
Supplementary information

The dental proteome of *Homo antecessor*

In the format provided by the
authors and unedited

Frido Welker[✉], Jazmín Ramos-Madrigal, Petra Gutenbrunner, Meaghan Mackie, Shivani Tiwary, Rosa Rakownikow Jersie-Christensen, Cristina Chiva, Marc R. Dickinson, Martin Kuhlwilm, Marc de Manuel, Pere Gelabert, María Martínón-Torres, Ann Margvelashvili, Juan Luis Arsuaga, Eudald Carbonell, Tomas Marques-Bonet, Kirsty Penkman, Eduard Sabidó, Jürgen Cox, Jesper V. Olsen, David Lordkipanidze, Fernando Racimo, Carles Lalueza-Fox, José María Bermúdez de Castro[✉], Eske Willerslev[✉] & Enrico Cappellini[✉]

Supplementary information for

The dental proteome of *Homo antecessor*

Frido Welker^{1,22,*}, Jazmín Ramos-Madrigal^{1,22}, Petra Gutenbrunner^{2,22}, Meaghan Mackie^{1,3}, Shivani Tiwary², Rosa Rakownikow Jersie-Christensen³, Cristina Chiva^{4,5}, Marc R. Dickinson⁶, Martin Kuhlwilm⁷, Marc de Manuel⁷, Pere Gelabert⁷, María Martínón-Torres^{8,9}, Ann Margvelashvili¹⁰, Juan Luis Arsuaga^{11,12}, Eudald Carbonell^{13,14}, Tomas Marques-Bonet^{4,7,15,16}, Kirsty Penkman⁶, Eduard Sabidó^{4,5}, Jürgen Cox², Jesper V. Olsen³, David Lordkipanidze^{10,17}, Fernando Racimo¹⁸, Carles Lalueza-Fox⁷, José María Bermúdez de Castro^{8,9,*}, Eske Willerslev^{18,19,20,21,*}, 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-ISCI de Evolución y Comportamiento Humanos, Madrid, Spain.

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

¹³ Departamento d'Història i Història de l'Art, Universitat 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 Tbilisi State University, Tbilisi, Georgia.

¹⁸ Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark.

¹⁹ Department of Zoology, University of Cambridge, Cambridge, United Kingdom.

²⁰ Wellcome Sanger Institute, Hinxton, United Kingdom.

²¹ Danish Institute for Advanced Study, University of Southern Denmark, Odense, Denmark.

²² These authors contributed equally.

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

1. ANTHROPOLOGICAL BACKGROUND

1.1 Atapuerca

María Martín-Torres, Juan Luis Arsuaga, Eudald Carbonell, José María Bermúdez de Castro

The Sierra de Atapuerca is placed about 15 km east of the city of Burgos in Northern Spain (3° 41'5922'' W, 42° 20'27'' N). It is a small hill with an area of about 25 km² and a maximum altitude of 1082 meters, placed between the basins of the river Duero to the southwest and the river Ebro to the northeast. The Sierra de Atapuerca has a strategic position at the end of the so-called Bureba corridor (Fig. S1), a relatively narrow passage between the Cordillera Cantábrica range (to the north) and the Sierra de la Demanda-Iberian range (located towards the south). The Bureba corridor was probably an important passage for the migrations of different species between Mediterranean areas and the interior of the Iberian Peninsula during the Pleistocene.

The Sierra de Atapuerca is made of Mesozoic limestone and from a structural point of view belongs to the Iberian range^{66,67}. The study of the geomorphology revealed that the Sierra de Atapuerca is an NNW-SSE anticline verging on the NE. This structure forms an anticline ridge made up of Upper Cretaceous limestone, dolomites, and calcarenites⁶⁸. The core of the anticline is made up of Keuper (Upper Triassic) shales and Triassic-Jurassic dolomites⁶⁹. The karst system, where the archaeological and paleontological sites are located, developed in the Upper Cretaceous. Unconformably overlying the Cretaceous limestone are calcareous conglomerates and red sandstone of Oligocene in age^{66,67}. Additional information on the geomorphology of the Sierra de Atapuerca can be found in Zazo *et al.*⁷⁰, and Parés and Pérez-González⁷¹. Ortega-Martínez⁷² carried out the main description of the multilevel karst system.

The construction of a railway trench at the southern slope of the Sierra de Atapuerca at the end of the XIX century, as well as the opening of a quarry during the mid-XX century in the abandoned trench, revealed the presence of some karstic fillings containing archaeological and paleontological information. From 1978 onwards, a systematic excavation of these fillings has produced rich archaeological and fossil records. One of the most important cavities is Gran Dolina. This cavity is 28 meters deep and is fully infilled with Early and Middle Pleistocene sediments. The first description of the Gran Dolina infilling was carried out by Gil *et al.*⁷³. These authors describe eleven main lithostratigraphic levels, TD1 to TD11 (from bottom to top; Fig. S2). Later, Parés and Pérez-González⁶⁹ and Parés and Pérez-González⁷¹ published a more detailed description of the stratigraphy of Gran Dolina. The finding of Early Pleistocene human fossil remains in level TD6 was very important for the scientific Atapuerca project. This finding, made in 1994 during the excavation of an archaeological pit of less than six square meters⁷⁴, initiated the systematic excavation and study of the fossiliferous filling of the Gran Dolina. An exhaustive stratigraphic study of the TD6 level has been carried out by Campaña *et al.*⁷⁵.

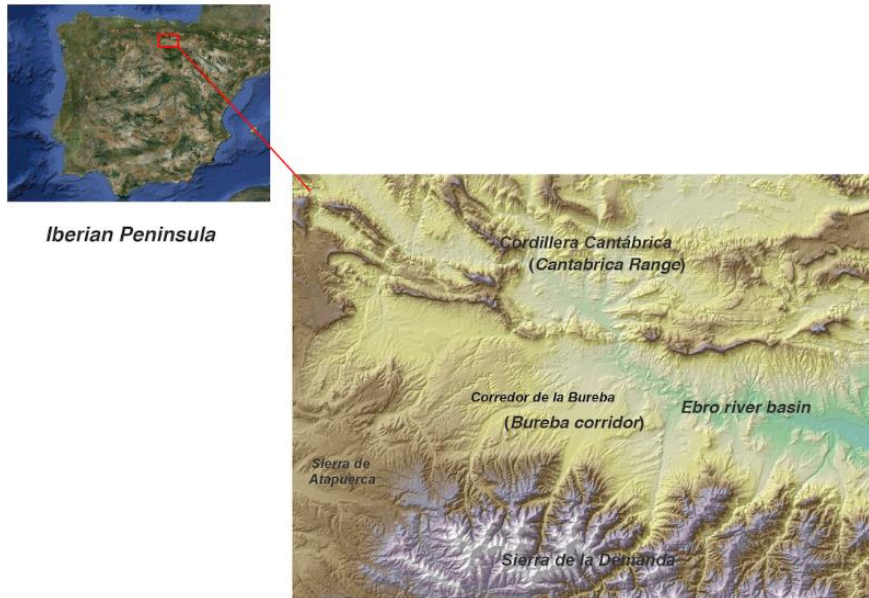


Figure S1. Location of the Sierra de Atapuerca in the Iberian Peninsula. Note that the Sierra de Atapuerca is located at the end of the Bureba corridor, which connects the basin of the Ebro River with the Castellana Plateau and the Duero river basin.

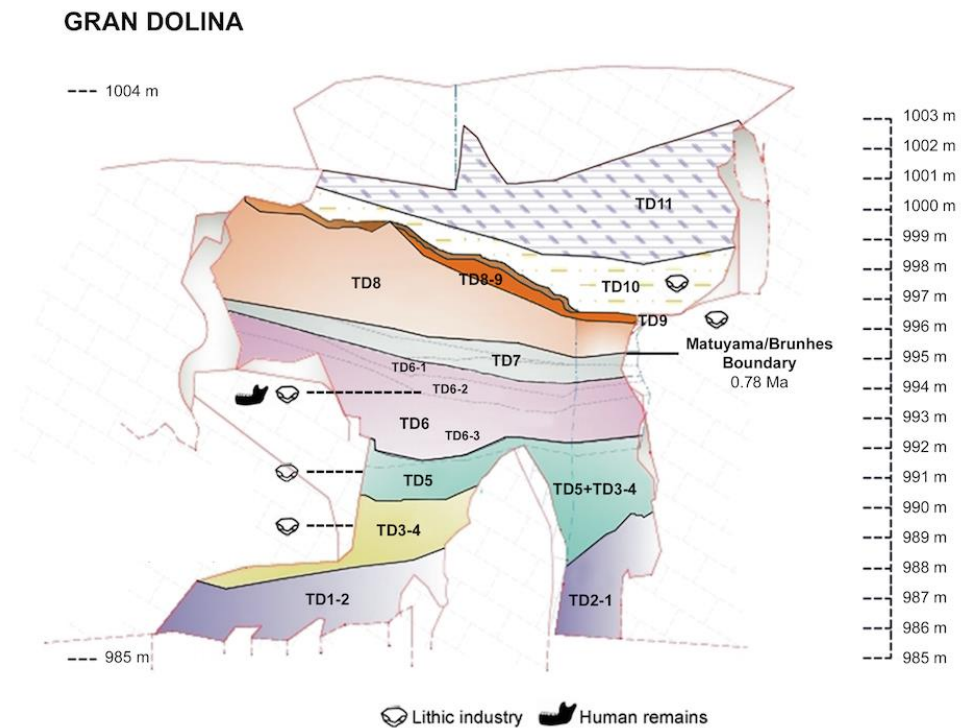


Figure S2. Stratigraphic profile of deposits and levels at the Gran Dolina cave site. A detailed view of the sublevels of TD6 can be found in Extended Data Figure 1b. This figure has been made by Jordi Rosell.

1.1.1 Atapuerca TD6.2 sublevel

The TD6 level has been divided in three sublevels: TD6.1, TD6.2, and TD6.3^{75,76} (Extended Data Fig. 1). The human fossils, as well as more than 830 artefacts and several thousands of micro- and macromammal fossil remains⁷⁷⁻⁸¹, come from the sublevel TD6.2^{69,71}. Parés and Pérez-González^{69,71} and more recently Álvarez Posada⁸² observed a polarity reversal between TD7 and TD8, interpreted as the Matuyama/Brunhes boundary, meaning that levels TD8 to TD11 were deposited during the Middle Pleistocene, whereas levels TD7 to TD1 were deposited during the Early Pleistocene. The combination of the paleomagnetic data and ESR/U-series ages suggests an age range between 0.78 and 0.86 million years ago (Ma) for the TD6 level⁸³. Thermoluminescence (TL) dates on samples taken at the TD7 level, one meter below the Brunhes/Matuyama boundary, give a weighted mean age of 0.96 ± 0.12 million years ago (Ma) for TD7⁸⁴. The ESR dating applied to optically bleached quartz grains from TD6 gives dates between 0.60 ± 0.09 Ma and 0.95 ± 0.09 Ma⁸⁵. These authors also obtained dates of 0.73 ± 0.13 Ma and 0.85 ± 0.14 Ma for the TD7 level, from samples taken under the Matuyama/Brunhes boundary. Using thermally transferred OSL (TT-OSL) dating of individual quartz grains, Arnold *et al.*⁸⁶ obtained a weighted mean age of 0.84 ± 0.06 Ma for the TD6 level. Arnold and Demuro⁸⁷ have undertaken a series of TT-OSL suitability assessments on known-age samples from TD6. Using this method, they obtained a weighted average age of 0.85 ± 0.04 Ma for TD6.3. The first direct Electron Spin Resonance (ESR) dating study of *Homo antecessor* using one hominin tooth has provided an age estimate ranging from 0.72 Ma to 0.95 Ma¹¹. Finally, a recent paleomagnetic study of the interior facies of TD1 places the TD6.2 hominins between the Matuyama/Brunhes boundary and the Jaramillo subchron⁸⁸. Summarizing, and taking into account the biostratigraphic information from TD6^{77,89}, we consider that the TD6 hominins can be assigned to the Marine Isotope Stage 21 (MIS 21).

1.1.2 ATD6-92

We sampled an isolated lower left first or second molar (ATD6-92). The specimen was excavated from square F14, sublevel TD6.2, in 2004. The fragment contains both enamel and dentine, although the root and dentine interior is largely absent. ATD6-92 derives from a stratigraphic unit containing over one hundred and seventy hominin fossils, all of which have been attributed to *Homo antecessor*.

The specimen has been μ CT scanned previously, as well as directly dated using a combination of Electron Spin Resonance (ESR), LA-ICP-MS U-series analysis, and bulk U-series dating¹¹. The dating analysis constrains the age of the tooth to between 624-949 thousand years ago (ka). As sublevel TD6.2 has reversed polarity, as indicated by paleomagnetic data, this age range can be shortened to 772-949 ka, and is in full agreement with other lines of chronological evidence for Gran Dolina TD6^{11,83,86} (see above).

1.2 Dmanisi

Ann Margvelashvili, David Lordkipanidze

The archaeological and paleontological site of Dmanisi, Georgia, is located in the South Caucasus, 55 km southwest of Tbilisi (41°20'10"N and 44°20'38"E). The site is situated on the promontory that overlooks the confluence of the Mashavera and Pinesauri rivers. Lower Pleistocene deposits are set between the Medieval ruins and above the 1.85-Ma Mashavera Basalt which originated from the Javakheti volcanic highland. Hominin fossils attributed to *Homo erectus ergaster georgicus* (Dmanisi) derive from the layer B1, which is dated to between 1.78 and 1.76 Ma, based on ⁴⁰Ar/³⁹Ar dates, paleomagnetism, and paleontologic constraints¹⁴. Besides the uniquely well-preserved and numerous hominin fossils, the site

has produced a rich paleontological assemblage, from some of which the recovery of ancient enamel proteins was possible. These specimens were discovered in different excavation blocks and/or have different preservation conditions⁶.

1.2.1 D4163

The isolated molar D4163 was discovered in M6, an excavation area located about 50 m northwest from Block II, where the majority of the hominin bones have been recovered. The excavations of M6 ceased after 2008, when the section reached the Mashavera basalt. The tooth was found in layer B1. This layer is dated to 1.77 Ma and correlates with the earliest Upper Matuyama chron, as it displays reversed polarity¹⁴. B1 ashes of all excavation blocks in Dmanisi have distinctive laminated calcretes – retarded or arrested water percolations - whereas these calcretes are absent in M6¹⁴.

D4163 is a right upper first molar. It is poorly preserved, with half of the crown missing and the third root broken off (Extended Data Fig. 2b). The tooth is worn down (grade 4-5 according to Molnar 1971⁹⁰). There is also the clear presence of taphonomic modifications of the dental surfaces. Together, these have resulted in the obliteration of the cusp shapes of the occlusal surface. Therefore, the fissure morphology of the occlusal surface is not observable, as the grooves are absent due to wear, taphonomy, and missing parts of the crown. The occlusal outline is asymmetrical, as far as can be assessed from the preservation state. The tooth is rather rhomboidal, instead of square. The remnants of a flat wear patch (contact point) are present on the distal surface of the crown. A more concaved wear patch can be traced on the mesial surface as well. Only two major cusps are present (paracone and metacone) with a faint buccal groove running between them. A shallow central fossa is present as well. The paracone is larger than the metacone, which is different from the other Dmanisi first molars. However, even though half of the crown is missing, it can be assumed that the bucco-lingual dimensions are larger than the mesio-distal dimensions. This pattern resembles the pattern observed for the first and second molars of the D4500 cranium¹⁵. Similarity between the D4500 molars and D4163 also include the massive roots present on all these specimens.

The roots slightly angle distally, and trifurcation is in the cervical third of the root. The roots are set widely apart from each other. The largest root (lingual/palatinal) is broken off. The bucco-distal root is the longest, whereas the bucco-mesial root is shorter, more triangular, and compressed mesio-distally.

The molar does not fit to any current Dmanisi individuals or their associated mandibles. It is therefore likely that D4163 belongs to another *Homo erectus* (Dmanisi) individual.

2. SUPPLEMENTARY METHODS and RESULTS

2.1 Amino Acid Racemization

Marc R. Dickinson, Kirsty Penkman

2.1.1 Amino acid racemization methods

Chiral amino acid analysis was undertaken on one Pleistocene sample from the hominin tooth from Dmanisi to test the endogeneity of the enamel protein through its degradation patterns. The current technique of amino acid analysis developed for geochronological purposes³⁶ combines a reverse-phase high-pressure liquid chromatography (RP-HPLC) method of analysis⁹¹ with the isolation of an 'intra-crystalline' fraction of amino acids by bleach treatment⁹². This combination of techniques results in the analysis of D/L values of multiple amino acids from the chemically protected (closed system) protein fraction within the biomineral, thereby enabling both decreased sample mass and increased reliability of the analysis.

The tooth chip was separated into the enamel and dentine portions and each 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 protein³⁷. 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*). After demineralization of the enamel, the pH of the solution was raised with KOH and then centrifuged for 5 min at 13,000 rpm, whereupon a biphasic solution formed³⁷. The supernatant was extracted and dried via centrifugal evaporation. Samples were analyzed in duplicate by RP-HPLC, with standards and blanks run alongside samples. During preparative hydrolysis, both asparagine (Asn) and glutamine (Gln) undergo rapid irreversible deamination to aspartic acid (Asp) and glutamic acid (Glu), respectively³⁸. It is therefore not possible to distinguish between the acidic amino acids and their derivatives and they are reported together as Asx and Glx respectively. The secondary amino acids proline and hydroxyproline are not detectable in the chiral amino acid analysis used here, and are therefore not reported below.

The D/L ratios of aspartic acid/asparagine, glutamic acid/glutamine, phenylalanine and alanine (D/L Asx, Glx, Phe and Ala) are assessed to provide an overall estimate of intra-crystalline protein decomposition (IcPD). In a closed system, the amino acid ratios of the FAA and the THAA subsamples should be highly correlated, enabling the recognition of compromised samples (e.g. Preece & Penkman, 2005⁹³; Dickinson *et al.*, 2019³⁷). The D/L of all the amino acids will increase with time, but each amino acid racemises at a different rate, therefore providing different resolution over different timescales.

2.1.2 Amino acid racemization results

The amino acid composition of the hominin enamel differs from dentine, but closely resembles that of enamel from other taxa from Dmanisi⁶ (e.g. Megacerini; Fig. S3), indicating the amino acids studied are derived from the same proteins entrapped within the intra-crystalline fraction of enamel.

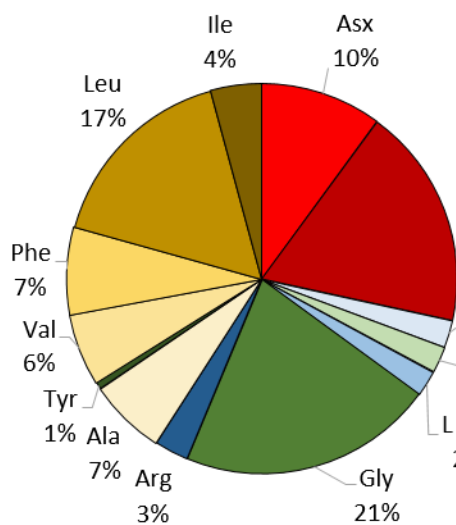
In a closed system, all of the products of protein degradation should be retained, so the racemization of the FAA and THAA should be highly correlated over geological time. More labile amino acids, such as those that are unbound (FAA) are more likely to leach out of an open system, so open system behavior tends to

result in data points falling away from the general closed system trend. Unfortunately, there is currently no directly comparative enamel data from the same species, but we have compared the hominin enamel both to other species from Dmanisi, and to Proboscidea data from the UK, for which there is a larger dataset (Extended Data Fig. 3).

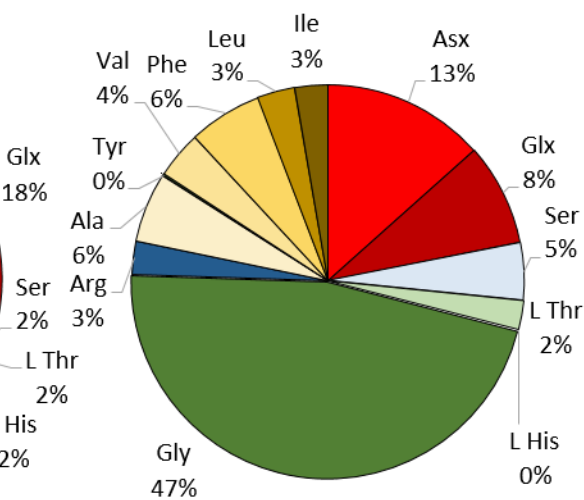
Whilst taxonomic effects influence the rates of racemization (so different taxa will have a different extent of racemization, despite being the same age), it is likely that the relative relationship between the racemization of FAA and THAA is similar enough between taxa to identify if significant leaching/contamination of the original protein has occurred. The FAA and THAA racemization values from the hominin enamel plot along the same trajectory as those from Proboscidea enamel, indicating that the intra-crystalline enamel amino acids from the hominin tooth are endogenous and show closed system behavior.

The extent of racemization in all four amino acids studied is comparable to other taxa from the Dmanisi deposits of the same age (Extended Data Fig. 3), but slightly higher, indicating that hominin enamel amino acids racemise at a faster rate. This fast rate of degradation is also consistent with the high percentage of FAA, implying higher rates of hydrolysis (Fig. S4). The extent of racemization and hydrolysis in the hominin enamel indicates a great antiquity, which is consistent with 1.8 Ma amino acids from this region. In summary, the protein composition, level of peptide bond hydrolysis, and extent of racemization in the hominin sample is consistent with the isolation of a closed system of endogenous protein from the intra-crystalline fraction of the enamel.

Hominin: Enamel



Dentine



Megacerini Enamel - 16629

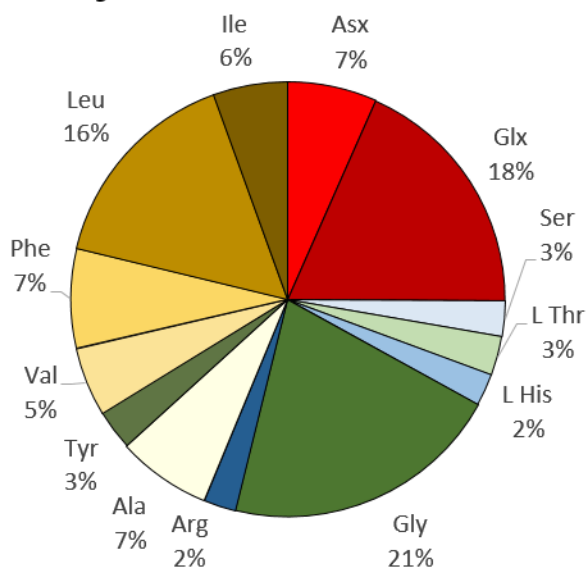


Figure S3. THAA amino acid compositions of enamel and dentine from the hominin tooth D4163. A profile of a Megacerini enamel sample from Dmanisi is shown for comparison (taken from Cappellini *et al.*, 2019⁶). The amino acid composition of the hominin enamel closely resembles that of the Megacerini enamel.

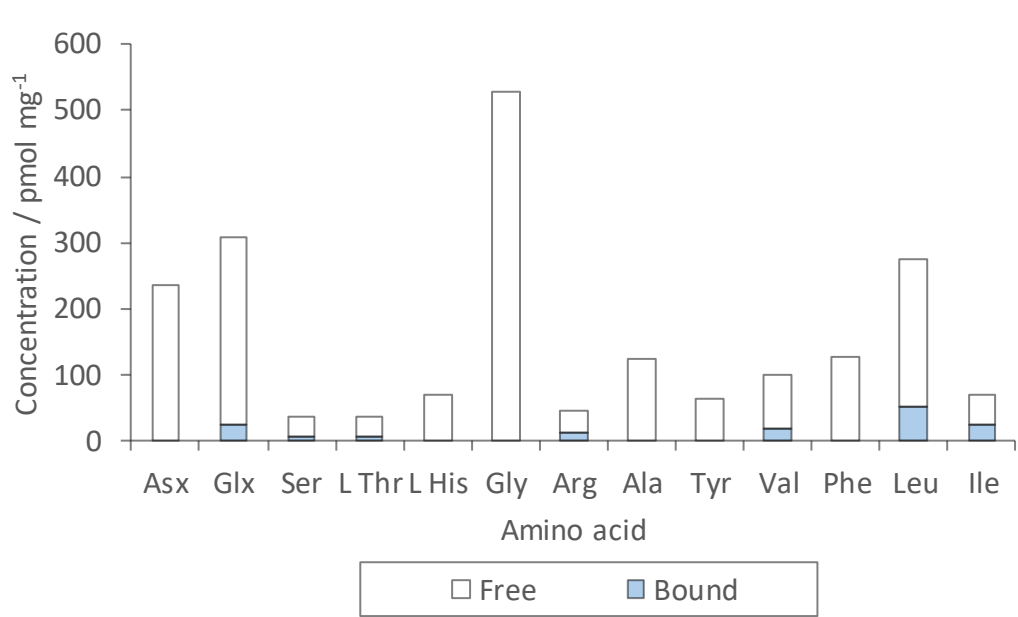


Figure S4. Absolute concentrations of the free and bound amino acids. Bound amino acid concentrations have been calculated by subtracting the FAA from the THAA concentrations. If the calculation results in a negative number, the bound concentration is assumed to be zero; this approach is likely to slightly overestimate the %FAA due to amino acid breakdown during the hydrolysis step for the THAA.

2.2 Proteomic Analysis

Frido Welker, Petra Gutenbrunner, Shivani Tiwary, Meaghan Mackie, Rosa Rakownikov Jersie-Christensen, Cristina Chiva, Pere Gelabert, Eduard Sabido, Jurgen Cox, Jesper V. Olsen, Enrico Cappellini

2.2.1 Protein extraction

We attempted protein analysis on enamel fragments from both hominins and our recent human control (Ø1952), and in addition also conducted protein extractions on a dentine sample from the Dmanisi hominin. Enamel powders might contain traces of dentine, as no dedicated chemical or physical separation was attempted after drilling. We did not perform subsuperficial acid-etching of exposed enamel surfaces^{16,20,94}. Three separate protocols were utilized. See Extended Data Table 1 for further details on the division of extraction approaches across samples.

Extraction method 1.

Approximately 250 mg of enamel was demineralized in 1.2 M HCl at 4°C, changing and saving the HCl supernatant every 24 hours. The peptides suspended in the HCl demineralization fractions were then directly cleaned, concentrated and immobilized, skipping reduction, alkylation and digestion steps, using in house assembled C18 solid-phase extraction Stage-Tips, using previously published methods⁶.

Extraction method 2.

After the demineralization of the Dmanisi sample from Extraction method 1, there were some remnants of the sample that did not appear to be mineral, and therefore digestion on this fraction was performed using a guanidine hydrochloride (GuHCl) solution, in case some protein could be extracted, following Mackie *et al.*³⁹. In short, the sample was reduced and alkylated using 2-Chloroacetamide (CAA) and Tris (2-carboxyethyl) phosphine (TCEP) within a GuHCl solution and incubated at 80°C. Subsequently, the sample was digested with the enzymes rLysC and Trypsin (Promega) overnight. Samples were then acidified to under pH 2 using 10% TFA in ultrapure water. Peptides were collected and cleaned on C18 Stage-Tips, as in method 1.

Extraction method 3.

A single sample of the Dmanisi *Homo erectus* was demineralized in 10% TFA Peptide clean-up on C18 StageTips followed the details outlined for Extraction method 1.

2.2.2 LC-MS/MS analysis

Protein extractions were analyzed in two independent laboratories. First, extracts were all analyzed at the Novo Nordisk Centre for Protein Research, University of Copenhagen (Denmark). Second, a significant subset of extracts were also analyzed at the Proteomics Unit of the Centre for Genomic Regulation, Pompeu Fabra University (Barcelona, Spain). Several of the extracts were also analyzed multiple times on the same instrument at the same facility, providing further depth to our proteomic data (Extended Data Tab. 1). For all analyses, injection blanks preceded and followed sample injections to minimize the risk of sample carry-over between runs. Furthermore, whenever possible, ancient enamel proteomes were only injected after extensive cleaning of the MS.

Copenhagen (Denmark)

Samples were analysed using in an EASY-nLC 1200 (Thermo Fisher Scientific (Proxeon), Odense, Denmark) coupled to either a Q-Exactive HF or HF-X (Thermo Fisher Scientific, Bremen, Germany) orbitrap mass

spectrometer. The peptides were eluted from the StageTips using 20 μ L 40% acetonitrile (ACN) and subsequently 10 μ L 60% ACN into a 96 well plate. For the Atapuerca sample, a quarter of the eluted sample was dried and sent to Barcelona. In Copenhagen, after elution, all samples were vacuum centrifuged at 40°C until approximately 3 μ L was left. The remainder of the Atapuerca samples were then rehydrated with 10 μ L of 0.1% TFA, 5% ACN. The Dmanisi samples were rehydrated with 9 μ L of 0.1% TFA, 5% ACN, 4 μ L of which being saved and sent to Spain. Different LC and MS methods were used in Copenhagen based on the instrument and column length. In addition, for the Dmanisi runs, a short test run of each sample was also performed before the main analysis. Columns were either 15 or 50 cm PicoFrit® columns (75 μ m inner diameter) in-house packed with 1.9 μ m C18 beads (Reposil-AQ Pur, Dr. Maisch). Buffer B consisted of 80% acetonitrile and buffer A of milliQ water, both containing 0.1% TFA. Tables S1-S3 summarise the different LC and MS methods used in Copenhagen.

Table S1. nLC operation parameters for samples when coupled to a Q-Exactive HF or Q-Exactive HF-X mass spectrometer (Copenhagen).

Sample set	Instrument	Column length (cm)	Injection volume (μ L)	Flow rate (nl/min)	Time (mm:ss)	Duration (min, from time)	% Buffer B
Atapuerca	Q-Exactive HF	50	5	200	00:00	0	2
					110:00	110	25
					135:00	25	40
					140:00	5	60
					145:00	5	60
					150:00	5	2
					165:00	15	2
Dmanisi	Q-Exactive HF test run	15	1	250	00:00	0	20
					11:00	11	35
					13:00	2	80
					14:00	1	80
					16:00	2	5
					26:00	10	5
	Q-Exactive HF	50	4	200	00:00	0	2
					110:00	110	25
					135:00	25	40
					140:00	5	60
					145:00	5	60
					150:00	5	2
					165:00	15	2
	Q-Exactive HF-X	15	5	250	00:00	0	5
					50:00	50	30
					60:00	10	45
					62:00	2	80
					67:00	5	80
					72:00	5	5
					77:00	5	5

Table S2. Mass spectrometer full scan operation parameters for samples on the Q-Exactive HF and Q-Exactive HF-X (Copenhagen). Resolution is at m/z 200. IT: injection time, ACG: auto gain control target.

Sample set	Instrument	TopN	Resolution	Mass range	ACG	Max IT
Atapuerca	Q-Exactive HF	10	120,000	300-1750 m/z	3e6	20 ms
Dmanisi	Q-Exactive HF test run	10	60,000	350-1400 m/z	3e6	20 ms
	Q-Exactive HF	10	120,000	300-1750 m/z	3e6	20 ms
	Q-Exactive HF-X	10	120,000	350-1400 m/z	3e6	25 ms

Table S3. Mass spectrometer fragment scan operation parameters for samples on the Q-Exactive HF and Q-Exactive HF-X (Copenhagen). Resolution is at m/z 200. IT: injection time, NCE: normalized collision energy, ACG: auto gain control target, DE: dynamic exclusion.

Sample set	Instrument	Resolution	Fixed first mass	ACG	Max IT	NCE	Isolation window	DE
Atapuerca	Q Exactive HF	60,000	100 m/z	2e5	108 ms	28	1.3 m/z	30 s
Dmanisi	Q Exactive HF test run	30,000	100 m/z	1e5	45 ms	28	1.3 m/z	10 s
	Q Exactive HF	60,000	—	2e5	108 ms	25	1.3 m/z	30 s
	Q Exactive HF-X	60,000	100 m/z	2e5	108 ms	28	1.2 m/z	20 s

Barcelona (Spain)

Samples were dissolved in 0.1% formic acid and analyzed by LC-MSMS using a LTQ-Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an EASY-nLC 1000 (Thermo Fisher Scientific (Proxeon), Odense, Denmark). Peptides were loaded directly onto the analytical column and separated by reversed-phase chromatography using a 50-cm column with an inner diameter of 75 μ m, packed with 2 μ m C18 particles (Thermo Scientific, San Jose, CA, USA). Chromatographic gradients started at 95% buffer A and 5% buffer B with a flow rate of 300 nl/min for 5 minutes and gradually increased to 22% buffer B and 78% A in 105 min and then to 35% buffer B and 65% A in 15 min. After each analysis, the column was washed for 10 min with 10% buffer A and 90% buffer B. Buffer A: 0.1% formic acid in water. Buffer B: 0.1% formic acid in acetonitrile.

The mass spectrometer was operated in positive ionization mode with nanospray voltage set at 2.4 kV and source temperature at 275°C. Ultramark 1621 was used for external calibration of the FT mass analyzer prior the analyses, and an internal calibration was performed using the background polysiloxane ion signal at m/z 445.1200. The acquisition was performed in data-dependent acquisition (DDA) mode and full MS scans with 1 micro scans at resolution of 120,000 were used over a mass range of m/z 350-1500 with detection in the Orbitrap mass analyzer. Auto gain control (AGC) was set to 4E5 and charge state filtering disqualifying singly charged peptides was activated. In each cycle of data-dependent acquisition analysis, following each survey scan, the most intense ions above a threshold ion count of 10000 were selected for fragmentation. The number of selected precursor ions for fragmentation was determined by the “Top Speed” acquisition algorithm and a dynamic exclusion of 60 seconds. Fragment ion spectra were produced via high-energy collision dissociation (HCD) at normalized collision energy of 28% and acquired in the Orbitrap mass analyzer. AGC was set to 3E4, and an isolation window of 1.6 m/z and maximum injection time of 100 ms were used. All data were acquired with Xcalibur software v4.1.31.9. Several blank samples were injected before and after each sample to avoid sample carryover.

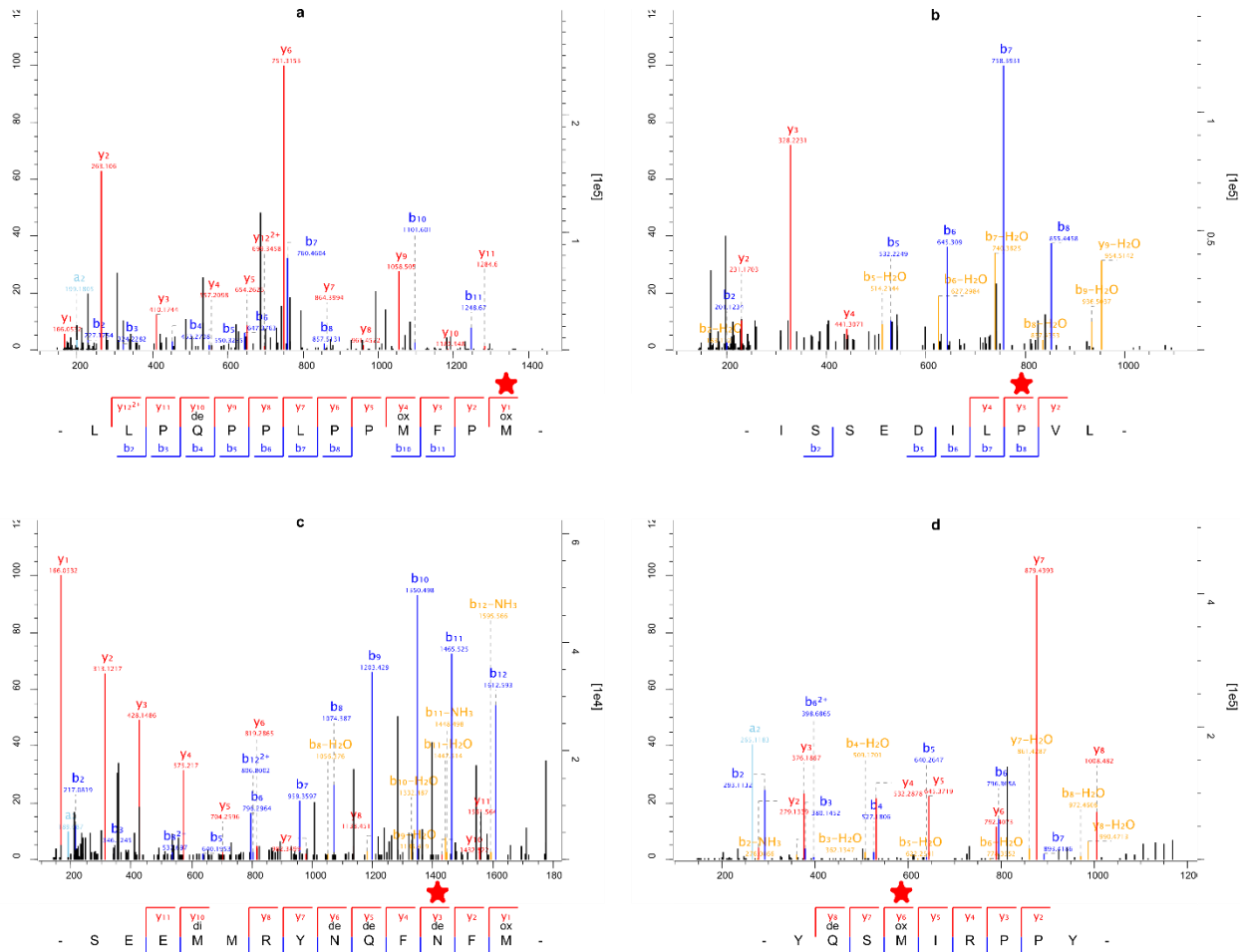


Figure S5. Examples of LC-MS/MS spectra of *Homo antecessor*. **a**, Ancestral SAP (L179M) in AMELY present in *Homo antecessor*. Stagtip number: 1069. Scan number: 31841. **b**, *Homo antecessor*-derived SAP (L1089P) in *COL17α1*. Stagtip number: 1069. Scan number: 31158. **c**, Derived SAP (T64N) in ENAM shared with *Homo sapiens*, Neanderthals, and Denisovans. Stagtip number: 1069. Scan number: 18166. **d**, AMELY-specific peptide. Stagtip number: 1069. Scan number: 9847. Note the presence of deamidation (de) and oxidation (ox, di) modifications in **a**, **c**, and **d**. Informative positions are highlighted (red star). Spectral acquisition was replicated across multiple extractions and injections for each informative position.

2.2.3 Protein sequence database

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. See §2.3.1 and §2.3.2 for details on the prediction of protein sequences for published genomes of great apes and extinct hominins. Accession numbers for UniProt or Genbank accessions of extant great apes and humans can be found in Table S4.

For each protein, we reconstructed the protein sequence of ancestral nodes in the Hominidae family through PhyloBot⁴⁴ to minimize cross-species proteomic effects⁴⁵, and added missing isoform variation based on the isoforms present for each protein in the human proteome as given by UniProt. Ancestral

Sequence Reconstruction (ASR) was conducted across the entire Hominoidea phylogeny using PhyloBot⁴⁴. Input sequences were constrained phylogenetically to (Macaca,(Nomascus,((Pongo abelii, Pongo pygmaeus),Gorilla,(Homo,((Pan paniscus, Pan troglodytes)))))). After obtaining protein sequences of extant, extinct, and ancestral nodes across Hominoidea, we imputed isoform variation known to exist for AMELX, AMELY, AMBN, AMTN, KLK4 and TUFT1 as we realized that isoforms for most of these proteins are not present in all great ape reference sequences available in UniProt. We assumed that alternative splicing would be placed identically for all great apes, took those protein positions from the human reference sequences (which does have isoform variation for each of these proteins), and placed them on the non-human great ape and ASR sequences. These manually created great ape and ASR isoforms were added to the protein sequence database, with sequence names appended with “ManIso2” and/or “_ManIso3”. 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 contaminants⁴⁶.

Supplementary Table S4. Accession numbers of publicly available protein sequences utilized in database construction and phylogenetic analysis.

Protein	<i>Homo sapiens</i>	<i>Pan paniscus</i>	<i>Pan troglodytes</i>	<i>Gorilla gorilla</i>	<i>Pongo abelii</i>	<i>Nomascus leucogenys</i>	<i>Macaca Mulatta</i>
COL1α1	P02452	XP_003817507	H2QDE6	G3RBN8	H2NVM9	XP_012358721	H9Z595
COL1α2	P08123	XP_003809763	H2QUY2	G3QT97	H2PMW7	G1RZZ2	H9Z2D1
COL17α1	Q9UMD9	XP_008949124.1	H2Q2J4	G3QE20	H2NBI5	G1RZC4	(absent)
ALB	P02768	XP_003832390	H2RBT1	G3S791	Q5NVH5	G1R8T8	Q28522
AMBN	Q9NP70	XP_003809040	H2R148	G3RCU1	H2PDI5	G1R841	F7HLX4
AMELY	Q99218	(absent)	Q861X8	C3UJP7	(absent)	(absent)	A0A1D5RDA1
AMELX	Q99217	XP_003805726	A5JJS6	G3SDK0	H2PUX0	G1RCS3	A5JJS8
ENAM	Q9NRM1	B2L7U5	H2QPM0	B2L7U8	H2PDI6	G1R843	F7H832
TUFT1	Q9NNX1	XP_003817293	K7CQG4	G3QY68	H2N5V2	G1RGY4	G7MDK9
KLK4	Q9Y5K2	XP_003813692	XP_009434410	G3QU55	A0A2J8U913	G1R1C5	G7NMD1
MMP20	O60882	XP_003828430	H2Q4M8	G3QLA8	H2NF32	G1R6B1	F7GQW6
AMTN	Q6UX39	XP_003809041	H2QPL9	G3RJV5	H2PDI4	G1R825	F6VN65
ODAM	A1E959	XP_003809049	A1YQ94	G3QY18	H2PDH6	G1R7Z0	A1YQ92
AHSG	P02765	XP_008953975	Q9N2D0	E1U7Q5	H2PC98	G1R4B1	F6VZ47

Accession numbers either refer to UniProt or GenBank.

2.2.4 wiNner peptide sequence and SAP validation

Dataset

The wiNner model was trained for the prediction of the phylogenetically informative peptide sequences in the ancient samples (Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*). Ancient samples (Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*) were divided into phylogenetically informative peptide sequences and the 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 fauna⁶. The dataset only has HCD fragmentation, so we build two models. HCD+2 contains 5,555 unique modification-specific peptides, and HCD+3 contains 692 unique modification-specific peptides. For each unique modification-specific peptide, the spectrum with the highest Andromeda score was taken. Spectra with an Andromeda score below 50 were discarded. The retained data for each model was split into a division of 80:10:10 for training, validation and test sets, respectively. We kept test data for evaluating the wiNner model by calculating the Pearson correlation coefficient (PCC) between true and predicted intensities for each peptide.

The training data has non-tryptic peptides. In addition, variable modifications such as Oxidation (M), Deamidation (NQ), Gln->pyro-Glu, Glu->pyro-Glu, Oxidation (P), Carbamidomethyl (C), Dioxidation (MW), Oxidation (W), His->GluOH (H), His->Asp (H), Arg Ornithine, Phospho (STY) and Phospho (S) were added when processed with MaxQuant, and peptide sequences containing these variable modifications were taken into account.

wiNNer model

The conventional neural network model wiNNer¹⁷ (window-based neural network being easily re-trainable) uses sequence windows to compute feature space around the backbone bond for which y- and b-ion intensities need to be predicted. Amino acids on the N- and C- terminus, the distance of the bond from the peptide termini, and the length of peptide, were also used as features. Each amino acid residue and modified amino acid residues were converted to a 38 binary feature by one hot encoding to include residue specific modification and sliding window extending termini. For example, to include deamidation (NQ), the modification of two extra residues N(de) and Q(de) were added for one hot encoding. The missing intensities were set to zero and the intensities of a single peptide was normalized by the maximum value among y- and b-ions. Then, they were log transformed: $\log_2(1 + Intensity * 10000)$. We trained two regression models, one for HCD+2 and one for HCD+3.

The architecture of the wiNNer model is built using Keras (version 2.0.8; <https://keras.io/>), a high-level neural network, to train the neural network model. Tensorflow version 1.3.0 was used as backend in Keras. The architecture of neural network includes 5 dense layers. The input layers contain 991 features for a window size of 24. The hidden unit is reduced from 600, 400, 200 to 50 in subsequent dense layers, and the output layer contains 2 units for y- and b-ion peak intensities. Hyper-parameters such as batch size, dropout, learning rate and number of epochs were optimized separately for different models.

The wiNNer model can be accessed on GitHub (<https://github.com/cox-labs/wiNNer.git>).

Results

The supplementary Figure S6a shows the PCC of HCD+2 and HCD+3 models were 0.85 and 0.81 respectively, between true and predicted intensities for each peptide in the test sets. These results are close to the wiNNer model for unmodified sequences, where the PCC is 0.88 for HCD+2 and 0.76 for HCD+3¹⁷. Figure S6b shows the PCC between true and predicted spectra intensity of all the phylogenetically informative peptide sequences in the ancient samples. The PCC of All peptides, Oxidation (M), Deamidation (NQ), Oxidation (P), Dioxidation (MW) and Phospho (S) were 0.77, 0.79, 0.76, 0.68, 0.76 and 0.76 respectively. Again, these distributions are similar to the overall test performance of our wiNNer model, and indicate accurate peptide sequence identification for our phylogenetically informative spectra.

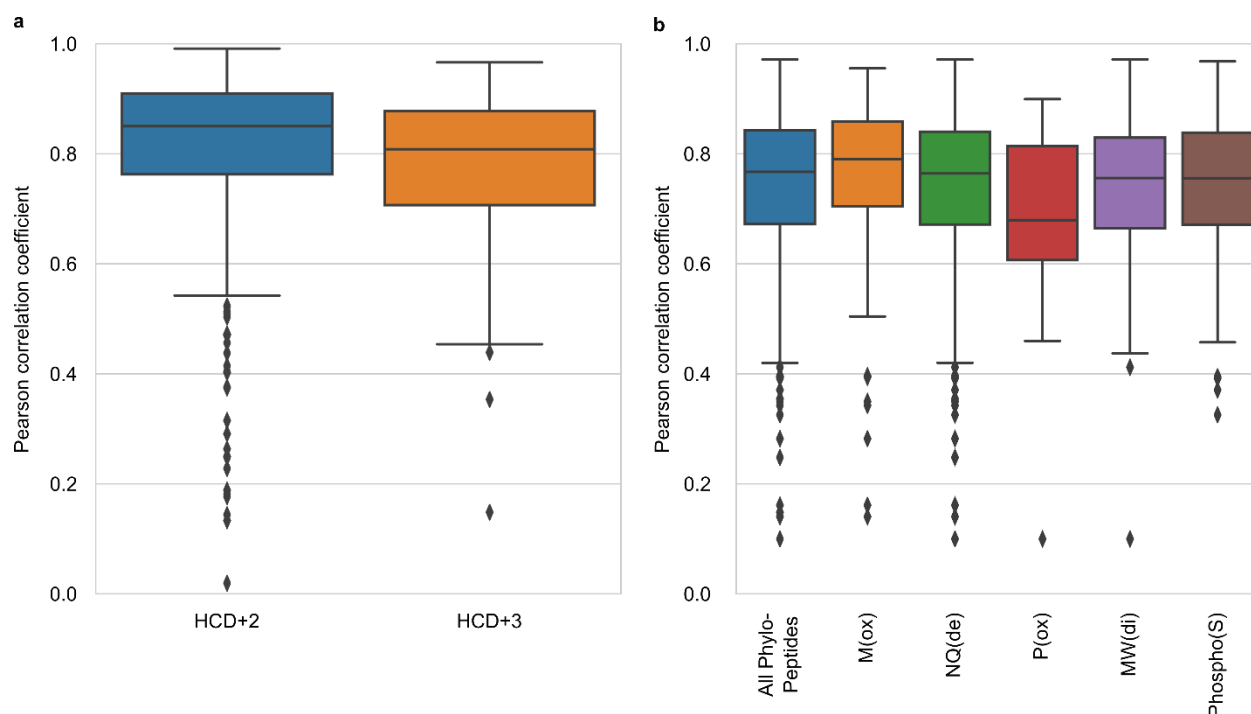


Figure S6. Results of wiNner on ancient enamel proteomes. **a**, Performance of the wiNner model for HCD+2 and HCD+3. The boxplots shows the PCC distribution between true and predicted spectrum intensity for each peptide sequence in the test data. The boxplots contains 556 unique PSMs for HCD+2 and 70 unique PSMs for HCD+3, respectively. **b**, PCC distribution of true and predicted intensities in phylogenetically informative peptides, and for specific classes of variable modifications. Distribution of PCCs of true and predicted y - and b -ion peak intensities in all the phylogenetically informative peptides ("All Phylo-Peptides"), and peptides including variable modifications, such as Oxidation (M), Deamidation (NQ), Oxidation (P), Dioxidation (MW) and Phospho (S). The boxplots contain 359, 125, 278, 22, 47, and 114 unique peptides, respectively. Lower mean PCC correlation for classes of variable modifications are generally related to smaller number of unique peptides included in wiNner test and validation. For **a** and **b**, each box extends from the lower to upper quartile values of the data. The line indicates the median. The whiskers extends up to 1.5 times the interquartile range. Values above and below the interquartile ranges are plotted in diamonds as outliers.

2.2.5 Proteomic data analysis

Proteomic data analysis focused on diagenetic degradation (deamidation and peptide bond hydrolysis), and the survival of *in vivo* modifications in the two Pleistocene hominin proteomes.

Deamidation analysis

Glutamine and asparagine deamidation has recently been investigated repeatedly in ancient proteome studies in an effort to understand patterns of protein damage between proteins and/or across time. We quantified glutamine and asparagine deamidation separately by following Mackie *et al.*³⁹ for MaxQuant. In short, protein-based deamidation rates are based on spectral intensities of matching PSMs, subsequently re-assessed through 1000 bootstraps to derive a 95% confidence interval. See the original paper for further details.

Peptide length distributions

Our digestion-free extraction approach allows us to explore peptide length distributions across and between samples. Here, peptide length distributions are based on MaxQuant peptide matches only. Comparisons are made with enamel proteomes extracted and analyzed in similar manners presented earlier^{6,8}.

MMP20 and KLK4 cleavage patterns

The existing literature on enamel proteomes and enamel proteome biomineralization describes the processes by which the enamel proteome is shaped through targeted protein hydrolysis by the proteolytic enzymes MMP20 and KLK4⁴⁹⁻⁵². These two enzymes digest the major enamel proteome components AMBN, AMELX, and ENAM into shorter amino acid chains. The *in vivo* cleavage locations of MMP20 and KLK4 on AMBN, AMEL(X/Y), and ENAM have been described in the literature⁴⁹⁻⁵². However, although various studies have previously obtained proteomic data on fully mineralized enamel samples, including ancient proteomes, no studies have explored whether *in vivo* MMP20 and KLK4 cleavage survives in (ancient) enamel proteomes.

Our digestion-free extraction approach allows us to explore whether our ancient proteome datasets preserve such *in vivo* patterns. As the Dmanisi sample contains several extracts either digested or alkylated, we explore MMP20 and KLK4 cleavage patterns solely using the Atapuerca dataset. In addition, we only focus on the proteins AMBN, AMEL(X/Y), and ENAM, since for these proteins we have large numbers of identified peptide-spectrum matches (PSMs), as well as sufficient literature on known MMP20 or KLK4 cleavage sites.

As a background, we calculated for each protein a theoretical matrix (21 x 21) containing the percentage that each possible amino acid (P1) – amino acid (P1') pair occurred in the UniProt sequence of that protein (AMBN_HUMAN, AMELX_HUMAN, AMELY_HUMAN, ENAM_HUMAN). For AMELX, this calculation was based on the isoform 3 variant, which is the longest of the three AMELX variants present in UniProt. Similarly, for AMELY, we selected isoform 1, because it is the longest isoform. For AMEL(X/Y), we had to construct a joined theoretical matrix, as AMELX and AMELY significantly overlap in peptide populations, while the sampled *Homo antecessor* individual is a male. We obtained this combined theoretical matrix by first joining the AMELX and AMELY matrices in a ratio of 9:1, which is close to the experimental proteomic and tRNA ratios observed in developing enamel⁹⁴⁻⁹⁶. The protein-derived ratios were then recalculated to a percentage for each amino acid pair. This theoretical background thereby is a simple model of protein fragmentation in which each amino acid pair (P1-P1') has an equal chance of being cleaved and being observed, regardless of the amino acids involved, their potential PTM modification, or their placement within the sequence or structure of the protein.

As for the observed peptide cleavage pairs, we took the C- and N- termini of the peptide-spectrum matches to each protein and counted the total occurrence of each amino acid pair. This matrix was then turned into percentages, and divided by the theoretical matrix to obtain a value indicating the fold-difference between the observed frequency of each hydrolyzed peptide bond (P1-P1') and random peptide bond cleavage (P1-P1') for each possible combination of amino acids in both positions.

The results of this exercise show that, against a general background of peptide bond hydrolysis, there are amino acid pairs that stand out by very large positive fold-differences in their occurrence (Extended Data Fig. 7a, b). When mapped to known MMP20 and KLK4 cleavage sites, bonds with high rates of fragmentation are without exception known cleavage sites of either of those enzymes, or are within 2 amino acids in either C- or N-terminal direction of such a known cleavage site (Extended Data Fig. 7b, c). This finding provides experimental evidence that *in vivo* MMP20 and KLK4 protein cleavage survives in Middle Pleistocene proteomes. In addition, for ENAM we observe a set of cleavage pairs that have highly enriched fold cleavages located closely together (positions 56-57, 58-59, 65-66, 66-67; accession ENAM_HUMAN in UniProt), suggesting that extensive hydrolysis, either *in vivo* or diagenetically, occurs at or around these positions (Extended Data Fig. 7b).

Furthermore, our data shows an absence of PSMs to known locations of N-linked glycosylation in ENAM (Extended Data Fig. 7c), as revealed in comparison with such modifications for ENAM included in UniProt. This might either be due to enhanced molecular diagenesis at such positions or aspects of our extraction and/or spectral identification approach preventing the recovery of PSMs from such sequence regions.

Phosphorylation occupancy

Phosphorylation occupancy was calculated using MaxQuant (v1.5.3.30) by measuring the intensity ratio of phosphorylated and non-phosphorylated peptides. The raw mass spectrometry data of both hominins, SK339 and Ø1952 was analyzed in a combined search in MaxQuant for occupancy comparison between the samples. For the database search, the same settings were applied as described in the Methods section. We modified the used sequence database in such a way that it only contains one protein sequence entry per species. The removal of AMELX and AMBN isoforms ensures that occupancy is calculated only once for each site. Additionally, we excluded AMELY protein sequences, since we did not confidently identify AMELY in the initial database search in the Dmanisi specimen.

In total, we could identify 18 phosphorylation sites. For three sites, ratio calculation was possible, since in all four specimen the phosphorylated peptides and its non-phosphorylated counterparts were present and quantified through label-free quantification (Tab. S5, Group 1). For nine sites, it was not possible to quantify the phosphorylated peptide in at least one out of all 4 specimens (Tab. S5, Group 2). For six sites, only the phosphorylated peptides were present but not the non-phosphorylated counterpart at least one out of all four specimens (Tab. S5, Group 3).

Table S5. Description of all identified phosphorylation sites in both ancient hominins and both modern human control samples, including their summed intensity ratios of phosphorylated and unmodified peptides. The sites are grouped, depending on the presence of both phosphorylated and non-phosphorylated peptides (group 1), the sole quantification of unmodified peptides in at least one of the samples (0, group 2), or the presence of phosphorylated peptides only (∞ , group 3). Ø1952 represents our recent human control. SK339 is an example of a recent control taken from Stewart *et al.*¹⁶.

Group 1: identification of both, phosphorylated and unmodified peptides							
	AMELX	AMBN	ENAM				
Phosphosite	32_AMELX	43_AMBN	191_ENAM				
<i>Homo erectus</i>	0.017859	0.23628	0.053114				
<i>Homo antecessor</i>	0.36654	0.10561	0.42624				
Ø1952	0.088134	0.073352	0.98228				
SK339	0.17361	0.28135	1.7769				
Group 2: 0 – quantification of only unmodified peptides							
	AMELX					AMBN	
Phosphosite	33_AMELX	42_AMELX	175_AMELX	181_AMELX	182_AMELX	34_AMBN	41_AMBN
<i>Homo erectus</i>	0.013984	0.001532	0.0053254	0	0	0.0028724	0.84794
<i>Homo antecessor</i>	0.0033069	0	0.0014622	0	0	0.0027449	0.020293
Ø1952	0.00016585	0	0	0.010832	0	0.00051682	0
SK339	0	0	0	0	0	0	0.22245
	ENAM						
Phosphosite	54_ENAM	216_ENAM					
<i>Homo erectus</i>	0	0.025636					
<i>Homo antecessor</i>	0.20649	0.01626					
Ø1952	0	0					
SK339	0	0					
Group 3: ∞ - identification of only phosphorylated peptides							
	AMTN		AMBN				
Phosphosite	115_AMTN	116_AMTN	101_AMBN	261_AMBN	262_AMBN	303_AMBN	
<i>Homo erectus</i>	∞	∞	∞	17.033	8.7485	∞	
<i>Homo antecessor</i>	∞	∞	∞	∞	∞	0.025266	
Ø1952	∞	∞	0	0.19077	0.10544	∞	
SK339	∞	∞	∞	∞	0.44953	∞	

2.2.6 Recent modern human control samples

To contrast our ancient hominin enamel proteomes with modern human data from less diagenetically-altered environments, we also processed a single Medieval human tooth from Copenhagen (Ø1952) using extraction methods 1 and 3 (two injections in total, one based on HCl and one based on TFA; see §2.2.1). LC-MS/MS set-up was identical as described above (§2.2.2). Ø1952 represents a male individual, derives from the former cemetery of the Almindeligt Hospital which was in use from approximately 1600-1800 A.D., and which was excavated in 1952. The specimen is stored and owned by the Laboratory of Biological Anthropology, Department of Forensic Medicine, University of Copenhagen. Additionally, we processed published human enamel proteome data from Stewart *et al.*¹⁶. The latter utilizes a 5% HCl acid-etching

procedure that is minimally invasive, but also only recovers proteins from the outer enamel surface. Stewart *et al.* used an extraction method without enzymatic digestion, and is therefore partly comparable to the workflow used for Ø1952 and the ancient samples processed using extraction 1 and 3. Their samples range in chronological age between approximately 5700 and 200 years old. Proteomic data from Stewart *et al.* and Ø1952 was processed using the same MaxQuant as the ancient samples against a protein sequence database restricted to *Homo sapiens*. All other search settings were as specified in the Methods, and are therefore identical between the modern human control samples and the ancient hominins.

Our analysis of Ø1952 and the Stewart *et al.* samples indicate that all these enamel proteome extracts are composed of the core enamel proteome (ENAM, AMBN, AMELX/Y, MMP20, AMTN) with the addition of additional collagens (COL1 α 1, COL1 α 2, COL17 α 1) and plasma proteins (AHSG, ALB) in some cases (Tab. S6). It is unclear whether the collagens and plasma proteins derive from residual dentine in the Ø1952 sample. It is likely these proteome components are endogenous to the enamel, however, as the acid-etch surface extractions performed by Stewart *et al.* also contains these proteins at low frequency. We confirm the sex assignment of the Stewart *et al.* samples by observing an absence of AMELY-unique peptides in female samples but the presence of AMELY-specific peptides in all male samples. We note that MMP20 is absent in acid-etching of enamel surfaces. Conversely, ODA and AMTN are only present in the acid-etches and not in the destructively-sampled Ø1952.

Analysis of proteome degradation performed on the *Homo antecessor* and *Homo erectus* datasets were contrasted with Ø1952 and SK339, a male individual of the Stewart *et al.* dataset. SK339 was chosen as it represents an approximation of the average protein sequence recovery across the Stewart *et al.* samples and is comparatively young in chronological age (19th century AD).

Protein sequence coverage obtained for our human control (Ø1952) and the ancient hominins is comparable, but on average lower compared to the recent controls generated through subsuperficial acid-etching of exposed enamel surfaces (Fig. S7). This is likely due to the deep-sequencing of the ancient samples across multiple LC-MS/MS injections, while data for SK339 is limited to a single LC-MS/MS injection and data for Ø1952 is limited to two LC-MS/MS injections. The decision to not employ subsuperficial acid etching of enamel surfaces, but opt for destructive sampling, is supported by the this observation of lower protein sequence recovery for the acid-etched datasets (Fig. S7). Still, major sequence regions for the enamel-specific proteins recovered in the ancient hominins are also present in the modern human controls, regardless of extraction approach (Extended Data Fig. 4).

We contrasted protein deamidation and peptide lengths of the recent human controls with those of the ancient hominin samples. We observe that protein deamidation is high for both recent controls, but less advanced compared with *Homo antecessor* and *Homo erectus* (Fig. S8). Interestingly, although the ancient hominins and Ø1952 have an expected pattern of more advanced asparagine (N) deamidation compared to glutamine (Q) deamidation, the SK339 proteome displays the opposite pattern of with more advanced glutamine deamidation. It is unclear why this would be the case. Enamel-specific proteins are fragmented *in vivo* through the combined action of MMP20 and KLK4^{49,52,53}. Nevertheless, we observe that the peptides recovered from the recent control samples are, on average, longer than those observed in the *Homo antecessor* and *Homo erectus* samples (Extended Data Fig. 6d; Ø1952 – *Homo antecessor*: t-test(6.87), df=1692, $p=9.0e^{-12}$; Ø1952 – *Homo erectus*: t-test(15.93), df=1573, $p<2.2e^{-16}$). We also observe that the peptide lengths for SK339 are significantly shorter than those obtained for Ø1952 (t-test(-6.67), df=570, $p=5.9e^{-11}$) or a Medieval ovicaprine enamel specimen published previously⁶ (t-test(-8.90), df=487, $p<2.2e^{-16}$). These observations are consistent with advanced enamel protein fragmentation *in vivo*, but also highlight the potential of significant variation in protein fragmentation between samples or analytical procedures during protein extraction and LC-MS/MS analysis.

Finally, we compare the amino acid cleavages between our recent controls and the *Homo antecessor* dataset. We observe that SK339 and Ø1952 display highly similar fold differences in cleavage site

occurrence compared to a random cleavage model for each protein (Fig S9). Large fold-differences in P1-P1' occurrence correspond to known MMP20 and/or KLK4 cleavage sites in these proteins. The *Homo antecessor* dataset contains a wider range of cleavage sites, compatible with diagenetic hydrolysis of peptide bonds. This has been observed previously for an ancient proteome as well⁷. Still, also for the *Homo antecessor* dataset, the observed P1-P1' pairs with a high fold difference compared to a random fragmentation model correspond to sites of known MMP20 and/or KLK4 activity, or peptide bond locations located in close proximity to such sites (see Extended Data Fig. 7).

Table S6. Sequence coverage (%) of proteins identified in the modern human enamel proteomes from Stewart *et al.* and Ø1952. For AMTN, AMELX, and AMELY, sequence coverage is calculated for the longest isoform present. Numbers in brackets indicate unique+razor counts per protein. Unique+razor peptide counts for AMELX are summed for the three isoforms present in the database – note that a large number of shared peptides between these three isoforms are therefore not taken into consideration here.

Protein	Guthlac_F	Seaham_F	SK363	SK378	Whitwell_F	Guthlac_M	Seaham_M	SK119	SK130	SK339	SK351	SK366	Whitwell_M	Ø1952
Sex	F	F	F	F	F	M	M	M	M	M	M	M	M	M
ALB		1.3 [1]			1.3 [1]	1.1 [1]		1.3 [1]	2.5 [2]	1.3 [1]		1.3 [1]	3.8 [1]	7.1 [4]
AHSG			2.5 [1]			2.2 [1]								12.8 [12]
COL1α1						6.3 [10]								67.3 [591]
COL1α2						1.8 [8]								44.8 [206]
COL17α1			1.1 [3]		1.7 [2]	1.6 [2]	0.9 [1]			0.8 [1]	0.8 [1]			17.2 [22]
ODAM		3.2 [1]	3.2 [1]		3.2 [1]		3.2 [1]	3.2 [1]		3.2 [1]		3.2 [1]	6.5 [2]	
AMTN				5.3 [1]				5.3 [2]		5.3 [2]	4.8 [1]		3.4 [1]	
AMELX	49.8 [124]	37.6 [89]	45.9 [131]	56.6 [120]	37.6 [112]	41 [77]	44.9 [113]	37.1 [121]	35.1 [93]	45.9 [147]	43.4 [147]	40.5 [132]	23.4 [36]	86.8 [413]
AMELY						22.4 [2]	22.9 [8]	31.8 [12]	27.6 [8]	33.3 [16]	32.8 [14]	29.2 [17]	17.7 [1]	64.6 [54]
AMBN	21.9 [50]	22.1 [62]	24.6 [53]	25.5 [68]	26 [80]	21.3 [48]	23.7 [69]	22.8 [58]	19.7 [50]	25.5 [67]	22.4 [55]	26.2 [62]	21 [54]	39.4 [312]
ENAM	10 [79]	9.1 [61]	9.9 [60]	7.6 [40]	10.7 [96]	9.1 [46]	9.6 [101]	9.8 [68]	8.9 [46]	10.6 [75]	10.7 [70]	10.3 [81]	6.7 [40]	15.5 [214]
MMP20														6.8 [5]

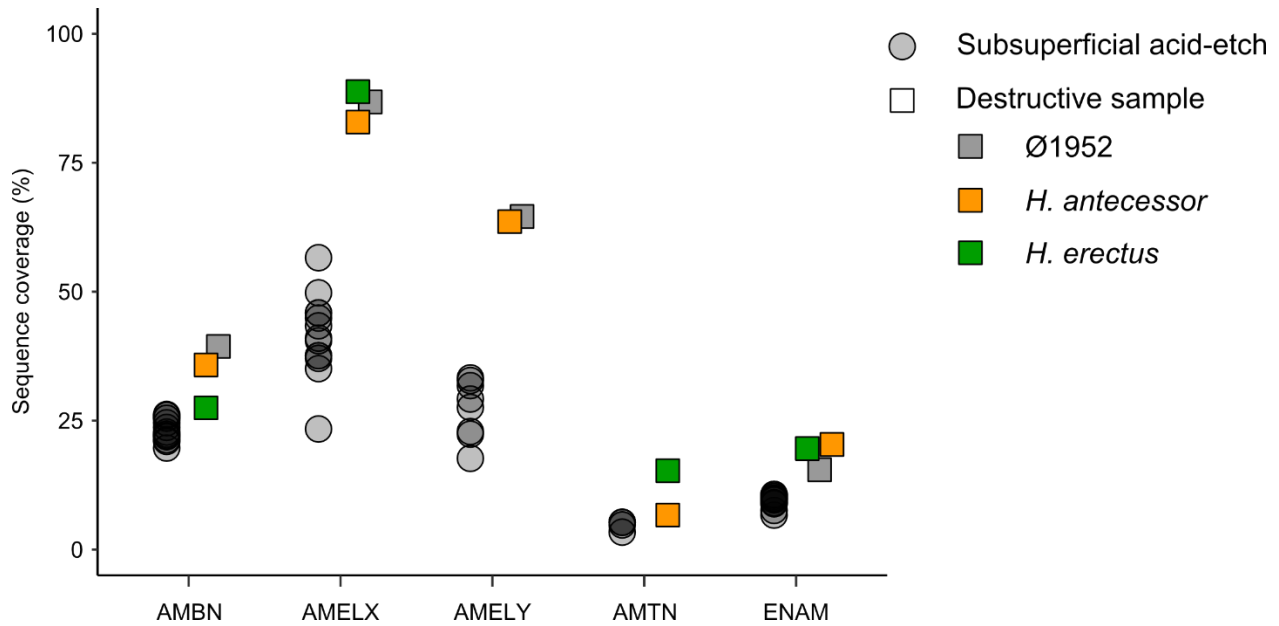


Figure S7. Comparison of protein sequence coverage between the subsuperficial acid-etches ($n=13$, Stewart *et al.*¹⁶) and destructive sampling ($n=3$, this study).

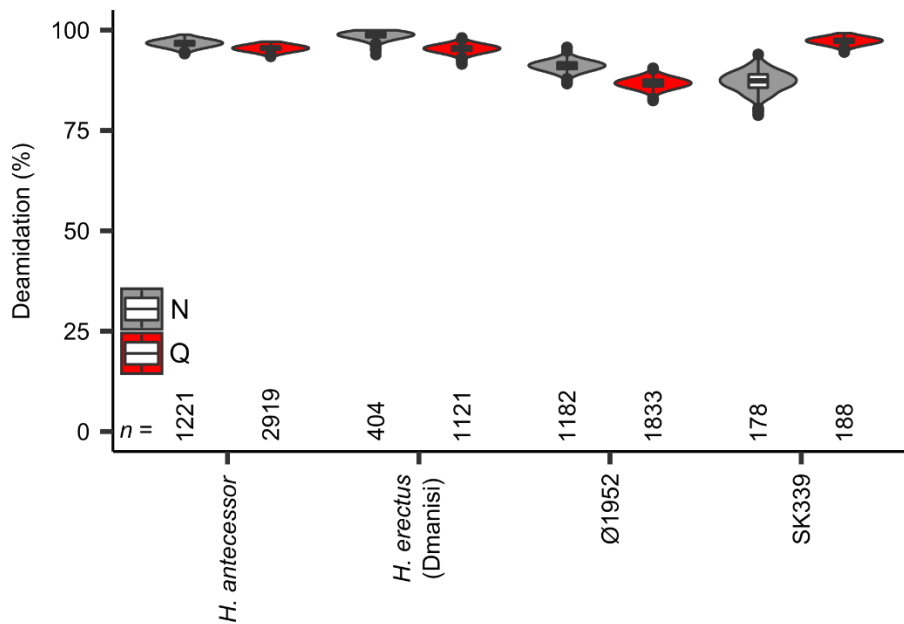


Figure S8. Enamel proteome deamidation of human controls and ancient hominin proteomes. Only peptides matching to AMELX, AMELY, AMBN, and ENAM are taken into account. Deamidation calculated following Mackie *et al.*³⁹ for 1,000 bootstrap replicates. The number of peptides (n) is given for each violin plot. 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). The boxplots define the range of the data (whiskers extending to 1.5 the interquartile range), 25th and 75th percentiles (boxes), and medians (dots).

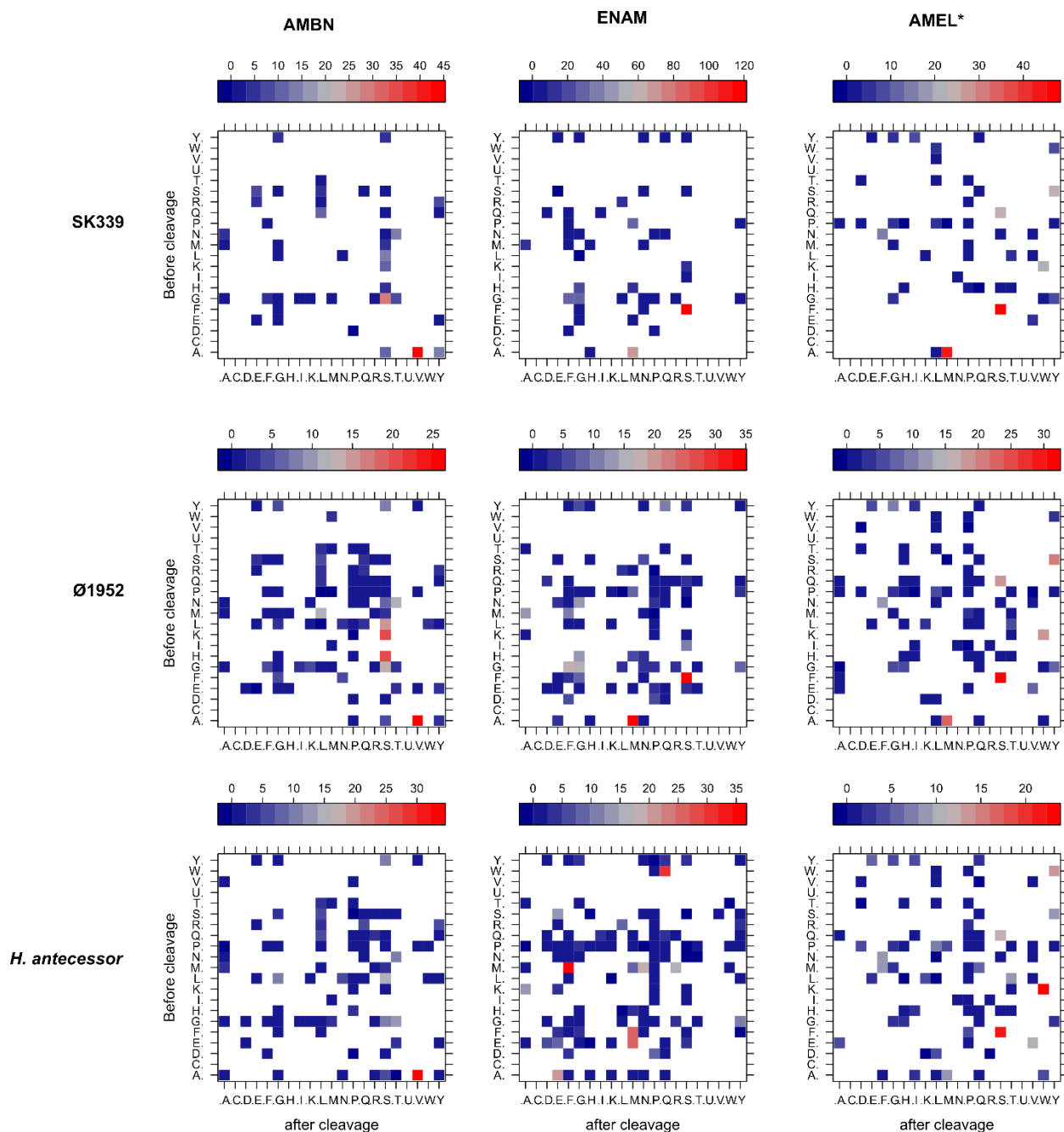


Figure S9. Cleavage matrices of SK339, Ø1952, and the *Homo antecessor* proteome for AMBN, ENAM, and AMELX/Y. Values indicate the fold difference between the observed cleavage frequency of each amino acid pair P1-P1' and its theoretical expectancy based on a random protein cleavage (PSMs). Counts based on peptide-spectrum-matches for each protein separately. The bottom row is identical to the top row in Extended Data Fig. 7. AMEL* contains PSMs matching to both AMELX and AMELY, with the theoretical matrix of both proteins joined in the ratio 9:1, respectively.

2.3 Phylogenetic analysis

Jazmín Ramos-Madrigal, Martin Kuhlwilm, Marc de Manuel, Tomas Marques-Bonet, Carles Lalueza-Fox, Eske Willerslev, Fernando Racimo

2.3.1 Reference data

We combined the new ancient protein sequences from *Homo erectus* and *Homo antecessor* with protein sequences from great apes^{41-43,97}, three Neanderthals^{31,40,55}, a Denisovan⁵⁶, and a panel of present-day humans, including 256 samples from the Simons Genome Diversity Panel (SGDP, IDs: ERZ312767-ERZ312784, ERZ312788-ERZ312789, ERZ312791-ERZ312797, ERZ312799-ERZ312848, ERZ324259-ERZ324404, ERZ324529-ERZ324546, ERZ324867-ERZ324876, ERZ325059, ERZ325062, ERZ329670-ERZ329671)⁵⁷ and 41 high-coverage individuals from the 1000 Genomes Project⁵⁸ (IDs: HG00096, HG00119, HG00183, HG00268, HG00419, HG00436, HG00640, HG00759, HG01051, HG01112, HG01136, HG01500, HG01565, HG01583, HG01595, HG01879, HG02568, HG02922, HG03006, HG03052, HG03642, HG03742, NA12413, NA12878, NA12891, NA12892, NA18525, NA18562, NA18939, NA18985, NA19017, NA19238, NA19239, NA19240, NA19625, NA19648, NA19685, NA19700, NA20502, NA20581, NA20845), representing worldwide populations (Tab. S7). Additionally, we included protein sequences from macaque (*Macaca mulatta*) and gibbon (*Nomascus leucogenys*) to root phylogenetic trees. When available, for some great apes, we obtained the protein sequences of interest from the UniProt database (Tab. S8), or, in the case of the Neanderthals and Denisovan, from published data⁴⁰. For the rest of the samples, including those from the SGDP and 1000 Genomes, we reconstructed the protein sequences from publicly available read alignments or genotype calls, as described below.

2.3.2 Reconstructing protein sequences from whole-genome sequencing data

Read alignments were obtained for 41 individuals from Phase3 of the 1000 Genomes Project (<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3>) and 256 individuals from the SGDP (ENA accession PRJEB9586). For each individual, we downloaded the reads overlapping the regions of interest using *samtools*⁹⁸ (v0.1.18) and built a majority-count-based consensus sequence using ANGSD⁹⁹ (v0.913); reads with mapping quality lower than 30 and bases with base quality lower than 20 were discarded. For great apes, VCF files with called diploid genotypes were provided by the authors of the corresponding original publications⁴¹⁻⁴³. In this case, we built FASTA files for each of the five genes of interest, for each sample. Indels were not considered and a random allele was chosen at heterozygous positions. For each of these consensus sequences, we identified and removed introns (*in silico* splicing) from each gene using the annotation of the reference human genome (h19), as provided by ENSEMBL¹⁰⁰. We then performed a *tblastn* search using the human reference protein as the query, and each of the previously 'spliced genes' as a subject, separately. Finally, protein sequences were obtained from the resulting translated alignments.

Table S7. Reference samples for which sequencing data was used to reconstruct the protein sequences.
Inds. = Individuals.

Species	Subspecies	Inds.	Common name	Reference(s)
<i>Gorilla beringei</i>	<i>Gorilla_beringei_beringei</i>	7	Mountain gorilla	Xue <i>et al.</i> ⁹⁷
	<i>Gorilla_beringei_grauer</i>	8	Eastern lowland gorilla	Prado-Martinez, <i>et al.</i> ⁴³ & Xue <i>et al.</i> ⁹⁷
<i>Gorilla gorilla</i>	<i>Gorilla_gorilla_diehli</i>	1	Cross River gorilla	Prado-Martinez, <i>et al.</i> ⁴³
	<i>Gorilla_gorilla_gorilla</i>	27	Western lowland gorilla	
<i>Homo sapiens</i>	<i>Homo_sapiens</i>	41	Present-day human	1000 Genomes Project ⁵⁸
		256		SGDP ⁵⁷
<i>Pan paniscus</i>	<i>Pan_paniscus</i>	10	Bonobo	Prado-Martinez, <i>et al.</i> ⁴³
<i>Pan troglodytes</i>	<i>Pan_troglodytes_elliotti</i>	10	Nigerian chimpanzee	Prado-Martinez, <i>et al.</i> ⁴³ , de Manuel, <i>et al.</i> ⁴¹ and Auton, <i>et al.</i> ¹⁰¹
	<i>Pan_troglodytes_schweinfurthii</i>	19	Eastern chimpanzee	
	<i>Pan_troglodytes_troglodytes</i>	18	Central chimpanzee	
	<i>Pan_troglodytes_verus</i>	12	Western chimpanzee	
	<i>Pan troglodytes verus/troglodytes</i>	1	Chimpanzee hybrid	
<i>Pongo pygmaeus</i>	<i>Pongo pygmaeus</i>	15	Bornean orangutan	Nater <i>et al.</i> ⁴² and Prado-Martinez, <i>et al.</i> ⁴³
<i>Pongo abelii</i>	<i>Pongo abelii</i>	11	Sumatran orangutan	
<i>Pongo tapanuliensis</i>	<i>Pongo tapanuliensis</i>	1	Tapanuli orangutan	
Ancient samples		Inds		Reference(s)
Denisovan		1		Castellano S, <i>et al.</i> ⁴⁰
Neanderthals		3		Prüfer K, <i>et al.</i> ^{31,55}

Table S8. Accession numbers of reference sequences obtained from the UniProt or Genbank databases.

Protein	ALB	AMBN	ENAM	COL17α1	MMP20	AMELY	AMELX
Chromosome (HG37)	4	4	4	10	11	Y	X
<i>Homo sapiens</i>	P02768	Q9NP70	Q9NRM1	Q9UMD9	O60882	Q99218	Q99217
<i>Pan paniscus</i>	XM_00383234 2.2	XM_00380899 2.1	B2L7U5	XM_00895087 6.1	XM_00382838 2.2		XM_00380567 8.1
<i>Pan troglodytes</i>	H2RBT1	H2R148	H2QPM0	H2Q2J4	H2Q4M8	Q861X8	A5JJS6
<i>Gorilla gorilla gorilla</i>	G3S791	G3RCU1	B2L7U8	G3QE20	G3QLA8	C3UJP7	G3SDK0
<i>Pongo abelii</i>	Q5NVH5	H2PDI5	H2PDI6	H2NBI5	H2NF32	A0A2J8W4N8	H2PUX0
<i>Nomascus leucogenys</i>	G1R8T8	G1R841	G1R843	G1RZC4	G1R6B1		G1RCS3
<i>Macaca mulatta</i>	Q28522	F7HLX4	F7H832		F7GQW6	A0A1D5RDA	A5JJS8

2.3.3. Protein alignments

We compared the protein sequences retrieved from the ancient samples and samples in the reference dataset. COL1 α 1 and COL1 α 2 were excluded from phylogenetic analysis, as their deamidation values could not exclude a contaminating origin of these proteins. AMTN was excluded since the retrieved sequences overlap a conserved sequence region with no phylogenetic SAPs within Hominidae. For the proteins considered endogenous and informative (AMBN, AMELX, AMELY, ENAM, MMP20, COL17 α 1, ALB), we present aligned fragment ion series and MS/MS spectra in the Supplementary Data 1 file.

For each of these proteins, and for each of the ancient samples (Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*), we generated two multiple sequence alignments using *mafft*⁵⁹: one including all individuals in Tables S5 and S6 and the corresponding ancient sample (Full alignments), and a second one including a single individual from each species/group (Single species alignments; Tab. S9). To account for isobaric amino acids (leucine and isoleucine), which cannot be distinguished with the methods used to obtain the protein sequences for ancient specimens, we did the following: 1) for positions where the ancient sample carried either a leucine or an isoleucine, and the reference samples carried only one of those amino acids, we set the ancient sample to the amino acid found in the reference samples, and 2) for positions where the ancient sample carried either a leucine or an isoleucine, and both amino acids were present among the reference samples, we set all individuals to carry a leucine.

Table S9. Description of the protein alignments used to compare the two ancient samples sequenced and the reference dataset.

Protein	Total sites	Non-missing sites in <i>Homo antecessor</i>	Unique amino acid substitutions in <i>Homo antecessor</i>	Polymorphic sites ³	Polymorphic sites (not singletons) ³	Polymorphic sites ³	Polymorphic sites (not singletons) ³	Substitutions uniquely shared between <i>H. antecessor</i> and any HND individual.
ALB	609	93	0	53 (10)	16 (2)	67 (13)	22 (4)	0
AMBN	447	169	0	64 (21)	16 (5)	77 (27)	35 (12)	0
AMELX	206	171	0	7 (4)	1 (0)	17 (12)	3 (2)	0
AMELY	206	141	0	28 (21)	17 (12)	36 (28)	26 (21)	0
COL17 α 1	1,498	112	1	62 (4)	21 (1)	124 (8)	62 (2)	0
ENAM	1,142	262	0	159 (31)	38 (8)	201 (39)	91 (16)	2
MMP20	483	66	0	32 (10)	8 (3)	50 (14)	21 (5)	1
Protein	Total sites	Non-missing sites in <i>Homo erectus</i>	Unique amino acid substitutions in <i>Homo erectus</i>	Polymorphic sites ³	Polymorphic sites (not singletons) ³	Polymorphic sites ³	Polymorphic sites (not singletons) ³	Substitutions uniquely shared between <i>H. erectus</i> and any HND individual.
ALB	609	245	0	53 (24)	16 (7)	67 (30)	22 (10)	0
AMBN	447	140	0	64 (18)	16 (7)	77 (23)	35 (12)	0
AMELX	206	182	0	7 (5)	1 (1)	17 (14)	3 (3)	0
COL17 α 1	1,498	67	0	61 (2)	21 (1)	123 (3)	62 (1)	0
ENAM	1,142	238	0	159 (26)	38 (4)	200 (37)	90 (12)	0
MMP20	483	99	0	32 (12)	8 (3)	50 (17)	21 (5)	0

¹ Single species alignment: built using a single individual per species/group. ² Full alignment: built using all individuals in the reference dataset (Tabs. S7, S8). ³ The numbers of sites where the ancient samples are non-missing are indicated in brackets. HND refers to any individual contained within the clade composed of *H. sapiens*, Neanderthals, and Denisovans.

a. Inspecting the protein alignments for informative amino acid substitutions in the ancient samples

We found one unique amino acid substitution in *Homo antecessor* and three substitutions that are uniquely shared among *Homo antecessor* and at least one other hominin (Neanderthal, Denisovan or present-day human; Tab. S9). In contrast, we found no substitutions that are unique to Dmanisi or are uniquely shared between Dmanisi and at least one other hominin. The unique amino acid substitution present in *Homo antecessor* requires one nucleotide change (Alanine (A) to Proline (P); codon GCT, most likely substituted to CCT; Tab. S10). To evaluate the plausibility of the unique amino acid substitution in *Homo antecessor*, we assessed whether the corresponding nucleotide position is segregating in any of the groups included in the full alignment and in the 1000 Genomes diversity panel. For the 1000 Genomes diversity panel, we used the Phase3 VCFs. We found that the three nucleotides coding for the amino acid present in great apes, including humans, are invariant; the A at position 1,089 in the protein *COL17A1* is coded by the GCT codon present in all samples in the alignment (Tab. S10). We further examined the frequency of this type of amino acid substitution in the set of proteins recovered from *Homo antecessor* and present in our protein alignments. We found 12 instances of an A-to-P substitution among the reference proteins analyzed. Alanine and proline are chemically compatible (they are both non-polar and hydrophobic), so this makes a plausible change to have occurred in the *Homo antecessor* lineage.

Finally, to get a sense of how likely it was to find one unique substitution in the *Homo antecessor* lineage given the sequences recovered, we estimated the average of unique substitutions in the modern and archaic hominin lineages. To do so, we counted the number of unique substitutions in each hominin sample when compared to *Homo antecessor*, *Gorilla*, *Pan*, and *Pongo* in the alignment comprising the seven concatenated proteins. In this case, we define an amino acid substitution as 'unique' in a given sample when present in that sample but not in *Homo antecessor*, *Gorilla*, *Pan*, or *Pongo*. To account for missing data in *Homo antecessor*, we only included the sites where the ancient sample is non-missing. We observed an average of 1.605 ± 0.68 unique substitutions for present-day humans, 2 unique substitutions for the Altai and Sidron Neanderthals and the Denisovan, and one unique substitution for the Vindija Neanderthal (Fig. S10).

b. Informative amino acid substitutions recovered in both Atapuerca *Homo antecessor* and Dmanisi *Homo erectus* samples

Among the proteins recovered for Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*, we obtained amino acid positions covered in both samples (Extended Data Fig. 4). Most of such positions correspond to non-segregating amino acids in hominids, or are singletons. These are not particularly informative for differentiating between groups. There are, however, three positions recovered for both hominins that are segregating in our alignment: position 281 in MMP20, 648 in *ENAM*, and 255 in *AMBN* (Tab. S11). The amino acid present at position 281 in the MMP20 protein is fixed to a threonine in *Pan* (n=72), *Gorilla* (n=44), and *Pongo* (n=28), Neanderthals (n=7) and the Denisovan, but it is segregating in present-day humans (both threonine (T) and glutamine (N) are present; n=302) (Tab. S11; Fig. S11a). In this position, both Dmanisi and *Homo antecessor* carry a threonine. Similarly, the amino acid present at position 648 in *ENAM* is fixed to a T (threonine) in *Pan* (n=72; with the exception of a single sample), *Gorilla* (n=44), and *Pongo* (n=28), but it is segregating in present-day humans (both threonine/isoleucine are present; n=302) and Neanderthals (n=7) and the Denisovan carry the derived amino acid (I). The sequence support of the single *Pan* individual carrying an isoleucine is restricted to just three sequence reads, and hence not well supported. Here, *Homo antecessor* carries an isoleucine - which is almost exclusively present in present-day humans, Neanderthals, while the Denisovan and Dmanisi carry a threonine (Tab. S11; Fig. S11b). Finally, the amino acid present at position 255 in the *AMBN* protein is fixed to an alanine in *Pan* (n=72), *Gorilla* (n=44), and *Pongo* (n=28), Neanderthals (n=6) and the Denisovan, but it is segregating in present-

day humans (both alanine (A) and valine (V) are present; n=302) (Tab. S11; Fig. S11c). In this case, both the Dmanisi *Homo erectus* and Atapuerca *Homo antecessor* carry an alanine.

Table S10. Nucleotide allele frequencies at the codons that translate to the amino acid where *Homo antecessor* carries a unique amino acid substitution.

	COL17α1- (chr10:105796813-105796815) A -> P ³		
Nucleotide	A	G	C
SGDP (n=256)	1.0	1.0	1.0
1000 genomes (n=41/2,504) ¹	1.0	1.0	1.0
Neanderthal (n=6) ²	1.0	1.0	1.0
Denisovan (n=1)	1.0	1.0	1.0
<i>Pan</i> ssp. (n=43)	1.0	1.0	1.0
<i>Gorilla</i> ssp. (n=70)	1.0	1.0	1.0
<i>Pongo</i> ssp. (n=27)	1.0	1.0	1.0

¹ None of these genomic sites are present in the 1000 genomes diversity panel. The frequencies of the alleles were obtained from the read alignments in the 41 high coverage genomes.

² Allele frequencies for the Neanderthal and Denisovan genomes were obtained from the read alignments in Castellano, *et al.*⁴⁰, Meyer, *et al.*⁵⁶, and Hajdinjak, *et al.*¹⁰². For samples in Hajdinjak, *et al.*, only the following samples were non-missing at the sites of interest: Mezmaiskaya_2, Les_Cottes_Z4-1514 and Goyet_Q56-1.

³ This amino acid substitution is only found in *Homo antecessor*. While all groups presented in the table carry the codon GCT (AGC, since the protein is coded in the reverse strand) which translates into Alanine (A), *Homo antecessor* most likely carries triplet CCT (AGG) that translates into Proline (P).

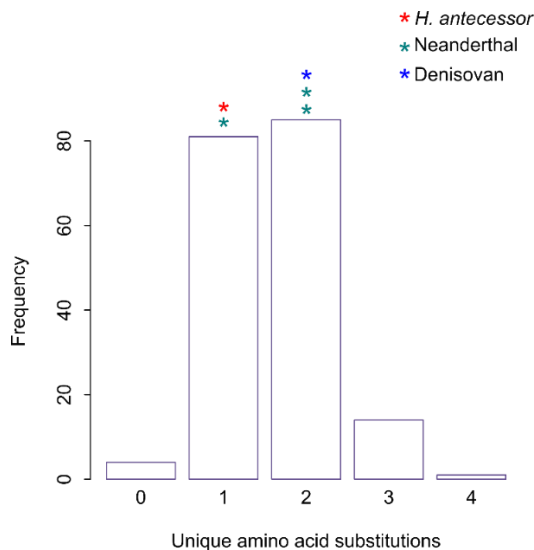


Figure S10. Frequency of unique amino acid substitutions among present-day humans. For each present-day human sample, we counted how many unique amino acid substitutions it carries when compared to *Homo antecessor*, *Gorilla*, *Pan*, and *Pongo*. Each bar corresponds to the number of present-day human samples that carry 0, 1, 2, 3 or 4 unique amino acid substitutions. The number of unique substitutions present in the archaic samples is indicated as stars in the top of the corresponding bar. This comparison only includes the sites where *Homo antecessor* is non-missing and the seven proteins recovered the ancient sample.

Table S11. Allele frequencies of the codons that translate to informative amino acid substitutions recovered for both Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*.

	ENAM+ (chr4:71509085-71509087) T->I			MMP20- (chr11:102477376-102477378) T->N			AMBN+ (chr4:71469603-71469605) A->V		
Nucleotide	A	C/T (rs7671281)	A	A	G/T (rs1784424)	T	G	C/T (rs7439186)	C
SGDP (n=256)	1.0	0.138/0.862	1.0	1.0	0.494/0.506	1.0	1.0	0.896/0.104	1.0
1000 genomes (n=41/2,504) ¹	1.0	0.182/0.818	1.0	1.0	0.580/0.420	1.0	1.0	0.881/0.119	1.0
Neanderthal (n=6-7) ²	1.0	0.0/1.0	1.0	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0
Denisovan (n=1)	1.0	0.0/1.0	1.0	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0
<i>Pan</i> ssp. (n=72)	1.0	0.977/0.023	1.0	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0
<i>Gorilla</i> ssp. (n=44)	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0
<i>Pongo</i> ssp. (n=28)	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0

¹ The allele frequencies at positions chr4:71509086 and chr11:102477377 were obtained from the 1000 Genomes diversity panel that comprise ~2,500 individuals. The frequencies of the invariant sites, that were not present in the 1000 Genomes VCFs, were obtained from the read alignments in the 41 high coverage genomes also from the 1000 genomes.

² Allele frequencies for the Neanderthal and Denisovan genomes were obtained from the read alignments in Castellano, *et al.*⁴⁰, Meyer, *et al.*⁵⁶, and Hajdinjak, *et al.*¹⁰². For samples in Hajdinjak, *et al.*, only the following samples were non-missing at the sites of interest: Spy_94a (only for ENAM and MMP20), Mezmaiskaya_2 (only for MMP20 and AMBN), Les_Cottes_Z4-1514 and Goyet_Q56-1 (only for ENAM and MMP20).

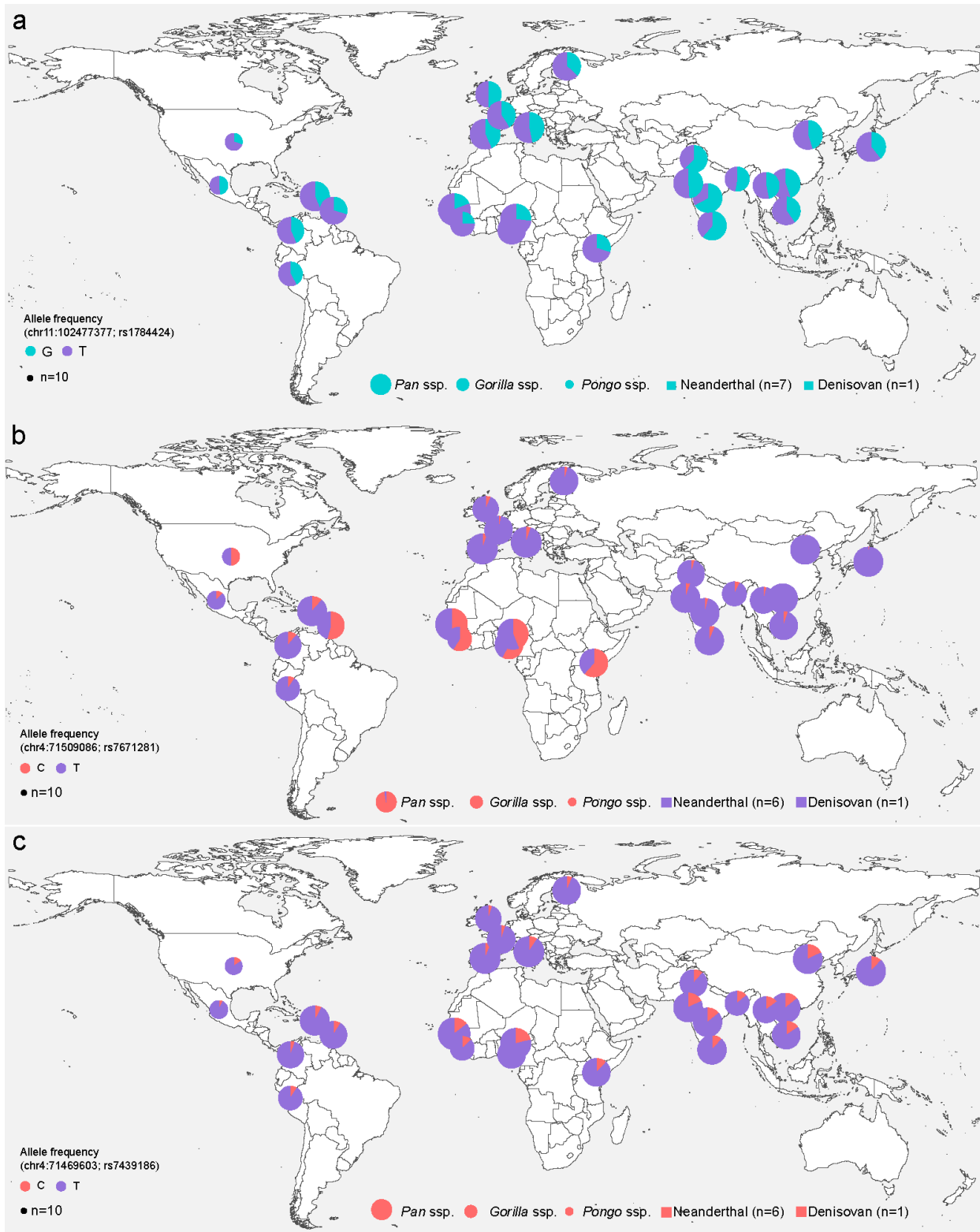


Figure S11. Allele frequencies in the 1000 Genomes diversity panel, among other great apes and in archaic hominins, at SNPs that code for three informative amino acids recovered for both Dmanisi *Homo erectus* and Atapuerca *Homo antecessor* samples. a, Genomic site corresponding to the second position of the codon that translates to a threonine (if G) or an asparagine (if T). This codon corresponds to position

281 in the MMP20 protein. Both Dmanisi *Homo erectus* and Atapuerca *Homo antecessor* samples carry a threonine at this position. **b**, Genomic site corresponding to the second nucleotide in the codon that translates into a threonine (if C) or an isoleucine (if T). This codon corresponds to position 648 of the ENAM protein. Dmanisi *Homo erectus* carries a threonine. While we cannot experimentally distinguish between isobaric amino acids (L/I) for *Homo antecessor*, it is most likely that the Atapuerca *Homo antecessor* carries an isoleucine, since only isoleucine is present among present-day human variation. **c**, Genomic site corresponding to the second position of the codon that translate to an alanine (if C) or a valine (if T). This codon corresponds to position 255 in the AMBN protein. Both Dmanisi *Homo erectus* and Atapuerca *Homo antecessor* samples carry an alanine at this position. The base map was generated using public domain data from <http://www.natureearthdata.com/>.

2.3.4. Pairwise-distance between the ancient and modern samples

We used a distance-based approach to compare the *Homo antecessor* and *Homo erectus* protein sequences with present-day humans, Neanderthals and the Denisovan. The latter three hominin groups are hereafter collectively referred to as the HND group. We estimated the pairwise distance between the Atapuerca *Homo antecessor* or Dmanisi *Homo erectus* and individuals of the HND group, while considering all the proteins retrieved for each ancient sample and excluding X and Y chromosome-located proteins (AMELX and AMELY). Pairwise distances between samples were estimated using the *phangorn* R package¹⁰³, considering the LG model¹⁰⁴ and pairwise-deletions. In this case, we used the full alignment comprising the complete set of samples in the dataset (multiple individuals per species).

The average distance between *Homo antecessor* and individuals from the HND group was 0.0036228 (0.0035712 when excluding AMELX and AMELY), whereas the average distance between members of the HND group is approximately three times smaller (0.0010827, or 0.0011429 when excluding AMELX and AMELY; Fig S9a, b). The distance between *Homo antecessor* and the HND group is significantly different to that between pairs of samples from the HND group (p -value<0.001; Mann-Whitney U Test). This suggests that the sequence is distantly related to the corresponding sequences in Neanderthals, Denisovans, and present-day humans.

In the case of Dmanisi *Homo erectus* individual, we obtained an average pairwise-distance between the ancient sample and individuals in the HND group of 0.0017148 (0.0020855, when excluding AMELX) and of 0.0011 (0.0011455 when excluding AMELX) between members of the HND group (Fig. S12c, d). In this case, we cannot exclude that Dmanisi belongs to the HND group, possibly due to the limited amount of informative positions retrieved for this sample, and the absence of a Dmanisi-unique SAP.

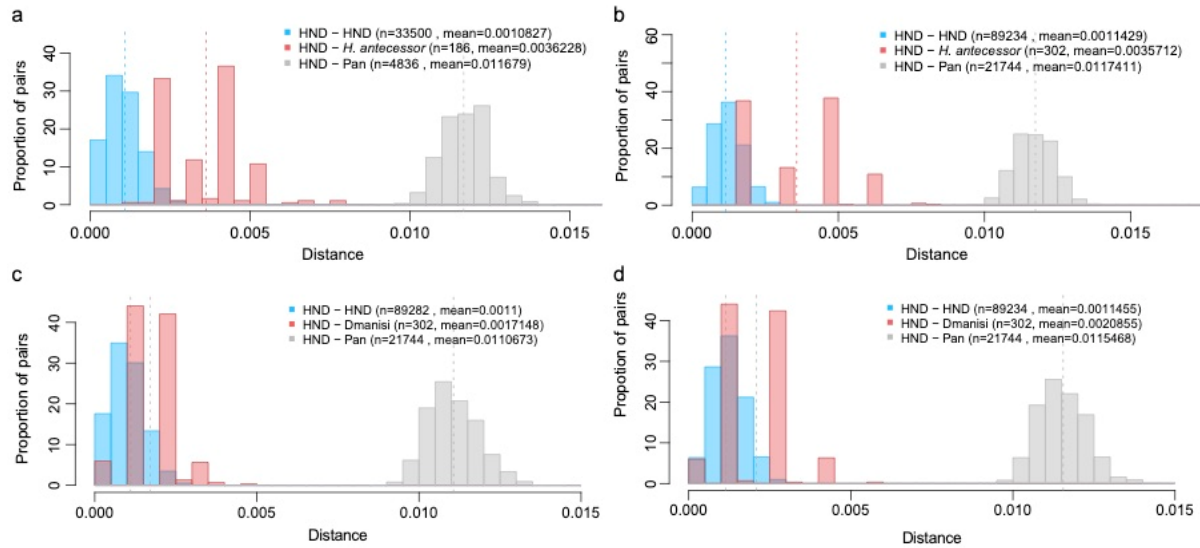


Figure S12. Pairwise-distances. **a**, Distribution of pairwise distances between individuals of the HND group (Humans, Neanderthals and Denisovan), *Homo antecessor*, and *Pan* for the seven proteins (ALB, AMBN, ENAM, MMP20, COL17 α 1, AMELX and AMELY), and **b**, excluding X- and Y-chromosome located proteins (AMELX and AMELY). **c**, Distribution of pairwise distances between individuals of the HND group, Dmanisi *Homo erectus*, and *Pan* for the six proteins (ALB, AMBN, ENAM, MMP20, COL17 α 1 and AMELX), and **d**, excluding the X-chromosome located protein (AMELX). The distance between pairs of samples is estimated based on the concatenated alignments and using the *phangorn* R package via the LG model and accounting for pairwise-deletions. HND refers to any individual contained within the clade composed of *H. sapiens*, Neanderthals, and Denisovans.

2.3.5. Phylogenetic placement of the ancient sequences

We sought to build phylogenetic trees using the aforementioned sequences, by applying two different approaches: a maximum likelihood and a Bayesian approach.

a. Maximum-likelihood (ML) approach

Using the single species alignments, we built ML trees for each protein and for a concatenated alignment of all seven of the available protein sequences using *PhyML* v3, for each of the ancient samples⁶⁰ (Figure S10 and S11). For each alignment we optimized the tree topologies, branch lengths and substitution rates (-o tlr) under the JTT model (-m JTT). Additionally, we obtained maximum likelihood estimates for the gamma distribution shape parameter (-a e) and the proportion of invariable sites (-v e). For each alignment, we started from three random trees and for each optimization step, we kept the best tree between those generated through the 'nearest neighbor interchange' and 'subtree prune and regraft' routines (-s BEST --rand_start --n_rand_starts 3). Support for each bipartition was obtained based on 100 non-parametric bootstrap replicates.

In the case of *Homo antecessor* (Fig. S13), each of the individual gene trees produced using the individual proteins provided poor resolution for resolving known species relationships. Four out of the seven gene trees place *Homo antecessor* within a hominin-only clade or as the closest outgroup to a hominin-only clade. The tree based on the concatenated alignment helps to resolve most of the known species relationships, and places the *Homo antecessor* sequence as the closest outgroup to Neanderthals, present-day humans and the Denisovan with 100% bootstrap support.

For the *Homo erectus* sample (Fig. S14), none of the individual gene trees produced a reliable topology for great ape relationships, and most inferred nodes have poor bootstrap support. In the tree based on the concatenated alignment, the Dmanisi *Homo erectus* is placed as the closest outgroup to present-day humans, Neanderthals and the Denisovan. However, the bootstrap support is, again, very poor (54%).

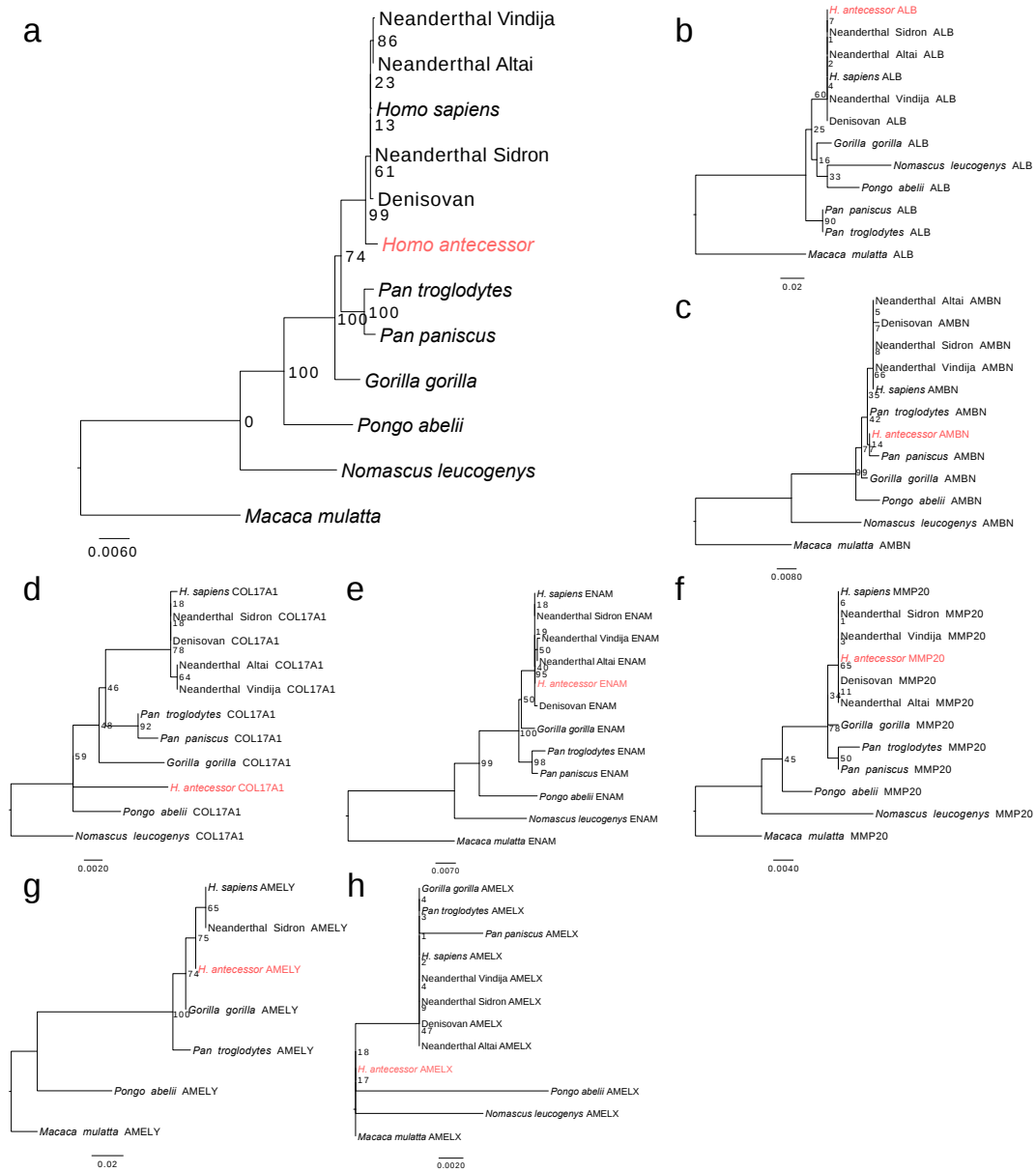


Figure S13. Atapuerca *Homo antecessor* maximum likelihood trees. a, Maximum-likelihood tree based on a concatenated alignment of the seven reconstructed *Homo antecessor* proteins. b-h, Individual protein trees. Support values for each bipartition were estimated through 100 non-parametric bootstrap replicates.

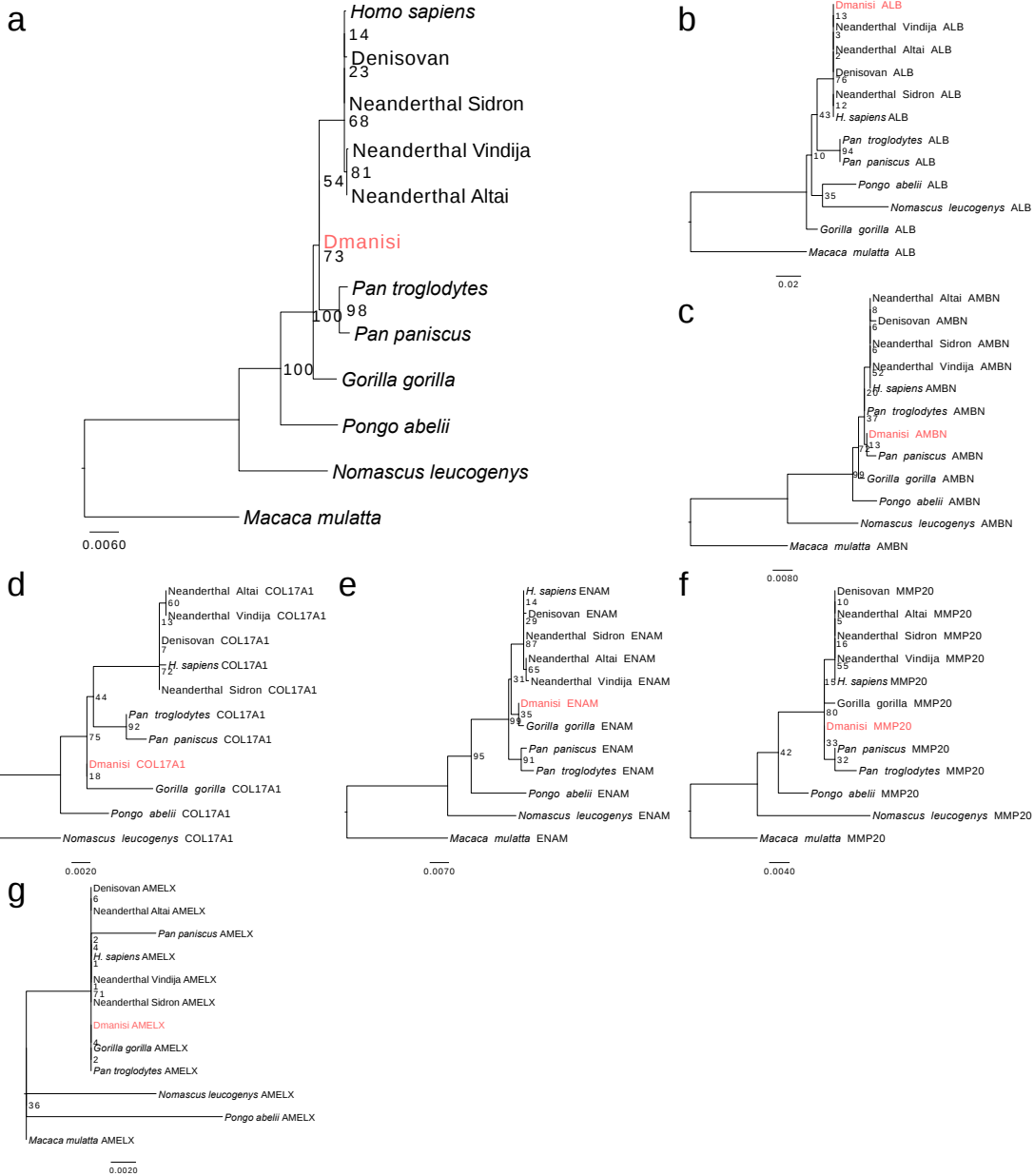
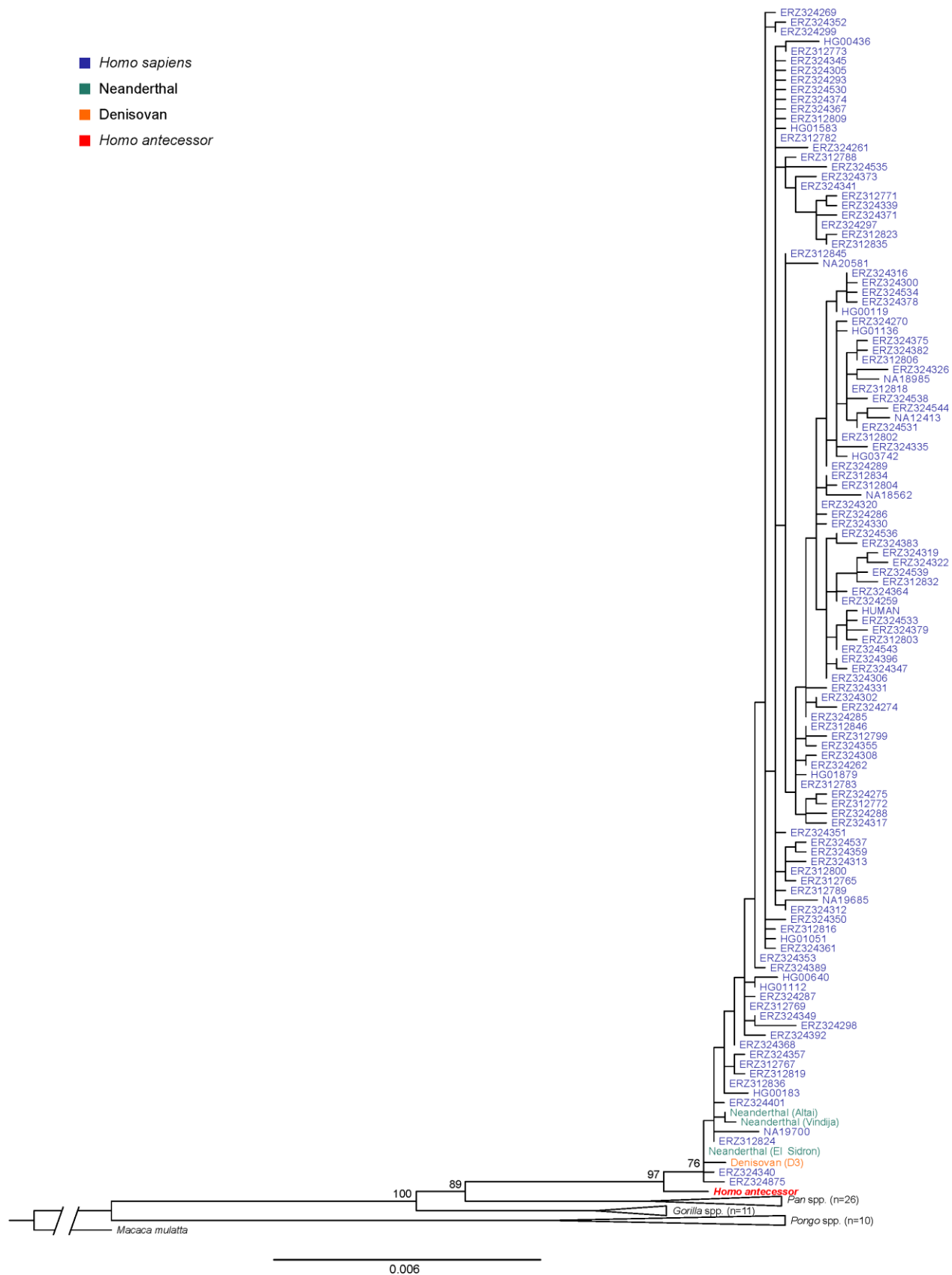


Figure S14. Dmanisi *Homo erectus* maximum likelihood trees. **a**, Maximum-likelihood tree obtained using PhyML and a concatenated alignment of the six reconstructed proteins for Dmanisi individual. **b-h**, Individual protein trees. Support values for each bipartition were estimated through 100 non-parametric bootstrap replicates.

b. Maximum-likelihood approach including all present-day human samples

In addition to the single-species tree, we estimated a maximum likelihood tree including all present-day human samples and other great apes present in the dataset for the Atapuerca *Homo antecessor* individual (Tab. S75). Note that, to include the AMELY protein recovered for *Homo antecessor*, we only included male individuals in the tree. We used the seven-protein concatenated alignment, PhyML v3 and the same parameters described above. Additionally, when more than one individual was identical across the seven proteins considered, we kept only one of them. Consistent with the single-species tree, the resulting phylogeny places *Homo antecessor* as an outgroup to all individuals in the HND clade (Fig. S15). However, the support for this bipartition is smaller: 41% of the trees place *Homo antecessor* as an outgroup to the HND clade, while in the remaining 59% *Homo antecessor* is placed within the HND clade.

We then tested, whether using different combinations of present-day human samples in the phylogenetic inference affects the position of *Homo antecessor*. We created 1,000 different alignments, each containing 7 randomly sampled present-day humans from the 305 in our dataset. For each alignment, we used PhyML v3 and the same parameters described above, and evaluated the position obtained for *Homo antecessor* with the respect to the HND clade. In 98.7% of the trees, *Homo antecessor* is placed as an outgroup to the HND clade, while in the remaining 1.3% it is placed within Neanderthals, forming a clade with the Vindija Neanderthal.



Previous page. **Figure S15. Maximum-likelihood tree based on a concatenated alignment of the seven protein sequences recovered for the Atapuerca *Homo antecessor* individual.** Support values were estimated through 100 non-parametric bootstrap replicates.

c. Effect of missing data on the phylogenetic inference

To assess the effect of missing data in *Homo antecessor*, and whether missing data could artificially yield the phylogenetic result we observe in the ancient sample, we performed a downsampling experiment by adding missing data to one present-day human sample (ERZ324268) and the Altai Neanderthal. For each of those samples, we created 100 independent replicates, where each contains a similar amount of missing data to *Homo antecessor*, distributed randomly in blocks of similar size as observed in the ancient sample. For each sample and replicate, we created a concatenated protein alignment consisting of the given replicate and all samples present in the single-species dataset, excluding the ancient sample. We then used *PhyML* (with the parameters described in the section above) to estimate a ML tree for each replicate and sample, and compared the resulting topologies to the topology obtained using the same alignment without missing data. For sample ERZ324268, we recover a topology similar to the one obtained without missing data in 99% of the replicates, while in 1% of the replicates the sample forms a polytomy with the Denisovan and the Sidron Neanderthal. For the Altai Neanderthal, we recover the topology obtained without missing data in 43% of the cases, while 57% of the replicates result in a polytomy involving all hominin samples.

d. Bayesian phylogenetic inference

To assess the robustness of the ML inference results, we also performed Bayesian phylogenetic inference based on the concatenated alignments using *mrBayes*⁶¹ (v3.2). In this case, we partitioned the alignments by gene and for each partition we estimated the substitution rates, the shape parameter of a gamma distribution to model across-site rate variation (exponentially distributed prior), and the proportion of invariable sites (uniformly distributed prior; unlink Statefreq=(all) Ratemultiplier=(all) Aamodel=(all) Shape=(all) Pinvar=(all)). The tree topology and branch lengths were inferred for all the partitions jointly. To obtain posterior probabilities of unknown parameters, we ran a Metropolis-coupled Markov-chain Monte Carlo (MCMCMC) algorithm with four chains and a temperature parameter of 0.2 for 5,000,000 cycles sampling every 500 steps, after which we discarded the first 1,250,000 runs as burn-in. Convergence of the algorithm was evaluated using Tracer⁶⁵ (v.1.7.0). In particular, we required that the effective sample sizes for each estimated parameter were greater than 200, and that the overall log-likelihood of the run did not fluctuate substantially. Bayesian inference performed via *mrBayes* was performed using the CIPRES Science Gateway⁶³. In agreement with the maximum-likelihood approach, the Atapuerca *Homo antecessor* and Dmanisi *Homo erectus* individuals are placed as an outgroup to the HND clade with a posterior probability of 1 and 0.613, respectively (Fig. S16; Extended Data Fig. 8).

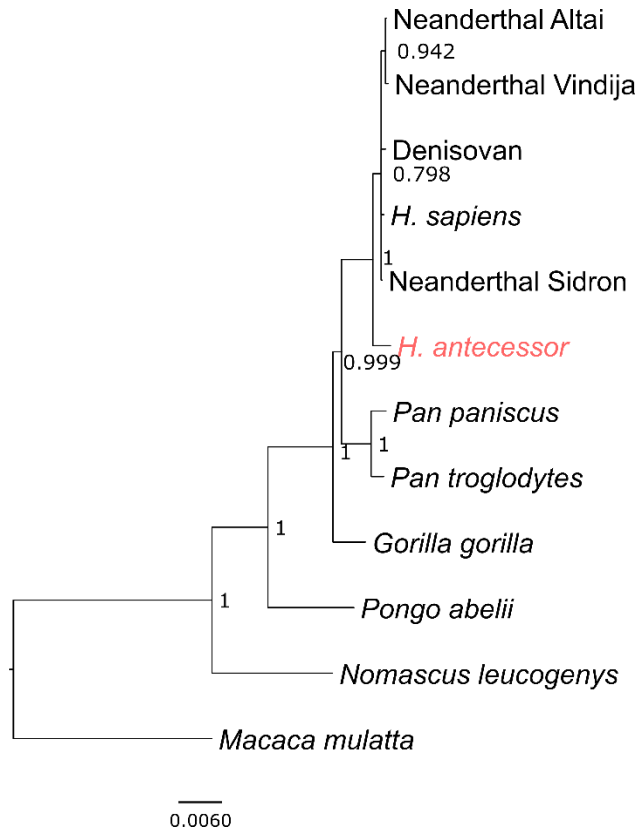


Figure S16. *Homo antecessor* Bayesian phylogenetic tree. Tree obtained using the seven-protein concatenated alignment and MrBayes. Posterior probabilities are indicated for each bipartition. *Macaca mulatta* was used as outgroup.

2.3.6. BEAST analysis

We used the Bayesian approach implemented in BEAST 2.5⁶² to infer sequence divergence times for the proteins sequenced and between the different groups in the dataset and *Homo antecessor*. For this analysis, we used the seven-protein concatenated alignment including the Neanderthals, the Denisovan, seven randomly chosen *Homo sapiens* sequences 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 ka⁵⁵, Altai Neanderthal: 112 ka⁵⁵, Denisovan: 72 ka⁵⁵ and *Homo antecessor* 860.5 ka¹¹) were used as tip dates for calibration. For each of the seven partitions 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*). Note that the overall topology of the tree was estimated for the seven partitions jointly. We assessed the convergence of the algorithm using Tracer⁶⁵ v1.7.0. Finally, since including all present-day humans in the analyses led to convergence problems, we instead repeated the analyses for 100 alignments, each of them consisting of seven different present-day humans chosen randomly. While the topology within the clade comprising present-day humans, Neanderthals and Denisovan (HND) was not consistent in all of the 100 alignments, 99 of them consistently placed *Homo antecessor* as an outgroup to the HND clade. The one alignment that did not place *Homo antecessor* as an outgroup, it placed it as an outgroup to Neanderthals and the Denisovan.

The inferred tree for one of the 99 replicates is shown in Figure 3a. We note that the estimated tree in Figure 3a does not place Neanderthals as a monophyletic clade, which is likely due to incomplete lineage sorting, and the poor phylogenetic resolution afforded by sampling only a few gene trees. We assessed the proportion of sampled trees from the MCMC chain in which the inferred HND-*Homo antecessor* divergence was more ancient than that estimated for the HND clade for each of the 100 replicates. We found that in 95% of the trees, the divergence of *Homo antecessor* predated the first divergence within the HND clade (Fig. 2b).

These estimates rely on gene trees reconstructed from only seven protein sequences for which it was possible to obtain data for *Homo antecessor*. We therefore caution that these divergence times only reflect an approximation to the average genetic divergence time. They are likely an upper boundary for the population split time between these groups, which is expected to be more recent than the average genomic divergence. Nevertheless, assuming incomplete lineage sorting is not prevalent in these gene trees, our results support the placement of *Homo antecessor* as an outgroup to the HND group.

References

- 1 Gabunia, L. *et al.* Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: taxonomy, geological setting, and age. *Science* **288**, 1019-1025, doi:10.1126/science.288.5468.1019 (2000).
- 2 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).
- 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, J. J. 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).
- 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).
- 9 Bermúdez de Castro, J. M. *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).
- 10 Carbonell, E. *et al.* Lower Pleistocene hominids and artifacts from Atapuerca-TD6 (Spain). *Science* **269**, 826-830, doi:10.1126/science.7638598 (1995).
- 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, S. E., 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).
- 13 Lacruz, R. S. *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).
- 16 Stewart, N. A., Gerlach, R. F., Gowland, R. L., Gron, K. J. & Montgomery, J. Sex determination of human remains from peptides in tooth enamel. *Proceedings of the National Academy of Sciences of the United States* **114**, 13649-13654, doi:10.1073/pnas.1714926115 (2017).
- 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).
- 18 Castiblanco, G. A. *et al.* Identification of proteins from human permanent erupted enamel. *European Journal of Oral Sciences* **123**, 390-395, doi:10.1111/eos.12214 (2015).

- 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).
- 20 Porto, I. M., Laure, H. J., de Sousa, F. B., Rosa, J. C. & Gerlach, R. F. 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).
- 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, V. S. *et al.* Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. *Science* **336**, 1150-1153, doi:10.1126/science.1217817 (2012).
- 24 Hu, J. C. C., Yamakoshi, Y., Yamakoshi, F., Krebsbach, P. H. & Simmer, J. P. Proteomics and Genetics of Dental Enamel. *Cells Tissues Organs* **181**, 219-231, doi:10.1159/000091383 (2005).
- 25 Glimcher, M. J., 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, G. A. *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 Martínón-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, J. M., Martínón-Torres, M., Arsuaga, J. L. & 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, J. M., Arsuaga, J.-L., Carbonell, E. & Polly, P. D. 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).
- 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).
- 32 Lacruz, R. S. *et al.* The evolutionary history of the human face. *Nature Ecology & Evolution* **3**, 726-736, doi:10.1038/s41559-019-0865-7 (2019).
- 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, R. C. *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).
- 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).
- 36 Penkman, K. E. H., Kaufman, D. S., Maddy, D. & Collins, M. J. Closed-system behaviour of the intra-crystalline fraction of amino acids in mollusc shells. *Quaternary Geochronology* **3**, 2-25, doi:10.1016/j.quageo.2007.07.001 (2008).

- 37 Dickinson, M., Lister, A. M. & Penkman, K. E. H. 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, R. L. Hydrolysis of proteins. *Advances in Protein Chemistry* **20**, 37-107 (1965).
- 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).
- 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).
- 43 Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471-475, doi:10.1038/nature12228 (2013).
- 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).
- 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).
- 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, Y. H. P. *et al.* Cleavage Site Specificity of MMP-20 for Secretory-stage Ameloblastin. *Journal of Dental Research* **89**, 785-790, doi:10.1177/0022034510366903 (2010).
- 50 Yamakoshi, Y., Hu, J. C. C., Fukae, M., Yamakoshi, F. & Simmer, J. P. 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).
- 51 Iwata, T. *et al.* Processing of Ameloblastin by MMP-20. *Journal of Dental Research* **86**, 153-157, doi:10.1177/154405910708600209 (2007).
- 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).
- 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).
- 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).
- 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).
- 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).

- 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).
- 58 The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*
526, 68-74, doi:10.1038/nature15393 (2015).
- 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).
- 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).
- 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).
- 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, M. A., 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, M. H. Direct estimation
of mutations in great apes reconciles phylogenetic dating. *Nature Ecology & Evolution* **3**, 286-292,
doi:10.1038/s41559-018-0778-x (2019).
- 65 Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in
Bayesian Phylogenetics Using Tracer 1.7. *Systematic biology* **67**, 901-904,
doi:10.1093/sysbio/syy032 (2018).
- 66 Olivé, A., Ramirez-Merino, J. L. & Ortega, L. I. (I.T.G.E. , Madrid, 1990).
- 67 Pineda, A. & Arce, J. M. (I.T.G.E., Madrid, 1997).
- 68 Pérez-González, A. *et al.* in *Human Evolution in Europe and the Atapuerca evidence* (eds J.M.
Bermúdez de Castro, J.L. Arsuaga, & E Carbonell) 99-122 (Junta de Castilla y León, Consejería de
Cultura y Turismo, 1995).
- 69 Parés, J. M. & Pérez-González, A. Magnetochronology and stratigraphy at Gran Dolina section,
Atapuerca (Burgos, Spain). *Journal of Human Evolution* **37**, 325-342, doi:10.1006/jhev.1999.0331
(1999).
- 70 Zazo, C., Goy, J. L. & Hoyos, M. Estudio geomorfológico de los alrededores de la Sierra de
Atapuerca (Burgos). *Estudios geológicos* **39**, 179-185 (1983).
- 71 Parés, J. M. & Pérez-González, A. Paleomagnetic age for hominid fossils at Atapuerca
archaeological site, Spain. *Science* **269**, 830-832, doi:10.1126/science.7638599 (1995).
- 72 Ortega-Martínez, A. I. *La evolución geomorfológica del karst de la sierra de Atapuerca (Burgos) y
su relación con los yacimientos pleistocenos que contiene* PhD thesis, University of Burgos, (2009).
- 73 Gil, E., Aguirre, E. & Hoyos, M. in *El hombre fósil de Ibeas y el Pleistoceno de la Sierra de Atapuerca*
(ed E. Aguirre, Carbonell, E. & Bermúdez de Castro, J.M.) (Junta de Castilla y León, Consejería de
Cultura y Turismo, 1987).
- 74 Carbonell, E. *et al.* The Pleistocene site of Gran Dolina, Sierra de Atapuerca, Spain: a history of the
archaeological investigations. *Journal of Human Evolution* **37**, 313-324,
doi:10.1006/jhev.1999.0282 (1999).
- 75 Campaña, I. *et al.* New interpretation of the Gran Dolina-TD6 bearing Homo antecessor deposits
through sedimentological analysis. *Scientific Reports* **6**, 34799, doi:10.1038/srep34799 (2016).
- 76 Bermúdez de Castro, J. M. *et al.* Early pleistocene human humeri from the gran dolina-TD6 site
(sierra de atapuerca, spain). *American Journal of Physical Anthropology* **147**, 604-617,
doi:10.1002/ajpa.22020 (2012).

- 77 Cuenca-Bescós, G., Laplana, C. & Canudo, J. I. Biochronological implications of the Arvicolidae (Rodentia, Mammalia) from the Lower Pleistocene hominid-bearing level of Trinchera Dolina 6 (TD6, Atapuerca, Spain). *Journal of Human Evolution* **37**, 353-373, doi:10.1006/jhev.1999.0306 (1999).
- 78 van der Made, J. Ungulates from Atapuerca TD6. *Journal of Human Evolution* **37**, 389-413, doi:10.1006/jhev.1998.0264 (1999).
- 79 García, N. & Arsuaga, J. L. Carnivores from the Early Pleistocene hominid-bearing Trinchera Dolina 6 (Sierra de Atapuerca, Spain). *Journal of Human Evolution* **37**, 415-430, doi:10.1006/jhev.1999.0325 (1999).
- 80 Carbonell, E. *et al.* The TD6 level lithic industry from Gran Dolina, Atapuerca (Burgos, Spain): production and use. *Journal of Human Evolution* **37**, 653-693, doi:10.1006/jhev.1999.0336 (1999).
- 81 Bermúdez de Castro, J. M. *et al.* The TD6 (Aurora stratum) hominid site. Final remarks and new questions. *Journal of Human Evolution* **37**, 695-700, doi:10.1006/jhev.1999.0334 (1999).
- 82 Álvarez-Posada, C. *et al.* A post-Jaramillo age for the artefact-bearing layer TD4 (Gran Dolina, Atapuerca): New paleomagnetic evidence. *Quaternary Geochronology* **45**, 1-8, doi:10.1016/j.quageo.2018.01.003 (2018).
- 83 Falguères, C. *et al.* Earliest humans in Europe: the age of TD6 Gran Dolina, Atapuerca, Spain. *Journal of Human Evolution* **37**, 343-352, doi:10.1006/jhev.1999.0326 (1999).
- 84 Berger, G. W. *et al.* Luminescence chronology of cave sediments at the Atapuerca paleoanthropological site, Spain. *Journal of Human Evolution* **55**, 300-311, doi:10.1016/j.jhevol.2008.02.012 (2008).
- 85 Moreno, D. *et al.* New radiometric dates on the lowest stratigraphical section (TD1 to TD6) of Gran Dolina site (Atapuerca, Spain). *Quaternary Geochronology* **30**, 535-540, doi:10.1016/j.quageo.2015.05.007 (2015).
- 86 Arnold, L. J. *et al.* Evaluating the suitability of extended-range luminescence dating techniques over early and Middle Pleistocene timescales: Published datasets and case studies from Atapuerca, Spain. *Quaternary International* **389**, 167-190, doi:10.1016/j.quaint.2014.08.010 (2015).
- 87 Arnold, L. J. & Demuro, M. Insights into TT-OSL signal stability from single-grain analyses of known-age deposits at Atapuerca, Spain. *Quaternary Geochronology* **30**, 472-478, doi:10.1016/j.quageo.2015.02.005 (2015).
- 88 Parés, J. M. *et al.* Chronology of the cave interior sediments at Gran Dolina archaeological site, Atapuerca (Spain). *Quaternary Science Reviews* **186**, 1-16, doi:10.1016/j.quascirev.2018.02.004 (2018).
- 89 Cuenca-Bescós, G. *et al.* Comparing two different Early Pleistocene microfaunal sequences from the caves of Atapuerca, Sima del Elefante and Gran Dolina (Spain): Biochronological implications and significance of the Jaramillo subchron. *Quaternary International* **389**, 148-158, doi:10.1016/j.quaint.2014.12.059 (2015).
- 90 Molnar, S. Human tooth wear, tooth function and cultural variability. *American Journal of Physical Anthropology* **34**, 175-189, doi:10.1002/ajpa.1330340204 (1971).
- 91 Kaufman, D. S. & Manley, W. F. A new procedure for determining DL amino acid ratios in fossils using reverse phase liquid chromatography. *Quaternary Science Reviews* **17**, 987-1000, doi:10.1016/S0277-3791(97)00086-3 (1998).
- 92 Sykes, G. A., Collins, M. J. & Walton, D. I. The significance of a geochemically isolated intracrystalline organic fraction within biominerals. *Organic Geochemistry* **23**, 1059-1065, doi:10.1016/0146-6380(95)00086-0 (1995).

- 93 Preece, R. C. & Penkman, K. E. H. New faunal analyses and amino acid dating of the Lower Palaeolithic site at East Farm, Barnham, Suffolk. *Proceedings of the Geologists' Association* **116**, 363-377, doi:10.1016/S0016-7878(05)80053-7 (2005).
- 94 Stewart, N. A. *et al.* The identification of peptides by nanoLC-MS/MS from human surface tooth enamel following a simple acid etch extraction. *RSC Advances* **6**, 61673-61679, doi:10.1039/c6ra05120k (2016).
- 95 Fincham, A. G. *et al.* Human developing enamel proteins exhibit a sex-linked dimorphism. *Calcified Tissue International* **48**, 288-290, doi:10.1007/BF02556382 (1991).
- 96 Parker, G. J. *et al.* Sex estimation using sexually dimorphic amelogenin protein fragments in human enamel. *Journal of Archaeological Science* **101**, 169-180, doi:10.1016/j.jas.2018.08.011 (2019).
- 97 Xue, Y. *et al.* Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **348**, 242-245, doi:10.1126/science.aaa3952 (2015).
- 98 Li, H. *et al.* The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **25**, doi:10.1093/bioinformatics/btp352 (2009).
- 99 Korneliussen, T., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356-356, doi:10.1186/s12859-014-0356-4 (2014).
- 100 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic acids research* **46**, D754-D761, doi:10.1093/nar/gkx1098 (2018).
- 101 Auton, A. *et al.* A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science* **336**, 193-198, doi:10.1126/science.1216872 (2012).
- 102 Hajdinjak, M. *et al.* Reconstructing the genetic history of late Neanderthals. *Nature* **555**, 652-656, doi:10.1038/nature26151 (2018).
- 103 Schliep, K., Potts, A. J., Morrison, D. A. & Grimm, G. W. Intertwining phylogenetic trees and networks. *Methods in Ecology and Evolution* **8**, 1212-1220, doi:10.1111/2041-210X.12760 (2017).
- 104 Le, S. Q. & Gascuel, O. An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution* **25**, 1307-1320, doi:10.1093/molbev/msn067 (2008).