Multi-protease analysis of Pleistocene bone proteomes

Liam T. Lanigan\textsuperscript{a}, Meaghan Mackia\textsuperscript{a,b}, Susanne Feine\textsuperscript{c}, Jean-Jacques Hublin\textsuperscript{d,e}, Ralf W. Schmitz\textsuperscript{c}, Arndt Wilckeb, Matthew J. Collins\textsuperscript{a,g}, Enrico Cappellini\textsuperscript{i}, Jesper V. Olsen\textsuperscript{b}, Alberto J. Taurozzi\textsuperscript{a,e}, Frido Welker\textsuperscript{h,i,e}

\textsuperscript{a} Evolutionary Genomics Section, Globe Institute, University of Copenhagen, Copenhagen, Denmark
\textsuperscript{b} Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark
\textsuperscript{c} LVR-LandesMuseum, Bonn, Germany
\textsuperscript{d} Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
\textsuperscript{e} Collège de France, Paris, France
\textsuperscript{f} Praunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany
\textsuperscript{g} Department of Archaeology, University of Cambridge, Cambridge, United Kingdom

Abstract

Ancient protein analysis is providing new insights into the evolutionary relationships between hominin fossils across the Pleistocene. Protein identification commonly relies on the proteolysis of a protein extract using a single protease, trypsin. As with modern proteome studies, alternative or additional proteases have the potential to increase both proteome size and protein sequence recovery. This could enhance the recovery of phylogenetic information from ancient proteomes. Here we identify 18 novel hominin bone specimens from the Kleine Feldhofer Grotte using MALDI-TOF MS peptide mass fingerprinting of collagen type I. Next, we use one of these hominin bone specimens and three Late Pleistocene Equidae specimens identified in a similar manner and present a comparison of the bone proteome size and protein sequence recovery obtained after using nanoLC-MS/MS and parallel proteolysis using six different proteases, including trypsin. We observe that the majority of the preserved bone proteome is inaccessible to trypsin. We also observe that for proteins recovered consistently across several proteases, protein sequence coverage can be increased significantly by combining peptide identifications from two or more proteases. Our results thereby demonstrate that the proteolysis of Pleistocene proteomes by several proteases has clear advantages when addressing evolutionary questions in palaeoproteomics. Significance: Maximizing proteome and protein sequence recovery of ancient skeletal proteomes is important when analyzing unique hominin fossils. As with modern proteome studies, palaeoproteomic analysis of Pleistocene bone and dentine samples has almost exclusively used trypsin as its only protease, despite the demonstrated advantages of alternative proteases to increase proteome recovery in modern proteome studies. We demonstrate that Pleistocene bone proteomes can be significantly expanded by using additional proteases beside trypsin, and that this also improves sequence coverage of individual proteins. The use of several alternative proteases beside trypsin therefore has major benefits to maximize the phylogenetic information retrieved from ancient skeletal proteomes.

1. Introduction

The study of ancient proteins, or palaeoproteomics, has recently seen its first applications to human origin studies. These range from the identification of faunal complexes associated with hominin occupancy through MALDI-TOF MS peptide mass fingerprinting of collagen type I, to the study of hominin phylogeny based on entire ancient skeletal proteomes [1–9]. Such palaeoproteomic studies primarily analyze the proteins preserved in the skeletal tissues (bone, dentine, enamel) from hominin fossils directly.

Bone and dentine proteomes are dominated by collagen type I (COL1), which comprises roughly 90% of the total organic content of bone. This dominance of COL1 is maintained in Pleistocene bone proteomes [10,11]. The COL1A1 and COL1A2 chains that compose the mature COL1 protein are largely composed of G-X-Y repeats, in which X and Y are frequently proline and hydroxyproline, respectively [12]. The maintenance of these G-X-Y repeats is required for proper collagen fibril formation [12,13]. Despite these structural and functional constraints on the COL1 triple helical regions, COL1 is a phylogenetically informative biomolecule that has been used in a number of studies on extinct mammals [14–16]. However, its utility to disentangle the phylogenetic relationships between closely related species or populations is limited due to its slow rate of evolution, a consequence of the restrictions on sequence variability within G-X-Y repeats.

In addition to collagens, the bone and dentine proteomes also contain a large number of non-collagenous proteins (NCPs; [10,17,18]). As an assemblage the NCPs have a heterogenous origin, including proteins deriving from the bone extracellular matrix, cellular proteins and plasma proteins. NCPs are commonly identified by fewer peptides when compared to COL1, and hence have lower sequence coverage.
However, they are more numerous in absolute protein numbers, numbering several hundreds to a few thousand different proteins in fresh tissues [19]. Some of these NCPs are also less evolutionarily conserved when compared to COL1 [20]. Therefore, theoretically, NCPs provide access to a larger, more diverse set of phylogenetic information.

As with most modern proteome studies, palaeoproteomic studies of bone proteomes almost exclusively use trypsin as the single digestive protease [21]. It has recently been demonstrated that the use of additional proteases, or different combinations of proteases, has the potential to access different “extractomes” of the same proteome, thereby significantly increasing both the total number of proteins identified and the combined protein sequence coverage obtained [22–24]. Therefore, the use of multiple proteases could have major benefits to the phylogenetic study of ancient proteomes. Alternative proteolysis strategies have so far been explored to a limited extent in palaeoproteomics. Examples include, among others, the cleavage of collagen with collagenase prior to trypsin proteolysis [11], parallel proteolysis with elastase and trypsin [14,25], and consecutive proteolysis using Lys-C and trypsin [26]. Despite their potential for improving protein identifications and/or sequence coverage in palaeoproteomic studies, a systematic study of the effects of alternative or additional proteases has so far not been formally investigated.

Here, our primary aim is to explore the use of five alternative proteases in addition to trypsin applied to four Late Pleistocene bone samples. We selected these four bone specimens (three horses and one hominin) randomly after MALDI-TOF MS peptide mass fingerprinting conducted at three different Late Pleistocene archaeological sites located in western Europe (Quinçay, La Ferrassie, the Kleine Feldhofer Grotte; Fig. 1). Equid1 was identified as part of previous MALDI-TOF MS COL1 peptide mass fingerprinting of morphologically unidentifiable bone specimens from Quinçay, France [27]. It derives from the Châtelperronian layer E at the site, and was recovered between 1968 and 1990 (excavations led by F. Lévêque). No recent radiocarbon ages are available for the site or its Châtelperronian layers, but the specimen can be expected to date between 40 and 50 ka BP [28]. Equid2 represents a non-diagnostic bone specimen from the Châtelperronian layer 6 at La Ferrassie, France, recovered during recent excavations (directed by Alain Turq). It is identified as Equidae as part of ongoing COL1 peptide mass fingerprinting of the bone assemblage deriving from layer 6. This layer is dated to approximately 42 ± 3 ka BP [29,30]. Finally, Equid3 and the Homo sp. specimens derive from sediment attributed to the Kleine Feldhofer Grotte, Germany. They were identified as part of a larger COL1 peptide mass fingerprinting study of 76 bone and dentine specimens that could represent potential hominin bone specimens (reported below). The deposits from the Kleine Feldhofer Grotte were initially removed during August 1856, resulting in the discovery and description of Homo neanderthalensis [31–33]. Archival research and archaeological excavations led by R.W. Schmitz and J. Thissen resulted in the rediscovery of these sediments in 1997, and subsequent excavations in 1997 and 2000 also provided additional faunal bone fragments, Neanderthal bone fragments, and Palaeolithic archaeology [34,35]. These specimens are dated to around 40 ka BP [35].

2. Methods

2.1. Protein extraction and proteolysis for MALDI-TOF MS

A total of 76 specimens (75 bone, 1 dentine) recovered from the backdirt excavation from the Kleine Feldhofer Grotte sediments [35] were analyzed using collagen fingerprinting via MALDI-TOF MS analysis (“ZooMS”) [36]. All bone specimens could either not be identified morphologically or were classified as potential hominins. Bone specimens maximum dimensions were in the range of 15 to 61 mm. Protein extraction proceeded in a two-step approach. First, soluble collagen was extracted using an ammonium-bicarbonate (Ambic; 50 mM) protocol without demineralizing the bone matrix [27]. Bone samples that failed to provide a conclusive taxonomic assignment (n = 15) or that were identified as potential hominins (n = 17) using the ammonium-bicarbonate protocol were subsequently demineralized in hydrochloric acid (0.6 M; [37]). In both protocols, solubilized proteins were digested overnight using trypsin (Thermo Scientific), acidified using an aqueous 5% v/v trifluoroacetic acid (TFA) solution, and purified using C18 ZipTips (Thermo Scientific) according to the manufacturer’s instruction. Samples were eluted using an aqueous solution comprising 0.5% v/v TFA and 50% v/v acetonitrile. One negative extraction control was processed alongside each extraction protocol to monitor the presence of protein contamination in the laboratory environment. COL1 peptides were not observed in these negative extraction controls. We note that peptide mass fingerprinting using MALDI-TOF MS is entirely based around trypsin digestion. In our study we maintained this approach and decided to not explore the use of alternative proteases for peptide mass fingerprinting purposes.

2.2. MALDI-TOF MS and spectra identification

1 μl of the eluted peptide mixture was spotted on a 384 stainless steel MALDI Target Plate together with 1 μl of CHCA matrix. Each sample was spotted in triplicate. The plate was air-dried and analyzed on a Autoflex LRF MALDI-TOF MS (Bruker) in reflector mode, positive polarity, matrix suppression of 590Da, and collected in the mass-to-charge range 800–4000 m/z. After spectra acquisition, the triplicates of each sample were merged and analyzed manually in mMass [38]. Spectral taxonomic identification proceeded via comparison against reference peptide markers that cover most medium- and larger mammal species in existence of the Late Pleistocene of western Europe [36,37]. Glutamine deamidation was calculated following Wilson et al. [39] for the peptides P1105 and P1706. Both peptides contain a single glutamine (Q) and no asparagines (N).
An equid (Equid3, ZooMS specimen number F-35) and a hominin specimen (ZooMS specimen number F-15) were selected randomly from the identified specimens. Here, the only criterion for selection was the presence of a MALDI-TOF MS spectrum allowing taxonomic identification. We added two Equidae specimens identified through peptide mass fingerprinting screening ongoing at La Ferrassie (Equid2, ZooMS specimen number LF-59) and Quinçay (Equid1, ZooMS specimen number Q-12497, [27]). The protein extraction, MALDI-TOF MS procedure, and spectral identifications were identical for these two specimens and the two Kleine Feldhofer Grotte specimens.

2.3. Protein extraction and proteolysis for LC-MS/MS

Protein extraction was based on existing protocols for palaeoproteomics [40], modified to ensure compatibility between the six different proteases. For each of the four selected bone specimens, around 60 mg of bone was weighed out and placed in a 1.5 ml microcentrifuge tube. To reduce contamination, 300 μl of 50 mM Ambic (pH 8) was added and incubated at 25 °C for 2.5 h, with 20 s of mixing every 20 min. A negative extraction control was generated for each proteolysis condition. Following this, the supernatant was removed, and replaced with 1800 μl of 0.5 M EDTA. The samples and negative controls were then rotated end-over-end until demineralisation was complete (1–3 days). They were then incubated for 30 min at 60 °C to ensure that the proteins were solubilized.

The samples were centrifuged at 10,000 x g for 10 min, and the EDTA was removed and split between four Amicon Ultra 3 kDa cut-off filters (Sigma-Aldrich, Denmark). The filters were prepared with a pre-rinse of 0.1 M NaOH suspended in molecular biology grade water (Accugene), with a subsequent rinse of molecular biology grade water. The sample was passed through the filters by centrifugation at 12,000 x g for 20–30 min. A buffer exchange was performed by washing each filter with 400 μl of 50 mM Ambic (pH 8) three times. After buffer exchange the upper fraction was eluted from each of the four filters for each sample and was then combined into a single collection tube for each sample. This was performed by sequentially removing and inverting each of the four Amicon filters and placing them in the collection tube followed by centrifugation at 1000 x g for 3 min. The same procedure was followed for the negative controls. The protein concentration was determined by BCA assay (Pierce Biotechnology, Rockford, IL, USA) according to the manufacturer's instructions.

Pellets left over from demineralization were resuspended in 100 μl of 50 mM Tris and incubated at 60 °C for 30 min, with 10 s mixing using a vortex every 10 min. Once the incubation was completed, the protein concentration of the supernatant was determined by BCA assay according to the manufacturer's instructions. Results from this analysis showed comparable protein concentrations; therefore the filtered EDTA and Tris fractions were combined.

A 0.5 M stock solution of TCEP (Sigma Aldrich) was added to each sample to reach a final concentration of 10 mM and incubated at 56 °C for 30 min. Subsequently, a 0.5 M stock solution of TCEP (Sigma Aldrich) was added to reach a final concentration of 20 mM, and incubated in the dark at room temperature for another 30 min. The supernatant of each sample and the extraction controls were then aliquoted into six equal fractions (ranging from 23 μl to 83 μl) to correspond with the six protease conditions. The divided samples were then heated to 60 °C for 30 min to promote solubilisation of the proteins.

All proteases were added in a protease:protein ratio of 1:50, and proteolysis was performed at 37 °C overnight, except for chymotrypsin which was incubated at 25 °C. Zinc sulphate (ZnSO₄) was added to Lys-N samples to a final concentration of 10 mM. Elastase, chymotrypsin, and pepsin were vacuum concentrated (Eppendorf, Denmark) to dryness and resuspended in 50 mM Tris-HCl at a pH of 8.5–9.5; 100 mM Tris-HCl and 10 mM calcium-chloride (CaCl₂) at pH 8.0; and 50 mM of Ambic (NH₄HCO₃) adjusted to a pH of one using HCl, respectively. All other samples and proteases were suspended in Ambic. A summary of extraction and proteolysis conditions can be found in Table S1. All samples were then immobilised on in-house made C18 StageTips in triplicate according to the parameters from [41], and sent for LC-MS/MS analysis at the Novo Nordisk Center for Protein Research.

2.4. LC-MS/MS

Samples were eluted from StageTips using 30 μl 40% ACN 0.1% formic acid (FA) into a 96 well MS plate. Samples were adjusted to contain approximately 2 μg protein, based on concentrations achieved by Nanodrop (Thermo, Wilmington, DE, USA) measurement at A205 nm, and all samples were brought to the same volume using 40% ACN 0.1% FA. Samples were then placed in a vacuum centrifuge at 40 °C until approximately 3 μl of solution was left and then rehydrated with 10 μl of 0.1% TFA and 5% ACN solution.

5 μl of each sample was then separated on an in-house made 15 cm column (75 μm inner diameter) packed with 1.9 μm C18 beads (Dr. Maisch, Germany) on an EASY-nLC 1200 (Proxeon, Odense, Denmark). The column was connected to a Q-Exactive HF-X orbitrap mass spectrometer (Thermo Scientific, Bremen, Germany) on a 77 min gradient. Buffer A was milliQ water, and the peptides were separated with increasing buffer B (80% ACN and 0.1% formic acid), going from 5% to 30% in 50 min, 30% to 45% in 10 min, 45% to 80% in 2 min, held at 80% for 5 min before dropping back down to 5% in 5 min and held for 5 min. Flow rate was 250 nl/min. The column temperature was kept at 40 °C with an integrated oven.

The Q-Exactive HF-X was operated in data dependent top 10 mode. Spray voltage was 3.7 kV, S-lens RF level was 40, and capillary was heated at 320 °C. Full scan mass spectra were recorded at a resolution of 120,000 at m/z 200 over the m/z range 350–1400 with a target value of 3e6 and a maximum injection time of 25 ms. MS/MS HCD-generated product ions were recorded with a maximum ion injection time set to 108 ms and a target value set to 2e5 and recorded at a resolution of 60,000. Normalized collision energy was set at 28% and the isolation window was 1.2 m/z with the dynamic exclusion set to 20 s.

A wash-blank method using 0.1% TFA and 5% ACN solution was run in between each sample to reduce cross-contamination.

2.5. MaxQuant database searching

Raw mass spectrometry data were searched in MaxQuant (version 1.6.3.4) against the horse (UP000002281) and human (UP000005460) proteomes (downloaded from UniProt on 07/07/2018 and 11/09/2018, respectively). Sequences of common protein contaminants were added to the database search in MaxQuant. For the Equidae, searches were performed separately for each protease. For the hominin, a single search was performed including all data files, grouped per protease. Infection replicates were kept separate during MQ data analysis. Each search was performed twice, once using semi-specific proteolysis mode and once using specific proteolysis mode. For each search, default settings of 20 ppm for the first search and 4.5 ppm for the final search were used, and a fragment mass tolerance of 20 ppm. Protease-specific cleavage settings can be found in Table S1. As ancient proteomes are commonly highly deamidated, we modified Glu-C specificity to also include C-terminal cleavage after glutamine (Q), alongside C-terminal cleavage of glutamic acid (E) and aspartic acid (D). This strategy worked, as approximately one quarter of the Glu-C-specific peptides derived from C-terminal cleavage of glutamines (Fig. S1; [42]). As variable post-translational modifications (PTMs), we selected oxidation (M), deamidation (N and Q), proline hydroxylation (P), and carbamidomethylation (C). No fixed modifications were selected. Peptide matches were accepted with a minimum score of 40. Peptide and protein FDR were both set to 1.0%.

Subsequently, during data analysis post-MQ protein groups were accepted as confidently identified for each sample only when at least
two replicates contained one unique + razor peptide, or when at least one replicate contained two or more unique + razor peptides. The UniProt accession numbers of protein groups considered possible contaminants and those present with at least two unique + razor peptides across all negative controls can be found in Table S2. These, alongside reverse protein entries, were removed from subsequent analysis. Sample-level deamidation rates were calculated for all remaining protein groups together using 1000 bootstraps [43]. Presented Venn diagrams were created using http://bioinformatics.psb.ugent.be/webtools/Venn/. The raw proteomic data files as well as MaxQuant search results are available in ProteomeXchange under accession number PXD018264 [44].

Fig. 2. New hominin bone specimens from the Kleine Feldhofer Grotte identified through MALDI-TOF MS. a, F-2. b, F-3. c, F-4. d, F-6. e + f, F-25. g, F-15. This bone was selected for subsequent LC-MS/MS analysis. h, F-26. i, F-34. j, F-38. k, F-40. l, F-42. m, F-51. n, F-53. o, F-54. p, F-55. q, F-63. r, F-72. s, F-76. Photos by J. Vogel (LVR-LandesMuseum Bonn).
3. Results and discussion

3.1. Collagen fingerprinting identifies novel hominin bone specimens

Collagen fingerprinting of 75 bone specimens and 1 dental specimen from the Kleine Feldhofer Grotte sediments identified 18 novel hominin bone specimens (Fig. 2; Table S3). The newly identified hominin bone specimens do not directly refit to any of the two Neanderthal individuals known from the Kleine Feldhofer Grotte hominin assemblage [35], but might represent parts of the missing skeletal elements from the Neanderthal type specimen or of the additional female Neanderthal.

In addition, specimens belonging to 13 additional taxa were also identified. These include extinct megafauna such as Elephantidae (likely Mammuthus sp.) and Rhinocerotidae, and a wider variety of medium- and large-sized herbivore specimens. A few specimens could be identified as carnivores. Extinct megafauna and the identified hominin specimens have slightly more advanced deamidation signatures compared to some, but not all, herbivores (Fig. S2) suggesting that the analyzed assemblage is potentially composed of more recent and Late Pleistocene bone specimens.

Extraction controls produced MALDI-TOF MS spectra devoid of collagen peptides, indicating that contamination by collagen did not occur in the laboratory environment. In addition, ammonium-bicarbonate and acid extractions from the hominin bone specimens produced identical taxonomic identifications. These results add to a growing number of studies that demonstrate the utility of proteomic screening of potential hominin and morphologically unidentifiable specimens in Pleistocene contexts [6,37,45].

3.2. Proteolysis with alternative proteases allows access to larger bone-proteomes

A homogenized bone sample from each specimen was digested using six proteases separately. Each digest was injected in triplicate (technical replicates). Peptide and proteome identification was then performed twice on the resulting datasets, first via semi-specific protease settings and a second time using specific protease settings. Overall proteome deamidation indicates greater deamidation for asparagine (N) compared to glutamine (Q; Fig. S3, S4), and above that observed for our negative extraction controls (Fig. S5). Multiple t-testing for a difference in deamidation for each extraction control-sample pair supports this conclusion (Table S4). Deamidation is therefore, given differences in protein and peptide numbers and proteome composition, relatively comparable across proteases for each bone specimen and suggests the recovery of endogenous, ancient bone proteomes from all four bone specimens.

We observe a significant difference (p < .000001) in extractome sizes between specific protease settings and semi-specific protease settings (means of 111.3 ± 66.5 proteins and 19.3 ± 7.4 proteins, respectively; Fig. 3; Table S5). For semi-specific searches, trypsin proteolysis resulted in larger extractomes than any other protease tested (mean of 33 proteins for trypsin and 17 proteins for the other five proteases). In contrast, chymotrypsin (mean = 178 proteins) and Glu-C (mean = 177 proteins) yielded extractomes larger than with trypsin proteolysis (mean = 131 proteins) in specific searches. Regardless of identification approach, proteolysis with elastase (mean = 51 proteins), pepsin (mean = 114 proteins), and particularly Lys-N
(mean = 16 proteins) resulted in smaller extractomes. The majority of
the proteins identified in the specific search are NCPs, while in the
semi-specific searches approximately 50% of the identified proteins are
collagens. This was observed in all conditions and samples.

Despite the smaller number of identified proteins, semi-specific
searches resulted in the identification of some proteins not observed in
specific searches. For the Equidae samples, these include decorin
(DCN), which is involved in collagen fibril formation [46], and lacto-
ylglutathione lyase (GLO1), which is necessary for normal osteo-
clastogenesis [58]. For the hominin specimen, a similar example is
represented by integrin alpha-9 (ITGA9), which together with integrin
beta-1 (ITGB1) interacts with osteopontin (SPP1; [47]). In turn, os-
teopontin is a common bone NCP that, like other small integrin-binding
ligand, N-linked glycoproteins (SIBLINGs), binds to hydroxyapatite
[48]. Therefore, despite their smaller size, semi-specific searches might
provide protein identifications of interest that could be missed during
protease-specific search settings. Additionally, protein and peptide-
identifications only recovered via semi-specific searches have the po-
tential to enlarge the phylogenetic information recovered from ancient
proteomes.

Alternative proteases also provide access to distinct extractomes
(Fig. 4). We observe that the majority of proteins are uniquely
identified by a single protease. This observation is true for all four bone
specimens that we analyzed. When comparing the protein composition
of our Equidae specimens for each of the utilized proteases, we also
observe that the majority of the proteins identified in each sample are
shared between the three specimens. Therefore, protein composition
identified using each protease is generally reproducible across bone
specimens (Fig. 5). Although the majority of proteins are represented by
low peptide counts, they are consistently identified by specific pro-
teases across Pleistocene bone specimens.

Strikingly, the trypsin extractome provided access to only a small
portion of the entire proteome per specimen (22.1%, 21.2%, 21.3%,
and 34.6%, respectively; Fig. 4). Examples of proteins not identified in
the Homo sp. trypsin digests include fibroblast growth factor-binding
protein 1 (FGFBP1, identified via Lys-N only) and von Willebrand factor
A domain-containing protein 1 (VWA1, identified via Glu-C and Lys-N).
FGFBP1 is an extracellular protein that controls the release of fibroblast
binding factors (FGFs), which play a role in skeletal growth and
maintenance [49,50]. VWA1 is expressed in chondrocytes and poten-
tially interacts with collagens, perlecan, as well as FGFs [51,52].

Our results imply that the majority of the ancient bone proteome is
potentially inaccessible to trypsin proteolysis. This observation is in
general agreement with previous modern proteomic studies of bone
non-bone tissues [21–24], that also demonstrate the recovery of a significant number of additional proteins through the use of non-trypsin proteases. Future work could potentially provide further clarification as to what determines the accessibility of the ancient bone proteome to trypsin proteolysis.  

3.3. Alternative proteases yield increased protein sequence coverage

In addition to proteome size and composition, the second determinant of the phylogenetic utility of ancient bone proteomes concerns protein sequence coverage. We therefore investigated protein sequence coverage for a subset of proteins consistently identified across bone specimens and proteases. We observe that protein sequence coverage of non-collagenous proteins (NCPs) identified across specimens and proteases is variable. In some cases the coverage retrieved from the alternative proteases is greater than the sequence coverage observed after trypsin proteolysis of the same bone specimen (Fig. 6c–f). More importantly, we observe examples where protease digests provide access to alternative sequence regions of the same protein. For example, alternative proteases almost double the sequence coverage of collagen type 5, alpha-2 (COL5A2, from 16 to 29%), biglycan (BGN, from 7 to 13%), and pigment epithelium-derived factor (SERPINF1, from 6 to 12%) in our *Homo* sp. sample (Fig. 7b, e, f). A similar observation can be made for proteins not identified by trypsin but present in several other digests of this sample. For example, ficolin-1 (FCN1), an extracellular lectin expressed in bone marrow [55], is absent in trypsin digests of our *Homo* sp. sample but has up to 20% sequence coverage across chymotrypsin, pepsin, and Glu-C extractomes (Fig. 7d). Another example concerns osteocalcin (OCN). OCN is thought to be one of the more abundant NCPs in bone and dentine proteomes but is variably detected in (trypsin-digested) ancient bone proteomes [56]. We identify no OCN peptides in our trypsin-digest either, but do retrieve peptide sequences spanning the C-terminal region of the mature protein in chymotrypsin, pepsin, and Glu-C digests (Fig. 7c). Again, this shows that (several) alternative proteases can demonstrate the presence of proteins not identified by trypsin proteolysis of the same protein extract.

Previous proteomic mass spectrometry studies have suggested that highly dominant proteins might mask the presence of low-abundance proteins [57]. Similarly, in skeletal proteomes, the dominance of COL1 might mask the presence of low-abundance proteins [54]. Indeed, we observe that COL1 has the highest sequence coverage of all identified proteins across all search conditions and bone specimens (Fig. 6a, b). As expected, sequence coverage is restricted to the mature, triple-helical regions of COL1A1 and COL1A2 (Fig. 7a). Previous attempts to deplete ancient collagen by collagenase proteolysis into peptide fragments too small for common nanoLC-MS/MS analysis proved unsuccessful in removing this dominant protein [11]. An alternative approach would be to avoid proteotypic cleavage of COL1. We observe that proteolysis with alternative proteases accesses the same mature, triple-helical region (Fig. 7a) but with lower sequence coverage (Fig. 6a, b). This is particularly true for chymotrypsin, which on average reduced COL1 sequence coverage by ≈50% (54.4% mature COL1 sequence coverage, compared to 94.2% mature COL1 sequence coverage for trypsin). This is unsurprising, as chymotrypsin primarily cleaves after large and aromatic amino acid residues (Y, W, F, M, L) which are relatively uncommon in COL1. Chymotrypsin proteomes are among the largest generated in our comparative set-up (Fig. 3). Enzymatic strategies that

![Fig. 5. Proteome overlap between Equidae specimens. a, Chymotrypsin. b, Elastase. c, Glu-C. d, Lys-N. e, Pepsin. f, Trypsin. The majority of proteins identified by each protease are identified in two or three of the Equidae specimens in all cases, with the exception of Lys-N. Data from specific search only.](image-url)
avoid or minimize the proteolytic cleavage of COL1 might therefore be a favourable route towards enhancing the detection of non-collagenous proteins in skeletal proteomics.

4. Conclusion

We identified 18 novel hominin bone specimens from the Kleine Feldhofer Grotte, potentially belonging to the Neanderthal type specimen, via MALDI-TOF MS peptide mass fingerprinting. We utilized one of these hominin bone specimens and three additional Pleistocene Equidae bone specimens and explored extractome size and composition as well as protein sequence coverage generated through the use of six proteases and nanoLC-MS/MS. We observed differences in size and composition for each of these extractomes. Importantly, we determined that the proteolysis of skeletal proteomes using only trypsin resulted in a systematic underestimation of bone proteome size. In our study, the proteases that complimented protein identification by trypsin most effectively were Glu-C and chymotrypsin. Additionally, combining sequence coverage generated via alternative proteases improved overall protein sequence coverage. Glu-C and chymotrypsin were found to be the most useful additional proteases for this application. For specific applications or targeted studies other alternative proteases might be better suited. These observations are in agreement with previously reported modern proteome studies on bone and non-bone tissues, and pave the way towards more complex ancient proteomes with improved phylogenetic value.

Funding sources

This research was supported by generous funding from the Leakey Foundation. FW was supported by the European Commission through a Marie Skłodowska-Curie (MSCA) Individual Fellowship (795569). E.C. is supported by the VILLUM FONDEN (grant number 17649). M.M., L.T.L., and A.J.T. are supported by the Danish National Research...
Foundation award PROTEIOS (DNRF128). F.W. and J.-J.H. thank the Max Planck Society for support. Work at the Novo Nordisk Foundation Center for Protein Research is funded in part by a donation from the Novo Nordisk Foundation (NNF14CC0001). Funders had no role in study design or any other aspect of the presented research.

Author contributions

F.W. and A.J.T designed the study. S.F., J.-J.H. and R.W.S. provided samples. A.W., M.J.C., E.C. and J.V.O. provided technical infrastructure. L.T.L. and A.J.T. performed the experiments. M.M. operated the nanoLC-MS/MS. F.W., L.T.L. and A.J.T. analyzed the data. F.W., L.T.L., A.J.T. and M.M. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Declaration of Competing Interest

The authors declare there is no conflict of interest.

Acknowledgements

M. Soressi, M. Roussel (Leiden University, Leiden, NL), S.P. McPherron (MPI-EVA, Leipzig, DE), T.E. Steele (UC Davis, USA), A. Turq (Musée national de Préhistoire, Les Eyzies, France), and S.


