, e65199, doi:10.1371/journal.pone.0065199 (2013).

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

15. Lordkipanidze D et al. A complete skull from Dmanisi, Georgia, and the evolutionary biology of early Homo. Science 342, 326–331, doi:10.1126/science.1238484 (2013). [PubMed: 24136960]

16. Stewart NA, Gerlach RF, Gowland RL, Gron KJ & Montgomery J Sex determination of human remains from peptides in tooth enamel. Proceedings of the National Academy of Sciences of the United States 114, 13649–13654, doi:10.1073/pnas.1714926115 (2017).

17. Tiwary S et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nature Methods 16, 519–525, doi:10.1038/s41592-019-0427-6 (2019). [PubMed: 31133761]

18. Castiblanco GA et al. Identification of proteins from human permanent erupted enamel. European Journal of Oral Sciences 123, 390–395, doi:10.1111/eos.12214 (2015). [PubMed: 26432388]

19. Asaka T et al. Type XVII Collagen is a Key Player in Tooth Enamel Formation. The American Journal of Pathology 174, 91–100, doi:10.2353/ajpath.2009.080573 (2009). [PubMed: 19036806]

20. Porto IM, Laure HJ, de Sousa FB, Rosa JC & Gerlach RF New techniques for the recovery of small amounts of mature enamel proteins. Journal of Archaeological Science 38, 3596–3604, doi:10.1016/j.jas.2011.08.030 (2011).

21. Gasse B, Chiari Y, Silvent J, Davit-Béal T & Sire J-Y Amelotin: an enamel matrix protein that experienced distinct evolutionary histories in amphibians, sauropsids and mammals. BMC Evolutionary Biology 15, 47, doi:10.1186/s12862-015-0329-x (2015). [PubMed: 25884299]

22. Demarchi B et al. Protein sequences bound to mineral surfaces persist into deep time. eLife 5, e17092, doi:10.7554/eLife.17092 (2016).

23. Tagliabracci VS et al. Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. Science 336, 1150–1153, doi:10.1126/science.1217817 (2012). [PubMed: 22582013]

24. Hu JCC, Yamakoshi Y, Yamakoshi F, Krebsbach PH & Simmer JP Proteomics and Genetics of Dental Enamel. Cells Tissues Organs 181, 219–231, doi:10.1159/000091383 (2005). [PubMed: 16612087]

25. Glimcher MJ, Cohen-Solal L, Kossiva D & de Ricqles A Biochemical analyses of fossil enamel and dentin. Paleobiology 16, 219–232, doi:10.1017/S0094837300009891 (1990).

26. Wagner GA et al. Radiometric dating of the type-site for Homo heidelbergensis at Mauer, Germany. Proceedings of the National Academy of Sciences 107, 19726–19730, doi:10.1073/pnas.1012722107 (2010).

27. Martinón-Torres M et al. Dental Evidence on the Hominin Dispersals during the Pleistocene. Proceedings of the National Academy of Sciences of the United States of America 104, 13279–13282, doi:10.1073/pnas.0706152104 (2007).

28. Bermúdez de Castro JM, Martinón-Torres M, Arsuaga JL & Carbonell E Twentieth anniversary of Homo antecessor (1997–2017): a review. Evolutionary Anthropology: Issues, News, and Reviews 26, 157–171, doi:10.1002/evan.21540 (2017).

29. Gómez-Robles A, Bermúdez de Castro JM, Arsuaga J-L, Carbonell E & Polly PD No known hominin species matches the expected dental morphology of the last common ancestor of Neanderthals and modern humans. Proceedings of the National Academy of Sciences 110, 18196–18201, doi:10.1073/pnas.1302653110 (2013).

30. Meyer M et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature 531, 504–507, doi:10.1038/nature17405 (2016). [PubMed: 26976447]

31. Prüfer K et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505, 43–49, doi:10.1038/nature12886 (2013). [PubMed: 24352235]

32. Lacruz RS et al. The evolutionary history of the human face. Nature Ecology & Evolution 3, 726–736, doi:10.1038/s41559-019-0865-7 (2019). [PubMed: 30988489]

33. Welker F et al. Middle Pleistocene protein sequences from the rhinoceros genus Stephanorhinus and the phylogeny of extant and extinct Middle/Late Pleistocene Rhinocerotidae. PeerJ 5, e3033, doi:10.7717/peerj.3033 (2017).

34. Hill RC et al. Preserved Proteins from Extinct Bison latifrons Identified by Tandem Mass Spectrometry; Hydroxylysine Glycosides are a Common Feature of Ancient Collagen. Molecular & Cellular Proteomics 14, 1946–1958, doi:10.1074/mcp.M114.047787 (2015). [PubMed: 25948757]

35. Wadsworth C & Buckley M Proteome degradation in fossils: investigating the longevity of protein survival in ancient bone. Rapid Communications in Mass Spectrometry 28, 605–615, doi:10.1002/rcm.6821 (2014). [PubMed: 24519823]

REFERENCES

36. Penkman KEH, Kaufman DS, Maddy D & Collins MJ Closed-system behaviour of the intra-crystalline fraction of amino acids in mollusc shells. Quaternary Geochronology 3, 2–25, doi:10.1016/j.quageo.2007.07.001 (2008). [PubMed: 19684879]

37. Dickinson M, Lister AM & Penkman KEH A new method for enamel amino acid racemization dating: a closed system approach. Quaternary Geochronology 50, 29–46, doi:10.1016/j.quageo.2018.11.005 (2019).

38. Hill RL Hydrolysis of proteins. Advances in Protein Chemistry 20, 37–107 (1965). [PubMed: 5334827]

39. Mackie M et al. Palaeoproteomic Profiling of Conservation Layers on a 14th Century Italian Wall Painting. Angewandte Chemie (International ed.) 57, 7369–7374, doi:10.1002/anie.201713020 (2018).

40. Castellano S et al. Patterns of coding variation in the complete exomes of three Neandertals. Proceedings of the National Academy of Sciences 111, 6666–6671, doi:10.1073/pnas.1405138111 (2014).

41. de Manuel M et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science 354, 477–481, doi:10.1126/science.aag2602 (2016). [PubMed: 27789843]

42. Nater A et al. Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species. Current Biology 27, 3487–3498, doi:10.1016/j.cub.2017.11.020 (2017). [PubMed: 29103940]

43. Prado-Martinez J et al. Great ape genetic diversity and population history. Nature 499, 471–475, doi:10.1038/nature12228 (2013). [PubMed: 23823723]

44. Hanson-Smith V & Johnson A PhyloBot: A Web Portal for Automated Phylogenetics, Ancestral Sequence Reconstruction, and Exploration of Mutational Trajectories. PLoS computational biology 12, e1004976, doi:10.1371/journal.pcbi.1004976 (2016).

45. Welker F Elucidation of cross-species proteomic effects in human and hominin bone proteome identification through a bioinformatics experiment. BMC Evolutionary Biology 18, 23, doi:10.1186/s12862-018-1141-1 (2018). [PubMed: 29463217]

46. Hendy J et al. A guide to ancient protein studies. Nature Ecology & Evolution 2, 791–799, doi:10.1038/s41559-018-0510-x (2018). [PubMed: 29581591]

47. Zhang J et al. PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. Molecular and Cellular Proteomics 11, M111.010587, doi:10.1074/mcp.M111.010587 (2012).

48. Cox J & Mann M MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology 26, 1367–1372, doi:10.1038/nbt.1511 (2008).

49. Chun YHP et al. Cleavage Site Specificity of MMP-20 for Secretory-stage Ameloblastin. Journal of Dental Research 89, 785–790, doi:10.1177/0022034510366903 (2010). [PubMed: 20400724]

50. Yamakoshi Y, Hu JCC, Fukae M, Yamakoshi F & Simmer JP How do enamelysin and kallikrein 4 process the 32-kDa enamelin? European Journal of Oral Sciences 114, 45–51, doi:10.1111/j.1600-0722.2006.00281.x (2006). [PubMed: 16674662]

51. Iwata T et al. Processing of Ameloblastin by MMP-20. Journal of Dental Research 86, 153–157, doi:10.1177/154405910708600209 (2007). [PubMed: 17251515]

52. Nagano T et al. Mmp-20 and Klk4 Cleavage Site Preferences for Amelogenin Sequences. Journal of Dental Research 88, 823–828, doi:10.1177/0022034509342694 (2009). [PubMed: 19767579]

53. Fukae M et al. Primary Structure of the Porcine 89-kDa Enamelin. Advances in Dental Research 10, 111–118, doi:10.1177/08959374960100020201 (1996). [PubMed: 9206327]

54. Colaert N, Helsens K, Martens L, Vandekerckhove J & Gevaert K Improved visualization of protein consensus sequences by iceLogo. Nature Methods 6, 786–787, doi:10.1038/nmeth1109-786 (2009). [PubMed: 19876014]

55. Prüfer K et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science 358, 655–658, doi:10.1126/science.aao1887 (2017). [PubMed: 28982794]

56. Meyer M et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338, 222–226, doi:10.1126/science.1224344 (2012). [PubMed: 22936568]

57. Mallick S et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538, 201–206, doi:10.1038/nature18964 (2016). [PubMed: 27654912]

58. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526, 68–74, doi:10.1038/nature15393 (2015). [PubMed: 26432245]

59. Katoh K, Misawa K, Kuma K. i. & Miyata T MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research 30, 3059–3066, doi:10.1093/nar/gkf436 (2002). [PubMed: 12136088]

60. Guindon S et al. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59, 307–321, doi:10.1093/sysbio/syq010 (2010). [PubMed: 20525638]

61. Ronquist F et al. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Systematic Biology 61, 539–542, doi:10.1093/sysbio/sys029 (2012). [PubMed: 22357727]

62. Bouckaert R et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS computational biology 15, e1006650, doi:10.1371/journal.pcbi.1006650 (2019).

63. Miller MA, Pfeiffer W & Schwartz T in Gateway Computing Environments Workshop (GCE) 1–8 (New Orleans, 2010).

64. Besenbacher S, Hvilsom C, Marques-Bonet T, Mailund T & Schierup MH Direct estimation of mutations in great apes reconciles phylogenetic dating. Nature Ecology & Evolution 3, 286–292, doi:10.1038/s41559-018-0778-x (2019). [PubMed: 30664699]

65. Rambaut A, Drummond AJ, Xie D, Baele G & Suchard MA Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Systematic biology 67, 901–904, doi:10.1093/sysbio/syy032 (2018). [PubMed: 29718447]
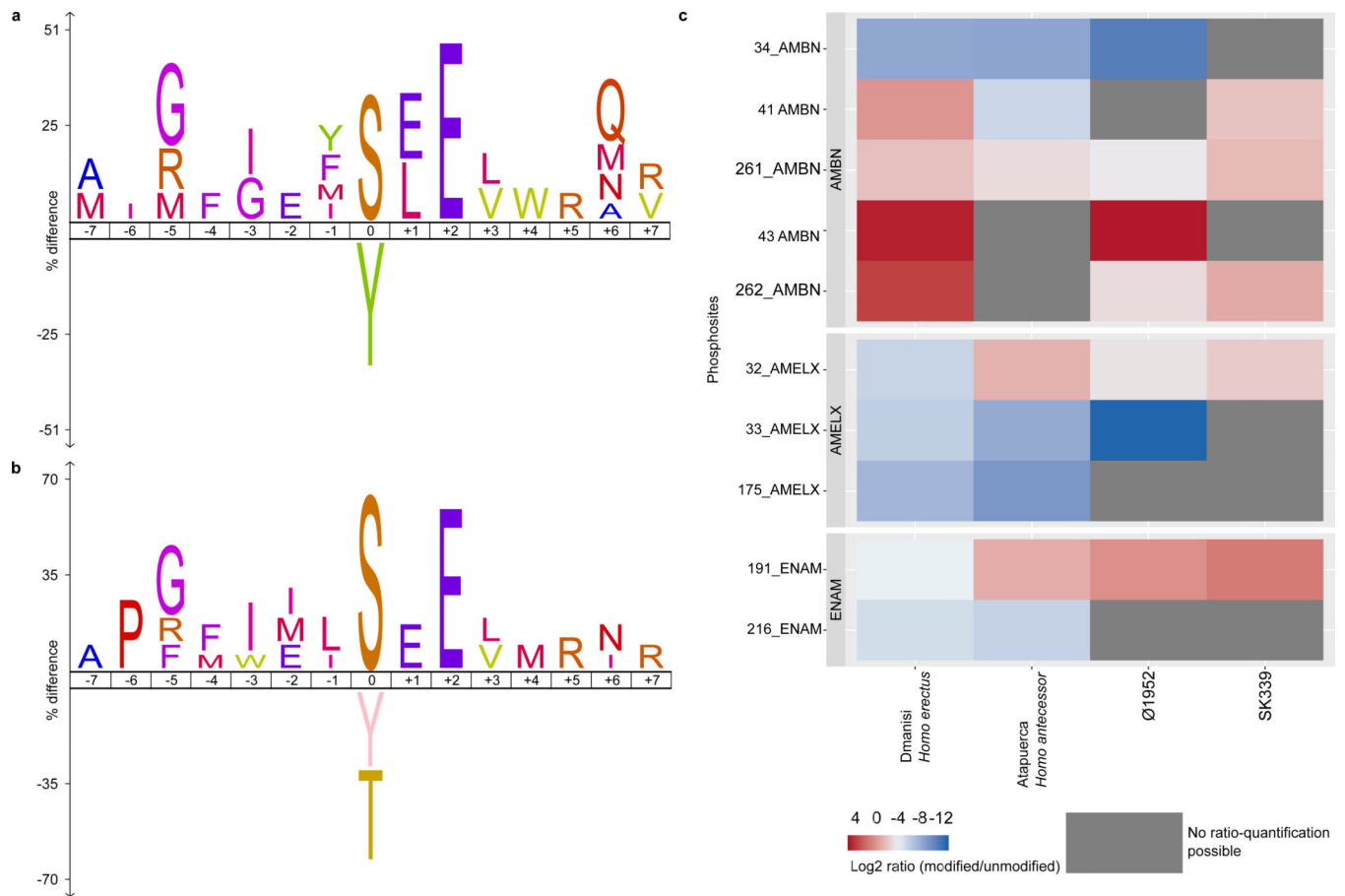
Figure 1. Hominin enamel proteome phosphorylation.
a, Phosphorylation sequence motif analysis of specimen ATD6–92 (*Homo antecessor* from Atapuerca). **b,** Phosphorylation sequence motif analysis of specimen D4163 (*Homo erectus* from Dmanisi). **c,** Phosphorylation occupancy comparison, expressed as the log2 of the summed intensity ratio of modified and unmodified peptides, for amino acid sites where data is available for at least two specimens. Y-axis labels indicate phosphorylated amino acid position per protein (UniProt accession numbers Q9NP70 (AMBN), Q99217 (AMELX), and Q9NRM1 (ENAM)).
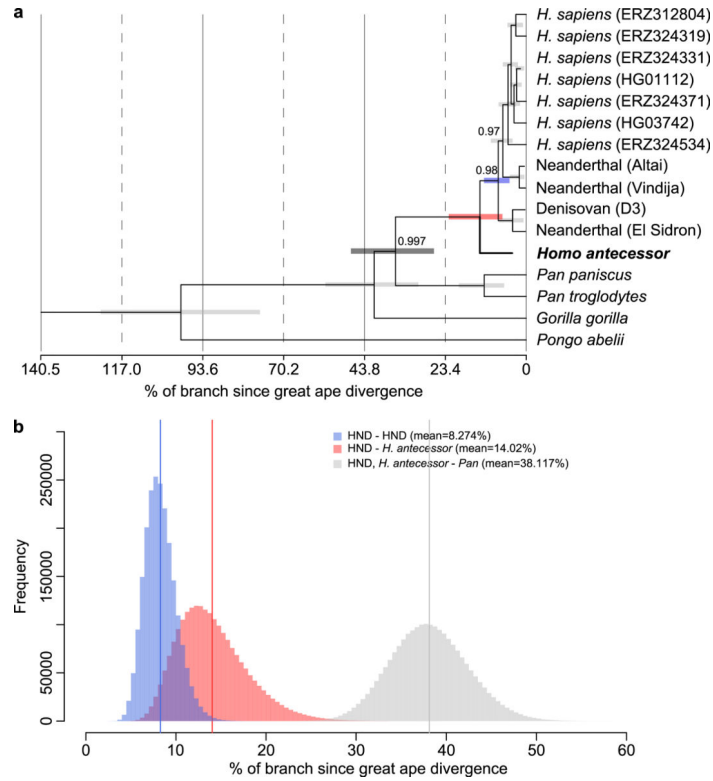
Figure 2. Phylogenetic analysis of *Homo antecessor* (ATD6–92, Gran Dolina, Atapuerca).
a, Maximum credibility tree estimated using BEAST and a concatenated alignment of seven protein sequences recovered for the ancient sample. Posterior Bayesian probabilities are indicated at nodes with a probability of ≤ 1. Horizontal error bars at each node indicate the 95% highest posterior density (HPD) intervals for the split time estimates. The position of *Homo antecessor* is consistent with that obtained via maximum-likelihood (Fig. S13) and Bayesian analysis (Fig. S16). **b**, Histograms of the divergence times obtained for the *Homo antecessor* – HND split (red), the HND – HND split (blue), and the *Pan* – (HND + *Homo antecessor*) split (grey). Divergence times **a** and **b** are shown as percentages since the divergence of all great apes.
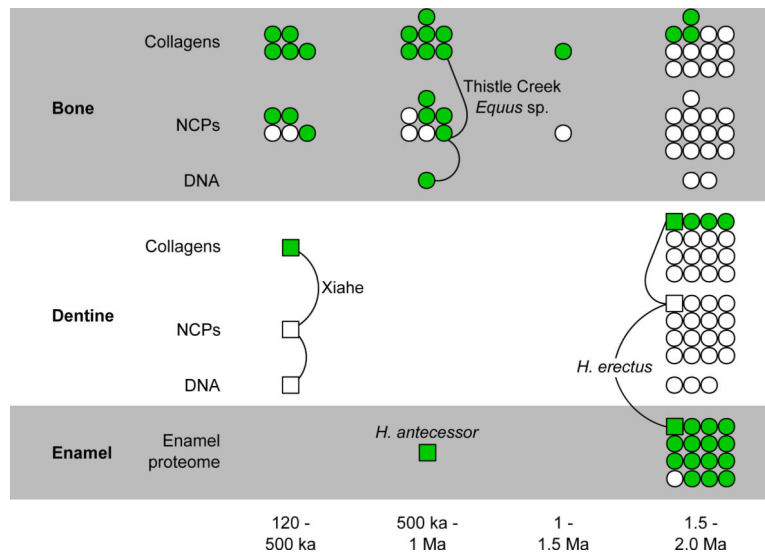
Figure 3. Skeletal proteome preservation in the Middle and Early Pleistocene (0.12 – 2.6 Ma).
For each sample, the presence (green) or absence (blank) of endogenous DNA, collagens, non-collagenous proteins (NCPs), or an enamel proteome is given. Only samples for which mammalian proteomes are published are considered[6–8,33–35]. Hominin samples are indicated with squares, other mammalian samples with circles. Selected specimens have their separate molecular components joined and are named.