Research Article

Rapid Communications in Mass Spectrometry

Received: 11 June 2013 Revised: 31 August 2013 Accepted: 5 September 2013 Published online in Wiley Online Library

Rapid Commun. Mass Spectrom. **2013**, *27*, 2685–2698
(wileyonlinelibrary.com) DOI: 10.1002/rcm.6730

Characterisation of novel α-keratin peptide markers for species identification in keratinous tissues using mass spectrometry

Caroline Solazzo[1,2,3]*, Marc Wadsley[1], Jolon M. Dyer[2,4,5], Stefan Clerens[2], Matthew J. Collins[1] and Jeffrey Plowman[2]

[1]BioArCh, Biology (S Block), Wentworth Way, University of York, York YO10 5DD, UK
[2]Proteins and Biomaterials, AgResearch Lincoln Research Centre, Private Bag 4749, Christchurch 8140, New Zealand
[3]Smithsonian's Museum Conservation Institute, 4210 Silver Hill Road, Suitland, MD 20746, USA
[4]Biomolecular Interaction Centre, School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
[5]Riddet Institute, Massey University, PB 11 222, Palmerston North 4442, New Zealand

RATIONALE: In ancient and/or damaged artefacts containing keratinous materials, the species of origin of the materials can be difficult to identify through visual examination; therefore, a minimally destructive methodology for species identification is required. While hair fibres from some species have seen substantial characterisation, others such as horn or baleen have received little or no attention, or lack protein sequences allowing formal identification using proteomics techniques.
METHODS: We used the PMF method (Peptide Mass Fingerprinting with matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS)) to catalogue and identify diagnostic peptide markers up to the genus level. Sequences were checked using nanoflow liquid chromatography/electrospray ionisation tandem mass spectrometry (nanoLC/ESI-MS/MS) and unidentified peptides were searched against a theoretical database generated by substituting amino acids in keratin sequences.
RESULTS: Specific peptides were identified by *m/z* and sequences characterised whenever possible for a range of species belonging to Bovidae and Camelidae, and for tissues such as baleen and horn. The theoretical database allowed an increase in the number of peptides of up to 10% in species with little genetic information.
CONCLUSIONS: A proteomics approach can successfully identify specific markers for the identification of materials to the genus level, and should be considered when identification by other means is not possible. Identification by PMF is fast, reliable and inexpensive. Copyright © 2013 John Wiley & Sons, Ltd.

Tissues made of keratins from mammals, birds and reptiles form an important class of archaeological and cultural artefacts. Wool, hair, horn, hoof, nail, baleen, claws and quills are all made of α-keratins (α-helical structure), while β-keratins (β-pleated sheet structure) are found in the hard keratinous tissues (scutes, scales, beaks and feathers) of sauropsids.[1] While analyses by Fourier transform infrared (FT-IR) and XRF are capable of discriminating α- from β-keratins in hard tissues (horn, tortoiseshell, etc.),[2] more specific identifications are traditionally obtained using microscopy.[3] For wool and hair fibres, species identification in ancient textiles is based upon the species-specific pattern of the overlapping scales forming the cuticle[3] and the diameter of fibres.[4,5] This is limited by variations between breeds, sometimes even in the fleece itself,[4] and by the degradation of the cuticle, especially in archaeological artefacts. For, horn, baleen and tortoiseshell the lack of structure on extensively worked or degraded materials can make positive identification difficult.[6,7] Animal fibres have recently been identified using specific peptide markers from mass spectra acquired by MALDI-TOF-MS (matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry); peptides are obtained after trypsin digestion of proteins and identified by their mass-to-charge ratio (*m/z*). The technique was successful in identifying the different elements of the clothing of Oetzi,[8,9] and to prove the existence of dog hair in Coast Salish blankets.[10] In both cases, the textiles proved to be made of a wide range of species and the proteomics approach was particularly useful in characterising blends of fibres (dog/goat or sheep/goat) in the Coast Salish textiles. Species identification of keratinous materials has important applications for cultural heritage (identification of textiles, horn, baleen and tortoiseshell artefacts), modern textile production (to detect frauds in textiles made of luxury fibres such as cashmere[11,12]), and the illegal trade of endangered species[2,13–15] (for instance, rhinoceros horn powder[16]).

We focus here on the identification of α-keratins (β-keratin-made artefacts will be described in a forthcoming publication), in particular in wool-producing animals, and in

* *Correspondence to:* C. Solazzo, Smithsonian's Museum Conservation Institute, 4210 Silver Hill Road, Suitland, MD 20746, USA.
E-mail: solazzo.c@gmail.com

the hard tissues baleen, porcupine quill, horn and hoof. Baleen was historically in great demand from the 17th to the 19th centuries mainly to make corsets and is also found in archaeological artefacts from Alaska and Japan.[6,17] Quill was a common material in the paraphernalia of Native American textiles and is used nowadays in Africa to make artefacts such as jewellery and lampshades.[18] Horn and hoof will have been used alongside bone for the fabrication of cultural objects (i.e. vessels, combs, drinking horns and helmets[19]) wherever horn-bearing species were scavenged or hunted. They can be preserved in metal corrosion products and thus become a more visible element of the archaeological record from the Bronze Age onwards. Rhinoceros horn was a prized material for centuries, in particular in Asia where it was used to make cups, bowls and other decorative carvings.[20] It continues to be sought today for putative beneficial properties claimed in traditional Chinese medicine (TCM), and illegal trade has increased in order to satisfy the important demand in Asia, to the point that horn artefacts have recently been the targets of thefts in museums.[21]

BACKGROUND

Structure of keratinous tissues

The cortex of mammalian hair is organised into cortical cells (paracortex, orthocortex, mesocortex) in which intermediate filaments assembled from α-keratins are arranged into different configurations and are responsible for the crimp and coarseness of fibres. In horn, hoof and baleen, intermediate filaments are organised with different orientations in tubules and intertubular material. A baleen plate grows from dermal papillae and ends with bristles (the role of which is to filter food from water). In cattle, horn grows from the epithelium layer of a bony core, an outgrowth located on the frontal bone of the skull, while rhinoceros horn is entirely a keratin structure and lacks the bony core. The difference between horn and hoof and between species is found in the density and diameters of the tubules, and the orientation of the intermediate filaments in tubules and intertubular material. Horn and hoof are organised in lamellar structures while in baleen an additional layer covers the tubular and intertubular core, forming a thin plate.[22–24] Cylindrical porcupine quills on the other hand are made of an outer shell with an inner foam core.[25]

The α-keratin families

The α-keratins are the main proteins found in intermediate filaments of epithelial cells. The keratins or intermediate filament proteins assemble into parallel dimers of one type I (acidic) protein and one type II (neutral to basic) protein. In humans, the keratin family contains 54 genes, expressing specifically in certain cell types and tissues;[26] type I K9-20, K23-24, and type II K1-8, K76-80 keratins are found in single epithelial cells of inner organs and in tissues such as the epidermis (stratified epithelia); type I K25-28, and type II K71-75 are expressed in the hair follicle (root sheath); and type I K31-40, and type II K81-87 (trichocyte or 'hard' α-keratins) are expressed in hair and hard tissues such as fingernails.[1] The hard α-keratins are assembled in microfibrils themselves embedded in a matrix of keratin-associated proteins (KAPs), a large group of proteins that contains up to 27 families of high-sulfur, ultra-high-sulfur and high glycine-tyrosine proteins, with most protein having multiple isoforms. Trichocyte keratins are different from the epithelial keratins in that they have a higher content of the sulfur amino acid, cysteine, in the head and tail domains forming abundant cross-links (disulfide bridges) between the keratins and with the KAPs. In wool this contributes to the stability of the fibre.[27] In horn and hoof, however, and although the protein composition is the same, the relative proportion of proteins is different; horn and hoof have a higher content of low-sulfur α-keratins (or microfibrils) but a lower content of matrix proteins and in particular of the high-sulfur KAPs.[28]

Recently, keratin genes and proteins in sheep have been sequenced[29,30] and the nomenclature for sheep wool keratins adjusted to the human model. It was found that K37 (a protein found only in the cortex of vellus hair[26]) was absent but a new protein, K87, was characterised. Protein expression varies with the structures of the wool fibre: in sheep K40, 82 and 84 are expressed in the cuticle and K31, 33a, 33b, 34, 36, 38, 39, 81, 83, 86, 87 in the cortex only.[30] The remaining proteins (K32, 35 and 85) are expressed in both cortex and cuticle. The KAPs are also abundantly found in the cuticle; however, these proteins are less reliable than keratins for species identification by proteomics. They contain fewer basic residues and therefore are not as well digested with trypsin as the keratins. They are less abundant, and less susceptible to identification by mass spectrometric analysis alone without prior chromatographic separation. They have not been characterised for many species, and it is yet to be established how conserved they are between species. Finally, identifying them in archaeological and degraded samples is tedious and unpredictable.[31] Therefore, we rely here on the diversity of the more abundant keratins to establish specific markers for species identification.

Protein databases

At the time of writing, α-keratins have been predicted through translation of genomic sequences for about 50 species including human, cow, pig, horse, dog, cat, rabbit, mouse, rat, and African elephant. *Ovis aries* was sequenced in 2011[30] and sequences compiled in the AgResearch in-house protein database, while more recently a number of sequences for yak, camel and White rhinoceros have been made available (see Supporting Information S1 for the complete list of species and the accession numbers of available keratins).

The PMF method and its limitations

Methodologies using peptide mass fingerprinting (PMF) for species identification[32] are based on the variations between species of the amino acid residue sequences of the proteins of interest. Keratins are a complex family of proteins: each type has several proteins (some with a high level of sequence homology) and consequently several variations of a peptide are possible, with some specific to one species only, some to more than one. With multiple variations of a peptide possible, the ones that belong to the most abundant

Rapid Communications in Mass Spectrometry

(or most expressed) proteins (typically type I K31, 33a, 33b and 34 and type II K81, 83, 85 and 86) are usually identified by PMFs. In contrast, a peptide that is present in one of the minor proteins for a given species or across a range of species might not be observed. Furthermore, in archaeological samples peptide profiles may vary due to protein damage (for instance, in cases of differential degradation: cuticle vs. cortex), resulting in peaks appearing or disappearing (the relative intensity of peaks is also affected); hence there is an advantage in knowing the complete pool of proteins for a species of interest.

In PMFs, peptides are identified by their *m/z* value; therefore, identification is possible only for those species that have had a spectrometric profile created for them previously and is problematic when identifying an unknown species. Sequence information can be used in addition to the peptide mass fingerprint using database searching algorithms such as Mascot[33] which match calibrated observed masses to known sequences with predicted masses. Using this method, peaks in a peptide mass fingerprint can be assigned to a sequence and a likely species match rapidly identified. Sequence database searching methods can also take into account post-translational modifications (PTMs) as well as limited sequence variation. It is therefore advantageous to be able to match a peak to its exact native primary sequence in order to also be able to predict the *m/z* values for likely PTMs of the peptide (for instance, deamidation that can shift the mass of a peptide marker of a few Da as observed in archaeological fibres[34]).

To be effective, the sequence libraries that form the Mascot databases must be populated with accurate sequences from a wide range of species. Here we identify specific markers for mammalian α-keratins in a range of species and extend the characterisation of new markers to hard keratinous tissues using a database of theoretical keratin sequences.

EXPERIMENTAL

Materials

Hair and wool from sheep, cashmere and mohair goat, cow, yak, camel, alpaca, angora rabbit, and dog were obtained locally from fibre suppliers and farms. Llama, vicuna, guanaco, alpaca, Bactrian camel, yak, horse, deer and reindeer were sampled from the fibre collection at the Museum Conservation Institute and the mountain goat from the Smithsonian's National Museum of Natural History (NMNH).[10] Mouflon (origin Poland), and baleen were available at AgResearch. Muskox was provided by the Anglo-Saxon Laboratory (York). Ibex and chamois hair were provided by the Salzburg Zoo, Austria. Markhor, Tahr and Urial sheep were provided by the Highland Wildlife Park, Kingussie, Scotland. Harbour seal and porcupine quill (origin South Africa) were available at the University of York, and the horn and hoof samples at the University of Bradford.

Chemicals

Urea, tris(2-carboxyethyl)phosphine (TCEP), iodoacetic acid (IAA) and acetone were provided by Sigma-Aldrich (USA), ammonium bicarbonate (AB) by BDH AnalaR (UK), acetonitrile (ACN), water (H_2O), trifluoroacetic acid (TFA) by Fisher Scientific (USA), formic acid by Ajax Finechem Pty Ltd (Univar analytical reagents), Thermo Fisher Scientific (NZ) and α-cyano-4-hydroxycinnamic acid (CHCA) by Bruker Daltonics. The 3500 MWCO Slide A Lyzer® mini dialysis units were obtained from Thermo Scientific (Rockford, USA).

Sample preparation

Samples (10 mg) were ground in liquid nitrogen and allowed to dry overnight. The samples were solubilised by overnight shaking in a 1 mL solution of 8 M urea, 50 mM Tris and 50 mM TCEP at pH 8.3. An aliquot of the supernatant was alkylated with 150 mM IAA and vortexed for 4 h in the dark. This was followed by 24 h dialysis with 100 mM AB on 3500 molecular weight cut-off (MWCO) dialysis units (two changes). An aliquot (about 25 µg of sample) was digested with 0.5 µg of trypsin, overnight at 37 °C. Samples were then dried down and re-solubilised in 10 µL of 0.1% TFA.

Peptide mass fingerprinting by MALDI-TOF-MS

A matrix solution was prepared by diluting 0.1 mg of CHCA in 97:3 acetone/0.1% TFA and 1 µL was applied onto an AnchorChip™ target (Bruker) and allowed to dry. A 1 µL aliquot of analytical solution was applied and removed after 1 min followed by 1 µL of washing buffer (0.1% TFA). The residual droplet was removed and 1 µL of recrystallisation solution (0.1 mg of CHCA in 6:3:1 EtOH/acetone/0.1% TFA) applied. The plate was loaded in a Ultraflex™ III mass spectrometer (Bruker), and analyses were carried out in positive reflector mode using a Nd:YAG laser operating at 355 nm. Spectra were acquired using flexControl 3.0 (Bruker) on a mass range of 700–4000 Da with an accumulation of 500 shots on the standards and 1000 shots on the samples. The calibration standard (Bruker) was prepared according to the manufacturer's instructions for instrument calibration and consisted of angiotensin I, ACTH clip(1-17), ACTH clip (18-39) and ACTH clip(7-38) peptides. The spectra were processed with mMass 5.4.1.0[35] after conversion of the raw files with flexanalysis 3.3 (Bruker). Peaks were accepted with a mass accuracy within 0.1 Da.

Peptide analysis by nanoLC/ESI-MS/MS

Peptide separation was carried out on an Ultimate nanoflow nanoLC equipped with a Famos autosampler and a Switchos column switching module (LC Packings, The Netherlands). A 10 µL sample was loaded on a C18 precolumn (Varian Microsorb 300 µm i.d., 5 µm particles, 300 Å pore size) at a flow rate of 8 µL/min. The precolumn was then switched in line with the analytical column (Microsorb C18, 20 cm, 75 µm i.d., 5 µm particles, 300 Å pore size), and eluted at a flow rate of 150 nL/min, with a gradient from 2% to 55% B in 50 min. Solvent A was HPLC-grade H_2O (Fisher Scientific, USA) with 0.2% formic acid, solvent B was LCMS-grade ACN with 0.2% formic acid. Using a stainless steel nanospray needle (Proxeon, Denmark), the column outlet was directly connected to a Q-STAR Pulsar i mass spectrometer (Applied Biosystems, USA) which was programmed to acquire tandem mass spectrometry (MS/MS) traces of 1+, 2+, 3+, 4+ and 5+ peptides. Mascot Daemon (Matrix Science, UK) was

used to extract peak lists from the data files. The peak lists from all *m/z* segments of each sample were concatenated and imported in ProteinScape v2.1 (Bruker).

Analyses of the reference samples of sheep horn, cattle horn and horse hoof were performed on a ultrahigh-resolution quadruple time-of-flight (UHR Q-TOF) instrument. HPLC was performed using a nanoAcquity UPLC system (Waters) equipped with a nanoAcquity Symmetry C18, 5 µm trap (180 µm × 20 mm; Waters) and a nanoAcquity BEH130 1.7 µm C18 capillary column (75 m × 250 mm; Waters). The trap wash solvent was 0.1% aqueous formic acid and the trapping flow rate was 10 µL/min. The trap was washed for 5 min after sample loading before switching flow to the capillary column. The separation used a gradient elution of two solvents: solvent A (0.1% formic acid) and solvent B (ACN containing 0.1% formic acid). The flow rate for the capillary column was 300 nL/min, column temperature was 60 °C and the gradient profile was as follows: initial conditions 5% solvent B (2 min), followed by a linear gradient to 35% solvent B over 20 min and then a wash with 95% solvent B for 2.5 min. The column was returned to initial conditions and re-equilibrated for 25 min before subsequent injections. The nanoLC system was interfaced to a maXis UHR Q-TOF mass spectrometer (Bruker) with a nano-electrospray source fitted with a steel emitter needle (180 µm o.d. × 30 µm i.d.; Thermo). Positive-ion MS and CID (collision-induced dissociation) MS/MS spectra were acquired using AutoMSMS mode. Instrument control and data acquisition were performed using Compass 1.3 SP1 software (micrOTOF control and Hystar, Bruker). Instrument settings were: ion spray voltage, 1400 V; dry gas, 4 L/min; drying gas temperature, 160 °C; and ion acquisition range *m/z*, 50–2200. AutoMSMS settings were for MS 0.5 s (acquisition of survey spectrum) and for MS/MS (CID with N_2 as collision gas) ion acquisition range *m/z*, 350–1400; 0.1 s acquisition for precursor intensities above 100,000 counts; for signals of lower intensities down to 1000 counts acquisition time increased linearly to 1.5 s; the collision energy and isolation width settings were automatically calculated using the AutoMSMS fragmentation table; three precursor ions; absolute threshold 1000 counts; preferred charge states 2–4; singly charged ions excluded. Two MS/MS spectra were acquired for each precursor and previously selected target ions were excluded for 60 s. Spectra were calibrated using a lock mass signal (*m/z* 1221.99064) prior to compound detection and peak list creation using DataAnalysis (Bruker Daltonics). The peak list files obtained were submitted for database searching to a locally-running copy of Mascot (Matrix Science Ltd., version 2.3.02).

Bioinformatics analysis

Mascot (Matrix Science) was used to search for matches against the database of publicly available sequences NCBI (National Center for Biotechnology Information, US), timestamp 24th July 2013. Searches were carried out using semi-trypsin as enzyme, two missed cleavages, peptide mass tolerance (MS) of 100/150 ppm (Q-star) or 20 ppm (maXis), and fragment mass tolerance (MS/MS) of 0.4 Da (Q-star) or 0.1 Da (maXis), carboxymethylation as a fixed modification (unless otherwise indicated) and acetyl (N-term), carbamyl (N-term), deamidated (NQ), Gln- > pyro-Glu (N-term Q), Glu- > pyro-Glu (N-term E) and oxidation (M) as variable modifications. Proteins were accepted with at least two unique peptides and peptide's cutoff score was set at 30 (Table S3, see Supporting Information). The same parameters were used for searching the house database using Mascot, with the exception that no missed cleavage was allowed and trypsin was used as enzyme.

RESULTS AND DISCUSSION

Creation of a theoretical database

Until more sequences are made publically available, spectrometric libraries or *de novo* sequencing are the only methods available to distinguish keratinous material from unsequenced species. Due to their importance as structural proteins, variation in the sequences of extracellular matrix proteins, such as collagen and keratin, is heavily constrained. An iterative approach for *de novo* sequencing of these proteins is feasible, though still computationally intensive with search space increased 191-fold and 16,452-fold for one and two substitutions, respectively.[36] Iterative approaches also require an initial match, which is difficult when the number of available sequences is limited as in the case of keratin. A theoretical database containing all permutations of sequences with observed variation was created, increasing the size of the database and therefore the chance of an initial match but informed by variation known to be possible. Search space is increased, resulting in speed and false positives increasing; however, given limited variations this was not a significant issue.

Sequences for the keratin proteins K31-40 and K81-86 were obtained from public databases (from *Homo sapiens*, *Ovis aries*, *Bos taurus*, *Mus musculus*, *Rattus norvegicus*, *Equus caballus*, *Canis lupus*, *Oryctolagus cuniculus* and *Sus scrofa*). Each group of keratin proteins was aligned using ClustalX[37] (alignments are shown in Supporting Information). Highly variable or missing sequences at the termini of the alignments were deleted up to the first and from the last arginine or lysine. These sequences contained either a missing sequence or were too long to be in the detection range used for the mass spectrometer so their exclusion would not cause fragments to be missed in the subsequent Mascot searches. A custom python script was used to create the theoretical keratin database. For each alignment the positions of all lysines and arginines in each sequence were mapped. The sequences were digested with trypsin *in silico* with each tryptic fragment stored with its start and end positions from the original alignment. All tryptic fragments were then collated with the start and end position as a key so that all tryptic fragments from the same section of the alignment were stored together. Fragments of fewer than five amino acids in length were discarded, as the masses of these peptides would fall outside the lower detection range limit. The variations between tryptic fragments from the same location in the alignment were discerned by a sequential search to create a set containing every observed variation at each position. The built-in itertools.product function in python was used to create every Cartesian product of the set of variations for each tryptic fragment. These fragments were concatenated to form a long sequence chain containing every permutation of

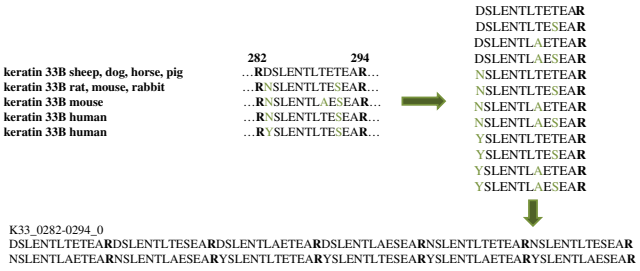

Figure 1. Example of a long sequence created after *in silico* digestion of a K33 sheep protein. Chosen permutations are applied on a peptide and a single sequence is created by concatenating the generated fragments.

observed variation for a given tryptic fragment. Within the chain there is no redundancy; fragments conserved in multiple species are only represented once to limit the inflation of the search space. Mascot limits on length of protein sequences meant that over length chains had to be broken down into blocks. For use in Mascot these concatenated tryptic fragment chains were given a header containing the protein of origin, the start position and end position in the original alignment and a block number (e.g. >K33_0282-0294_0, see Fig. 1 and Supporting Information).

Wool-producing animals

Most peaks on a PMF can be identified to known peptides for those species that have had their keratins sequenced. Figure 2 shows an example of a PMF for domestic sheep; peptides were confirmed by MS/MS analysis in separate nanoLC/ESI-MS/MS analysis (data not shown). Peaks were mainly identified as peptides derived from keratins with few peaks tentatively identified as coming from KAPs by comparison with MS/MS data. These peaks serve as references to identify homologous peptides in unknown species. The peptide markers most useful for species identification in wool-producing animals were identified in a range of species and are given Table 1 with their sequence and calculated *m/z* (the presence of the markers in other species available in public database is given in S1, see Supporting Information).

A typical MS profile shows an abundance of peaks in the *m/z* 700–1600 region, with a few peaks appearing in the *m/z* 2000–3000 area. One peptide usually predominates all PMFs for the species studied: type II LGLDIEIATYR at *m/z* 1263.69. Consistently present in the analysis of placental mammalian fibres, it is also found in all species available in the public database (S1, see Supporting Information) except for the gray short-tailed opossum, where the sequence is LGLDIEIATY**K**. That this sequence is also present in the Tasmanian devil suggests its association with marsupials. In most species' type II family of proteins, LGLDIEIATYR has no other homologue, which may account for it being the most intense peak. Two homologous peptides (L**A**LDIEIATYR and LGLD**V**EIATYR) are found in sheep and a few more species, but are not observed in PMFs. Other common markers found in the wool-producing species analysed here are peptides QNQEYQVLLDVR, *m/z* 1504.77, LNVEVDAAPTVDLNR, *m/z* 1625.85, and DSLENTLTETEAR, *m/z* 1478.70, which are found in all except harbour seal (Table 1).

Separation at the family level

The most common marker characterising even-toed ungulates (Cetartiodactyla) is TVNALE**V**ELQAQHNLR (*m/z* 1834.98),[9] found in all Bovidae, Cervidae and Camelidae (as well as pig and platypus). Its homologue is TVNALE**I**ELQAQHNLR, *m/z* 1848.99, observed in dog, horse and rabbit (other species, see S1, Supporting Information). The elephant has both markers, as well as harbour seal (highest intensity for *m/z* 1848.99).

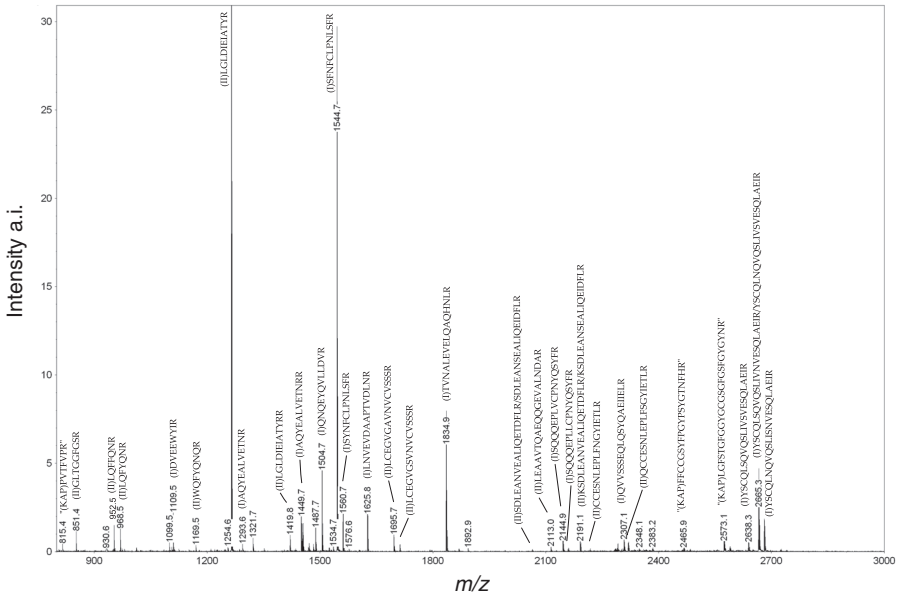

Figure 2. Peptide mass fingerprint of domestic sheep (Icelandic breed). (I) and (II) indicate type I and type II, respectively, KAP is for keratin-associated protein. Sequences indicated in brackets were tentatively identified after comparison with nanoLC/MS/MS data.

Table 1. Main markers for species identification observed by PMFs

	Family	Bovidae												Cervidae		Camelidae		Equidae	Leporidae	Canidae	Phocidae		Hystricidae	
	Sub-family	Caprinae										Bovinae		Cervinae	Odocoileinae									
		Urial (*Ovis vignei*)	Mouflon (*Ovis aries musimon*)	Domestic sheep (*Ovis aries*)	Ibex (*Capra ibex*)	Markhor (*Capra falconeri*)	Domestic goat (*Capra hircus*)	Mountain goat (*Oreamnos americanus*)[d]	Chamois (*Rupicapra rupicapra*)	Tahr (*Hemitragus jemlahicus*)	Muskox (*Ovibos moschatus*)	Cow (*Bos Taurus*)	Yak (*Bos grunniens*)	Deer (*Cervus elaphus*)	Reindeer (*Rangifer tarandus*)	Camel (*Camelus bactrianus*)	Lama and vicugna[b]	Horse (*Equus caballus*)	Rabbit (*Oryctolagus cuniculus*)	Dog (*Canis familiaris*)	Harbour seal (*Phoca vitulina*)[d]	Baleen (unknown species)	Porcupine (*Hystrix africaeaustralis*)	Unique peptides
N	Number of specimens sampled	1	5	27[e]	1	1	5[e]	1	2	1	1	1	3	1	1	4	9[e]	1	2	1	1	1	1	
m/z[a]	(Type)Peptide																							
952.50	(II)LQFFQNR	+	+	+	+	+	+			+		+	+	*				+	+				+	
968.49	(II)LQFYQNR	+	+	+	+	+	+	+	+	+	+	+	+	*	*	+	*	+			~		+	
1041.49	(II)WQFYQNR																	+		+	*	+		
1169.55	(II)WQFYQNQR	+	+	+	+	+	+	+	+	+	+	+	+			+	*	+	+	+	*			
1051.48	(I)DVEEWFAR															+	*							
1109.53	(I)DVEEWYIR	+	+	+	+	+	+	+	+	+	+	+	+		*						*			
1081.49	(I)DVEEWFTR																					+		
1896.86	(I)DVEEWFTTQTEELNR																	+						
1235.69	(II)LGLDIEIATYK																						+	
1263.69	(II)LGLDIEIATYR	+	+	+	+	+	+	+	+	+	+	+	+	*	*	+	*	+	+	+	*	+	+	
1463.70	(I)NSLENTLTESEAR															~	*		+					
1478.70	(I) DSLENTLTETEAR	+	+	+	+	+	+	+	+	+	+	+	+	*	*	+	*	+	D	+				
1353.65 1336.60	(I)QNHEQEVNSLR + Gln->pyro-Glu (N-term Q)																	H						
1504.77 1487.70	(I)QNQEYQVLLDVR + Gln->pyro-Glu (N-term Q)	+	+	+	+	+	+	+	+	+	+	+	+	*	*	+	*	+	+	+	*			
1544.72	(I)SFNFCLPNLSFR + acetyl N term	+	+	+	+	+	+		+	+	+													
1547.65	(I)SYSCCLPNLSFR + acetyl N term											+	+											U
1547.65	(I)SYSCFLPNLSCR + acetyl N term																	+						U
1560.71	(I)SYNFCLPNLSFR + acetyl N term	+	+	+	+	+	+	+	+	+	+	+	+	*	*									
1611.83	(I)LNVEVDAAPSVDLNR																						+	
1625.85	(I)LNVEVDAAPTVDLNR	+	+	+	+	+	+	+	+	+	+	+	+	*	*	+	*	+	+	+	~		+	
1655.82	(I)LNVEVDAAPTEDLNR[c]																					+		U
1834.98	(I)TVNALEVELQAQHNLR	+	+	+	+	+	+	+	+	+	+	+	+	*	*	+	*				*			
1848.99	(I)TVNALEIELQAQHNLR																	+	+	+	*		+	
1857.99	(I)TVHALEVELQAQHNLR[c]																					+		U
2047.03	(II)SDLEANAEALIQEIDFLR																						+	
2175.13	(II)KSDLEANAEALIQEIDFLR[c]																						+	U
2063.03	(II)SDLEANSEALIQEIDFLR/ SDLEANVEALIQETDFLR		+	+ +			+		+	+	+	D			*	D						+		
2191.12	(II)KSDLEANSEALIQEIDFLR KSDLEANVEALIQETDFLR		+	+ +				+	+		+	D			*	D								
2191.12	(II)KSDLEANVEALTQEIDFLR														*			H						
2075.06	(II)SDLEANVEALIQEIDFLR											H	+								*			
2203.16	(II)KSDLEANVEALIQEIDFLR											H	+								*			
2072.02	(II)LEAAVTQAEQQGEAALTDAR																			+				U
2074.00	(II)LEAAVTQSEQQGEAALSDAR																		+					
2086.04	(II)LEAAVSQAEQQGEVALTDAR																	H						U
2101.02	(II)LEAAVTQAEQQGEATVNDAR															+	*							U

(Continues)

Table 1. Continued

2113.05	(II)LEAAVTQAEQQGEVALNDAR	+	+	+	+	+	+		+	+	+	+							*				
2114.07	(II)LEVAVSQAEQQGEVALTDAR																H						U
2115.98	(I)SQQQEPLLCPSYQAYFR[e]																					+	U
2130.95	(I)SQQQDPLVCPNYQSYFR[c]																				+		U
2144.97	(I)SQQQEPLVCPNYQSYFR	+	+	+	+	+	+	+	+	+	+	+	+			+	*						
2234.91	(I)CQQQEPEVCPNYQSYFR																+						U
2131.99	Sequence not known												*	*									
2154.07	(I)GDLEANAHSLVEEVCVLGLK											H											U
2164.11	Sequence not known																		*				
2258.16	(I)QVVSSSEQLQSNQAEIIELR															+	*						U
2307.18	(I)QVVSSSEQLQSYQAEIIELR		+	H						+		H	+					+	+				
2517.33	(I)YSSQLAQIQGLIGNVEAQLSEIR[e]																				+		U
2519.29	(I)YSSQLAQMQGLIGNVEAQLAEIR			H								H											
2547.34	(I)YSSQLSQVQALISNVEAQLAEIR[c]																					+	U
2637.32	(I)YSSQLSQVQCLISNVEAQLAEIR[c]																					+	U
2563.34	(I)YSSQLSQVQGLITNVESQLAEIR																+						U
2577.30	(I)YSCQLAQVQGLIGNVESQLAEIR											+	+	*	*								U
2583.35	(I)YGSQLSQVQGLITNVEHQLAEIR															+	*						U
2593.27	(I)YGSQLAQMQCLIGNVEAQLAEIR																	+					U
2712.30	(I)YSSQLNQVQCMITNVESQLAEIR																	+					U
2638.34	(I)YSCQLSQVQSLIVSVESQLAEIR	+	+	+	+		+			+	+												U
2665.35	(I)YSCQLSQVQSLIVNVESQLAEIR	+	+	+	~					+	+												U
	YSCQLNQVQSLIVSVESQLAEIR																						U
2680.32	(I)YSCQLNQVQSLISNVESQLAEIR	+	+	+	+	~	+	+	+	+									*				
2692.36	(I)YSCQLNQVQSLIVNVESQLAEIR				+	+	+	+	+	+	+												U
2576.36	Sequence not known																		*				
2662.30	Sequence not known																		*				
2786.33	Sequence not known															+	*						
2796.57	Sequence not known																	+					

+ indicates a peptide observed on the PMF and sequence confirmed by MS/MS or from available sequences in public database, * indicates a peak observed on the PMF but sequence not confirmed, and ~ indicates a low intensity peak.
D stands for a peptide appearing in sequences in public database, but that was not observed, H is for a peak present in horn/hoof but not observed in hair, and U indicates sequences that are unique to one species, using BLAST (Basic Local Alignment Search Tool from NCBI;[46] see details in S1, Supporting Information).
[a]*m/z* values calculated as $[M+H]^+$ using PeptideMass from ExPASy,[47] SIB Swiss Institute of Bioinformatics (Lausanne, Switzerland), with carboxymethylation modification on cysteine (+ 58 Da), or observed *m/z* for unknown peaks.
[b]Includes Lama and Vicugna genera: llama (*Lama glama*), alpaca (*Vicugna pacos*), vicugna (*Vicugna vicugna*), guanaco (*Lama guanicoe*).
[c]Sequences from York database and *de novo* sequencing from nanoLC/ESI-MS/MS data.
[d]Species alkylated with iodoacetamide resulting in cysteines modified by carbamydomethylation (+57 Da).
[e]The sheep breeds include Merino (n = 5), Shetland (n = 5), Soay (n = 2), North Ronaldsay (n = 5), Icelandic (n = 5) and Wensleydale (n = 5); the domestic goat breeds include Cashmere (n = 3) and Mohair (n = 2); the camelids include Lama (n = 1), Guanaco (n = 1), Vicugna (n = 1) and Alpaca (n = 6).

Peptide DVEEWYIR (*m/z* 1109.53)[9] is a very common marker for Bovidae and Cervidae (also found in harbour seal at low intensity), found in K31, K33a and K33b in *Ovis aries*, K31 in *Capra hircus* and *Bos grunniens mutus*, and K31, K33a, K33b and K34 in *Bos taurus* (public database data). Although this is the main marker in sheep, other rarer variations are present, such as DVEEWFTR (*m/z* 1081.49) in K32. This variant is present in *Ovis aries* only, according to the public sequences, and is seen in low abundance in the sheep PMFs. Other tryptic peptides are DVEEWFNK in K32 *Bos taurus* (*m/z* 1066.48) and DVEEWFIK in K34 *Equus caballus* (*m/z* 1065.52). Homologues of this peptide in other species are non-tryptic (the C-terminus is not an R or a K), for instance the main peptide in horse is DVEEWFTTQTEELNR at *m/z* 1896.86. Both *m/z* 1065.52 and 1896.86 are seen on the horse PMF. The homologue for Camelidae is *m/z* 1051.48 DVEEWFAR (from *Camelus ferus* K33b, and also *Sus scrofa* K31 and K34).

Peptide SQQQEPLVCPNYQSYFR at *m/z* 2144.97 is an intense peak in modern fibres but tends to break down quickly in ancient fibres; it is present in all Bovidae and Camelidae species (and pig), but not in the Cervidae. One peak is observed in this area at *m/z* 2131.99 for reindeer, but has not been sequenced yet. Peptide LEAAVTQAEQQGEVALNDAR, at *m/z* 2113.05, is a marker for Bovidae (also present in elephant). Its homologous peptide in Camelidae (*Camelus ferus* K81 sequence LEAAVTQAEQQGEATVNDAR), at *m/z* 2101.01, is present in all Camelidae species analysed.

Separation at the genus level

Peptide LQFYQNR at *m/z* 968.49 is a marker common to most species of Table 1 except for dog and rabbit. Its homologues LQFFQNR at *m/z* 952.50 is present in rabbit, deer (but not reindeer) and in Caprinae (Table 1) except for chamois,

mountain goat and muskox, thereby indicating that the absence of this marker could be used to distinguish these species from other Caprinae species. The absence of the peptide was confirmed by MS/MS analysis (data not shown). The homologous peptide WQFYQNR, *m/z* 1041.49 is one of the most intense peaks in horse, dog and harbour seal (both dog and seal are from the suborder Caniformia). Finally, the peak at *m/z* 1169.55 attributed to WQFYQNQR is observed in all species, except for the Cervidae.

Peptides SFNFCLPNLSFR (*m/z* 1544.72 + acetyl N-term) and SYNFCLPNLSFR (*m/z* 1560.71 + acetyl N-term) are common peptides visible on PMFs, both belong to Caprinae (*m/z* 1544.72 was not detected in *Oreamnos americanus*). In Cervidae, only *m/z* 1560.71 is observed, while in Bovinae *m/z* 1560.71 is observed along with SYSCCLPNLSFR (*m/z* 1547.65 + acetyl N-term). Peptide *m/z* 1547.65 is also present in horse as SYSCFLPNLSCR.

The most diagnostic markers however are the peptides corresponding to the sheep K33b peptide YSCQLSQVQSLIVN-VESQLAEIR (*m/z* 2665.35). As mentioned in previous publications,[9,10] this peptide is highly variable, both in the same genus and among different genera (Table S2 in Supporting Information highlights the most variable residues and the most common substitutions, it also illustrates that

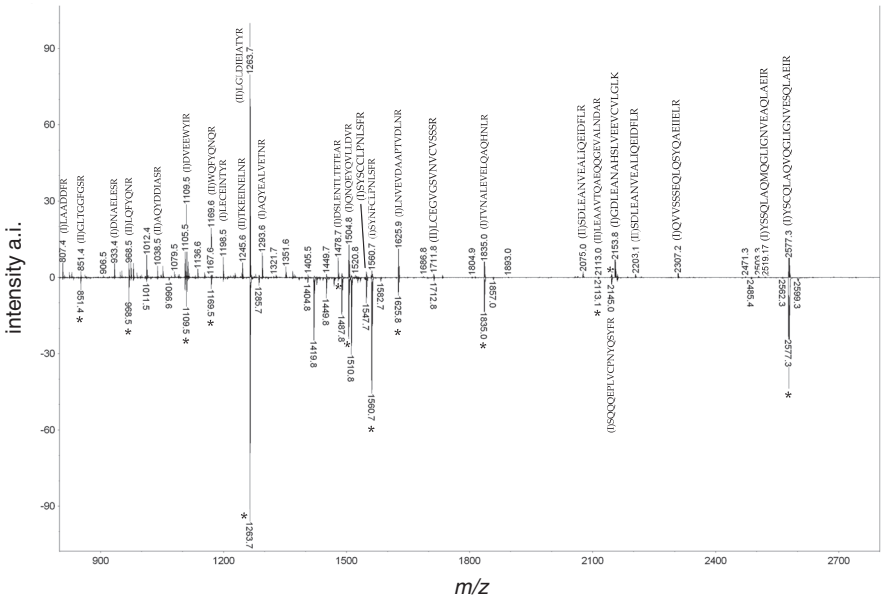

Figure 3. Peptide mass fingerprints of cow horn compared to cow hair (reversed spectra), the stars refer to the sequences indicated in the top panel.

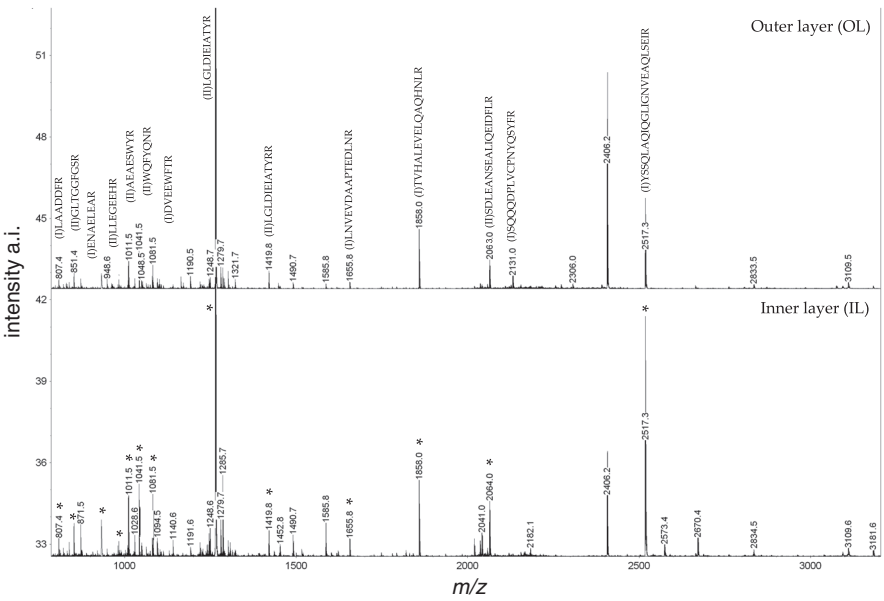

Figure 4. Peptide mass fingerprints of baleen (unknown species): outer layer (top) and inner layer (bottom), the stars refer to the sequences indicated in the top panel.

the glutamine, leucine and isoleucine residues are the most constant residues in the sequence). The sheep peptide was found to have a homologue in domestic goat's K33a at *m/z* 2692.36 (YSCQLNQVQSLIVNVESQLAEIR). The homologous main peak for Camelidae is YGSQLSQVQGLITNVEHQLAEIR (from *Camelus ferus* K33b), at *m/z* 2583.35 (Fig. 6(a)), the main horse marker is YSSQLSQVQGLITNVESQLAEIR (*Equus caballus* K31, K33a and K34) at *m/z* 2563.34, and the main bovine marker is YSCQLAQVQGLIGNVESQLAEIR (*Bos taurus* K31, K33a and K34) at *m/z* 2577.30; a *m/z* 2577.30 peak was also identified in reindeer. Harbour seal is characterised by a peak at *m/z* 2662 (after carbamidomethylation of cysteine).

As has been previously reported with collagen,[38] differentiating species beyond the genus level is more problematic due to recent evolutionary divergence times

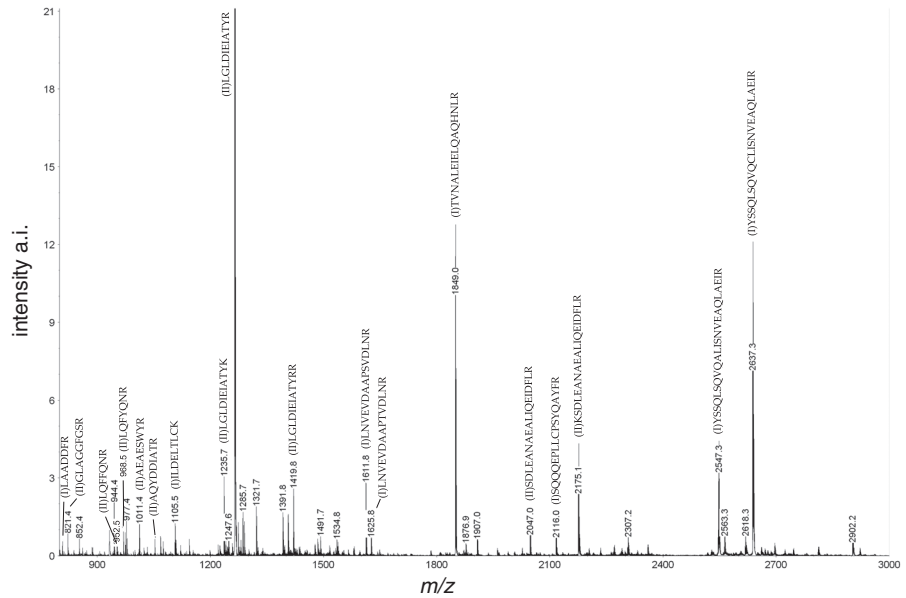

Figure 5. Peptide mass fingerprint of porcupine (*Hystrix africaeaustralis*) quill.

Table 2. New peptides with Mascot score identified for baleen, porcupine and white rhinoceros using computer-generated keratin sequences and *de novo* sequencing from nanoLC/ESI-MS/MS data. $[M+H]^+$ calculated using PeptideMass from ExPASy,[47] SIB Swiss Institute of Bioinformatics (Lausanne, Switzerland), with cysteines treated with iodoacetic acid (carboxymethylation)

		Baleen			White rhinoceros	
$[M+H]^+$	Sequence	IL	OL	Porcupine	filaments	shavings
969.50	ISSGYSVTR					49
995.52	YNTELSLR			59		
1065.50	GLCGTTGGVSR	72	64			
1111.56	LESEITTYR	74	71			
1655.82	LNVEVDAAPTEDLNR	88	105			
1857.99	TVHALEVELQAQHNLR	72	73			
2073.01	LEAAVTQSEQQGEAALNDAK			136		
2115.98	SQQQEPLLCPSYQAYFR			91		
2130.95	SQQQDPLVCPNYQSYFR		91			
2175.13	KSDLEANAEALIQEIDFLR			*de novo*		
2353.17	QVVSSSEQLQSYQVEILEMR				*de novo*	
2517.33	YSSQLAQIQGLIGNVEAQLSEIR	124				
1328.72	IGNVEAQLSEIR		77			
2520.33	YSSQLAQVQGLITSVESQLAEIR				139	134
2547.34	YSSQLSQVQALISNVEAQLAEIR			*de novo*		
2552.30	YSSQLAQVQGLITSMESQLAEIR				*de novo*	
2637.32	YSSQLSQVQCLISNVEAQLAEIR			*de novo*		
Total new peptides		6		6	4	

between species, and we found only two peptides that showed variations in the Caprini clade. *Capra*, *Ovis* and *Hemitragus* (belonging to the Caprina subclade of Caprini[39] that diverged between 5 and 8.9 Mya[40]) contains the peptide LQFFQNR at *m/z* 952.50; the peptide is absent in *Rupicapra*, *Oreamnos* and *Ovibos*. The *Capra* and *Ovis* genera can be distinguished by only one peptide (*m/z* 2692.36 for *Capra*, and *m/z* 2665.35 for *Ovis*) with the substitution of only one amino acid, and both contain a homologous peptide at *m/z* 2680.32 (YSCQLNQVQSLISNVESQLAEIR from K31 *Ovis aries* and *Capra hircus*). However, within the *Capra* (split time ~1.5 Mya) and *Ovis* (split

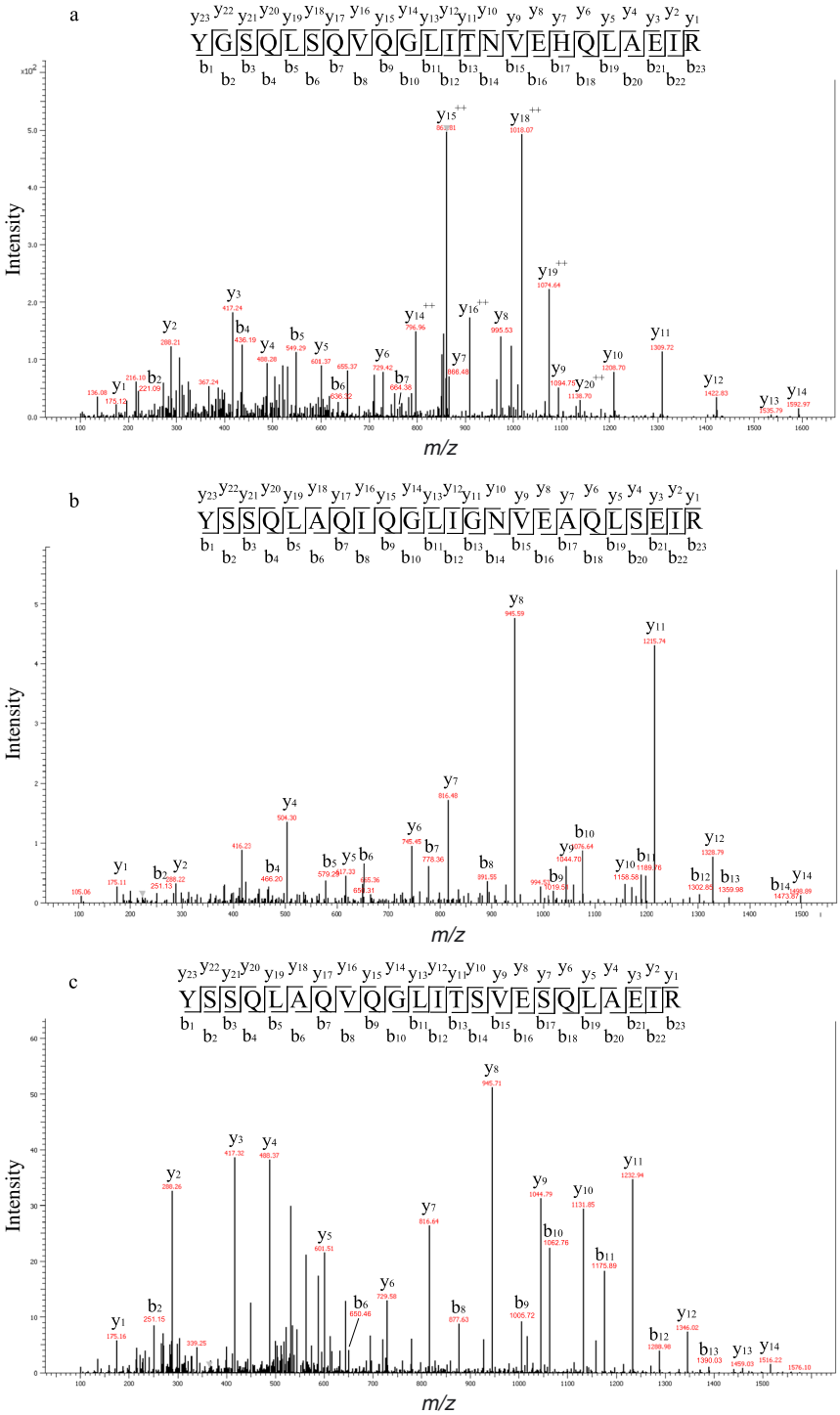

Figure 6. MS/MS spectra (ESI) of (a) YGSQLSQVQGLITNVEHQLAEIR *m/z* 861.8380 triply charged (Mr 2582.4922) from camel, (b) YSSQLAQIQGLIGN VEAQLSEIR *m/z* 839.7785 triply charged (Mr 2516.3136) from baleen, and (c) YSSQLAQVQGLITSVESQLAEIR *m/z* 840.9409 triply charged (Mr 2519.8009), from rhinoceros shavings.

time~2.2 Mya)[41] genera, we could not find unique markers to distinguish between species. As in *Capra*, the *Rupicapra* (split time ~1.6 Mya)[41] and *Oreamnos* (mountain goat) genera both have *m/z* 2680.32 and 2692.36. *Hemitragus jemlahicus* (tahr), which split from the *Capra* genus between 2.5 and 4.7 Mya,[40] has all three peptides, in contrast to *Ovibos* (muskox) which lacks *m/z* 2680.32 but contains the other two sequences. In studies of collagen, chamois, tahr and muskox had a collagen peptide specific to *Ovis*, distinct from its homologue *Capra* sequence.[42,43] The presence of *m/z* 2665.35 in *Ovis* and *Ovibos*, of *m/z* 2692.36 in *Capra* and *Rupicapra*, and of both in *Hemitragus* suggests that the peptides (together with *m/z* 2680.32) were present at the time of the split of Caprini tribe around 8.7–11.9 Mya.[40]

Hard keratinous materials

Pieces of modern horse hoof, sheep and cattle horn were ground and PMFs generated as described above. Figure 3 shows the PMFs for cattle horn compared to its woollen equivalent in cow, while sheep horn and horse hoof are shown in the Supporting Information (the main protein hits obtained by MS/MS analysis are reported in Table S3). Markers found in horn and hoof only are indicated with H in Table 1. A piece of baleen plate of unknown origin (AgResearch archives) was sampled in two places: the covering layer or outer layer (OL, yellow) and the inner layer of tubules and intertubular material (IL, black). Both layers were scraped and analysed separately. The porcupine quill came from South Africa (Cape Porcupine *Hystrix africaeaustralis*) and was entirely ground. The PMFs of the two layers of baleen and the porcupine quill are shown Figs. 4 and 5. The absence of sequence information for these materials and associated species resulted in no specific match after MS/MS analysis (Table S3, see Supporting Information) and the non-identification of many peptides (identified peptides are reported in S4, see Supporting Information). Baleen, which belongs to the Mysticeti sub-order of the even-toed ungulates (Cetartiodactyla), has several matches to various species such as *Sus scrofa*, *Bos taurus* and *Ovis aries*. In comparison porcupine matches *Rattus norvegicus* best; both species belong to the Rodentia order. Using the computer-generated database of theoretical keratin sequences and manual sequencing new peptides were identified (Table 2) and sequences reported (Table 1 and Figs. 4 and 5); this allowed an increase in the number of peptides of 7% for baleen and 10% for porcupine, reflecting the generally high conservation of keratins. The MS/MS spectrum for the new marker YSSQLAQIQGLIGNVEAQLSEIR (Mr 2516.3136) from baleen is shown Fig. 6(b), while Fig. 7 shows the

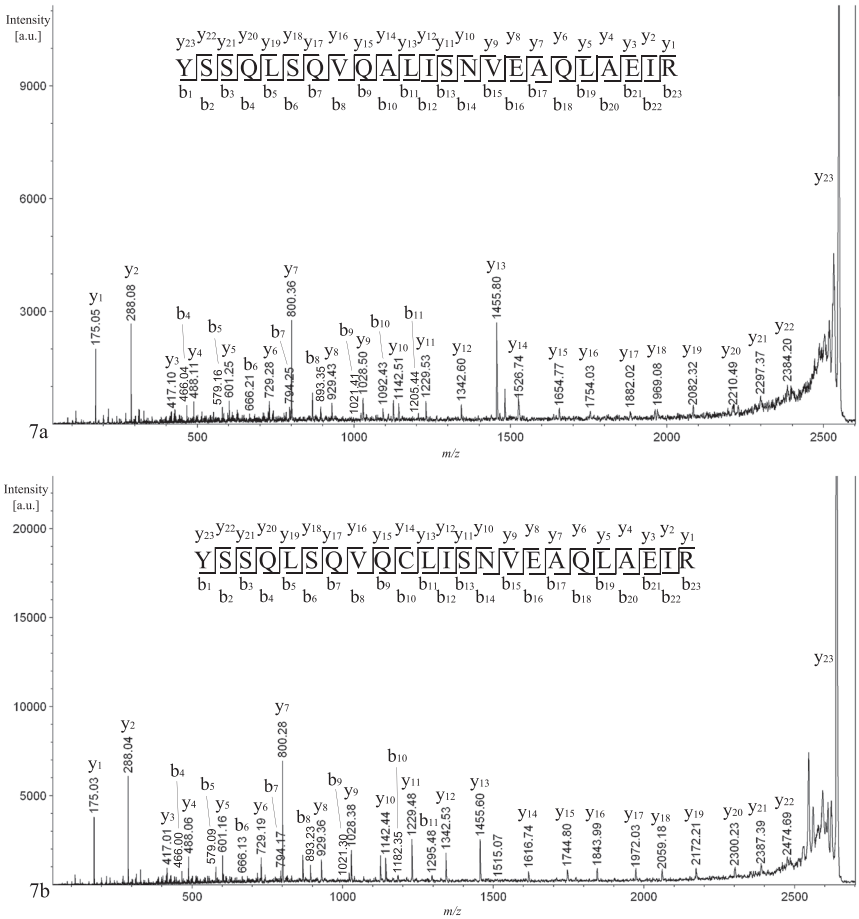

Figure 7. MS/MS spectra (MALDI) of (a) YSSQLSQVQALISNVEAQLAEIR *m/z* 2547.3854 singly charged and (b) YSSQLSQVQCLISNVEAQLAEIR *m/z* 2637.3957 singly charged (MALDI) from porcupine quill

Table 3. Homologous peptides to the markers identified Table 1 for the White rhinoceros analysed by nanoLC/ESI-MS/MS (complete set of identified peptides given in S4, see Supporting Information). Rhinoceros keratins available at the time of writing are K32 (GI:478520681), K35 (GI:478520679), K36 (GI:478520677), K84 (GI:478509697), K83 (GI:478509691), K85 (GI:478509693, GI:478509695) and K86 (GI:478509689)

$[M+H]^+$	Sequence	Accession # *Ceratotherium simum simum*	Unique peptides
968.49	(II)LQFYQNR	GI:478509689; GI:478509691; GI:478509693	
1065.54	(II)WSFLQEQK	GI:478509697	
1169.55	(II)WQFYQNQR	GI:478509695	
1896.86	(I)DVEEWFTTQTEELNR	uncharacterized protein XP_004434926	
1263.69	(II)LGLDIEIATYR	GI:478509689; GI:478509691; GI:478509693; GI:478509695; GI:478509697	
1478.70	(I)DSLENTLTETEAR	GI: 478520681	
1353.65	(I)QNHEQEVNSLR	uncharacterized protein XP_004434926	
1504.77	(I)QNQEYQVLLDVR	GI: 478520679	
1583.83	(I)LNVEVDAAPSVDLNK	GI: 478520677	U
1639.86	(I)LNIEVDAAPTVDLNR	uncharacterized protein XP_004434926	
1800.94	(I)TVNALEIELQAQQSTR	GI: 478520677	
1834.98	(I)TVNALEVELQAQHNLR	uncharacterized protein XP_004434926	
2028.98	(II)LEAAVAEAEQQGEAALSDAR	GI:478509689; GI:478509691; GI:478509693	
2307.18	(I)QVVSSSEQLQSYQAEIIELR	uncharacterized protein XP_004434926	
2353.17	(I)QVVSSSEQLQSYQVEILEMR	Theoretical database	U
2520.33	(I)YSSQLAQVQGLITSVESQLAEIR	Theoretical database	U
2552.30	(I)YSSQLAQVQGLITSMESQLAEIR	Theoretical database	U
2563.32	(I)YSSQLAQMQGLITNVEAQLAEIR	GI: 478520677	U

MS/MS spectra of YSSQLSQVQALISNVEAQLAEIR *m/z* 2547.3854 (Fig. 7(a)) and YSSQLSQVQCLISNVEAQLAEIR *m/z* 2637.3957 (Fig. 7(b)) from porcupine quill (all other peptides shown in Supporting Information).

In order to further test our theoretical database, unpublished MS/MS results on rhinoceros horn were generously provided by AgResearch. The rhinoceros horn material was obtained from a white rhinoceros (*Ceratotherium simum*) and was provided by Orana Wildlife Park, Christchurch, New Zealand. Samples of horn internal filaments and horn shavings were obtained from a horn broken off during a fight between two animals and from fence posts around the enclosure, respectively. The samples have a majority of matches for *Equus caballus* keratins (Table S3, see Supporting Information), a logical match as both horse and rhinoceros belong to the odd-toes ungulates (Perissodactyla). Table 3 shows the main markers for rhinoceros horn identified by MS/MS and verified against recently sequenced keratins from *Ceratotherium simum simum* (K32, 35, 36 and K83, 84, 85, 86). Four new markers are reported in Table 2 and all MS/MS results are reported in S4 (see Supporting Information). The MS/MS spectrum for the new marker YSSQLAQVQGLITSVESQLAEIR (Mr 2519.8009) from rhinoceros is shown in Fig. 6(c) (all other peptides are shown in the Supporting Information).

Although horn and hoof tissues are made of the same keratin sequences that the hair/wool from the same species, differences are observed on their PMFs. For instance, *m/z* 1263.69 is low in sheep horn and horse hoof (but not in cattle). Figure 3 shows peaks at *m/z* 2075.06, 2203.16 and 2154.07 in cattle horn that were not visible in hair. The peptide at *m/z* 2519.29 YSSQLAQMQGLIGNVEAQLAEIR, a peptide found in K36 of yak, cow and sheep, is present in horn of cattle and sheep while it has not been detected in fibres. K36 and K84, two proteins difficult to identify in wool due to low abundance, are here identified in all horn and hoof materials (Table S3, see Supporting Information). In horn (but not baleen and porcupine), cytoskeletal keratins are detected but few keratin-associated proteins are (in wool up to 13 cytoskeletal keratins, and 36 KAPs[29] can be found). Cytoskeletal keratins had also been detected in the horn of Saiga.[44] Separation by two-dimensional (2D) gels conducted at AgResearch on rhinoceros horn showed a lower abundance of proteins in the KAPs area compared to wool fibres (unpublished data). Although this is beyond the scope of this study, differences in expression of the different keratin proteins must play an important role in the organisation of intermediate filaments and rigidity of the hard keratinous materials.

CONCLUSIONS

We demonstrated in this study the possibility of distinguishing materials made of mammalian α-keratins to the genus level by selecting peptides that are diagnostic to a range of species important for the textile industry, modern and ancient, and for artefacts of cultural heritage. The limitations imposed by the lack of keratin sequences in the public database were overcome by searching theoretical sequences created by substitution of variable residues, thus producing a large range of possible new sequences. The best matches were manually confirmed and important species markers were characterised; one in particular exists in many variations in every genus (YSCQLSQVQSLIVNVESQLAEIR in *Ovis*, Table S2, see Supporting Information). New sequences of this peptide were characterised for unknown species, for instance in a baleen sample of unknown origin. The Mysticeti suborder that separated about 28–30 Mya[45] has 16 species belonging to six genera (*Balaena*, *Eubalaena*, *Balaenoptera*, *Megaptera*, *Eschrichtius* and *Caperea*). With a range of divergence times of the six genera from 0.8 to 23 Mya,[45] it should be possible to distinguish

between important species used in the past, for example the bowhead whale (*Balaena* genus, split time ~5.4 Mya) from the grey whale (*Eschrichtius* genus, split time ~9 Mya) based on this single peptide.

Protein recovery from archaeological keratinous samples can be challenging, especially when artefacts have spent a long period of time in the ground. Species identification is highly dependent on the condition of the sample and burial history.[31] However, we have evidence that protein markers can survive in samples up to a few thousand years.[9] For instance, we succeeded in identifying the unique *Ovis* (*m/z* 2665) and *Bos* peptides (*m/z* 2577) by PMF in greatly mineralised textiles and horn artefacts associated with copper (forthcoming publications). In cases where unique markers are degraded, however, a combination of markers can allow to narrow down the identification to a family or range of species. In addition to archaeological artefacts, the methodology can be applied to a large range of modern materials, and be used in problems of fibre adulteration, fraud and illegal trade of keratin-made objects.

Acknowledgements

This research was supported by a Marie Curie International Outgoing Fellowship (FP7-PEOPLE-IOF-GA-2009-236425, THREADs) between the University of York, UK, and AgResearch, New Zealand (2009–2012) and by a Smithsonian's Museum Conservation Institute (MCI) post-doctoral grant (2008 and 2012). We particularly thank Dr. Sonia O'Connor for reviewing this manuscript. The following people are thanked for their assistance in providing samples: Penelope Walton-Roger from The Anglo-Saxon Laboratory, York, UK; Mary Ballard from the Smithsonian's Museum Conservation Institute, USA; Sonia O'Connor from the University of Bradford, UK; Douglas M. Richardson from the Highland Wildlife Park, Kingussie, Scotland; Kathrin Mayr from the Salzburg Zoo, Austria; Denis O'Connell from AgResearch, Lincoln, New Zealand; and Lucinda Backwell from the University of Witswatersrand, Johannesburg, South Africa. Analyses on the Ultraflex III and QStar were carried out at AgResearch, New Zealand, and we are very grateful to the Proteins & Biomaterials team. We thank Dr. David Ashford at the Technology Facility, University of York, for carrying analyses on the maXis instrument. Finally, we would like to acknowledge the following persons who have participated in the White Rhinoceros study at AgResearch, supported by FRST contract C10X0710: Drs. Santanu Deb-Choudhury, Anita J. Grosvenor, Charisa D. Cornellison, Henning Koehn and Ancy Thomas.

REFERENCES

[1] H. H. Bragulla, D. G. Homberger. Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia. *J. Anatomy* **2009**, *214*, 516.
[2] E. O. Espinoza, B. W. Baker, C. A. Berry. The analysis of sea turtle and bovid keratin artefacts using drift spectroscopy and discriminant analysis. *Archaeometry* **2007**, *49*, 685.
[3] H. M. Appleyard. *Guide to the Identification of Animal Fibres*. British Textile Technology Group, Leeds, **1978**.
[4] M. L. Ryder. A survey of European primitive breeds of sheep. *Genet. Sel. Evol.* **1981**, *13*, 381
[5] M. L. Ryder. The history of sheep breeds in Britain. *Agric. Hist. Rev.* **1964**, *12*, 1.
[6] J. A. Lauffenburger. Baleen in museum collections: its sources, uses, and identification. *J. Am. Inst. Conserv.* **1993**, *32*, 213.
[7] S. O'Connor, in *Archaeological Bone, Antler and Ivory, (Occasional Paper no. 5)*, (Eds: K. Starling, D. Watkinson), United Kingdom Institute for Conservation, London, **1987**, p. 9.
[8] K. Hollemeyer, W. Altmeyer, E. Heinzle, C. Pitra. Species identification of Oetzi's clothing with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry based on peptide pattern similarities of hair digests. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 2751.
[9] K. Hollemeyer, W. Altmeyer, E. Heinzle, C. Pitra. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry combined with multidimensional scaling, binary hierarchical cluster tree and selected diagnostic masses improves species identification of Neolithic keratin sequences from furs of the Tyrolean Iceman Oetzi. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 1735.
[10] C. Solazzo, S. Heald, M. W. Ballard, D. A. Ashford, P. T. DePriest, R. J. Koestler, M. J. Collins. Proteomics and Coast Salish blankets: a tale of shaggy dogs? *Antiquity* **2011**, *85*, 1418.
[11] K. Hollemeyer, W. Altmeyer, E. Heinzle. Identification and quantification of feathers, down, and hair of avian and mammalian origin using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Anal. Chem.* **2002**, *74*, 5960.
[12] K. Kerkhoff, G. Cescutti, L. Kruse, J. Müssig. Development of a DNA-analytical method for the identification of animal hair fibers in textiles. *Text. Res. J.* **2009**, *79*, 69.
[13] K. Hollemeyer, W. Altmeyer, E. Heinzle. Identification of furs of domestic dog, racoon dog, rabbit and domestic cat by hair analysis using MALDI TOF mass spectrometry. *Spectrosc. Eur.* **2007**, *19*, 8.
[14] E. Espinoza, B. Baker, T. Moores, D. Voin. Forensic identification of elephant and giraffe hair artifacts using HATR FTIR spectroscopy and discriminant analysis. *Endangered Species Res.* **2008**, *9*, 239.
[15] B. C. Yates, E. O. Espinoza, B. W. Baker. Forensic species identification of elephant (Elephantidae) and giraffe (Giraffidae) tail hair using light microscopy. *Forensic Sci. Med. Pathol.* **2010**, *6*, 165.
[16] K. Nowell. Assessment of rhino horn as a traditional medicine, in *Species Trade and Conservation: Rhinoceroses*, 62nd Meeting of the Standing Committee (SC62), CITES, Geneva, SC62 Doc. 47.2, Annex (Rev.2), **2012**, pp. 1–43.
[17] M. Rijkelijkhuizen. Whales, walruses, and elephants: Artisans in ivory, baleen, and other skeletal materials in seventeenth- and eighteenth-century Amsterdam. *Int. J. Hist. Archaeol.* **2009**, *13*, 409.
[18] N. Chevallier, B. Ashton. A report on the porcupine quill trade in South Africa. IFAW (International Fund for Animal Welfare), **2006**, p. 1.
[19] A. MacGregor. *Bone, Antler, Ivory and Horn: The Technology of Skeletal Materials Since the Roman Period*. Croom Helm, London, **1985**.
[20] J. Chapman. *The Art of Rhinoceros Horn Carving in China*. Christie's Books, London, **1999**.
[21] T. Milliken, J. Shaw. *The South Africa - Viet Nam Rhino Horn Trade Nexus: A deadly combination of institutional lapses, corrupt wildlife industry professionals and Asian crime syndicates*. TRAFFIC, Johannesburg, South Africa, **2012**.

[22] D. S. Fudge, L. J. Szewciw, A. N. Schwalb. Morphology and development of blue whale baleen: An annotated translation of Tycho Tullberg's classic 1883 paper. *Aquatic Mammals* **2009**, *35*, 226.
[23] L. J. Szewciw, D. G. de Kerckhove, G. W. Grime, D. S. Fudge. Calcification provides mechanical reinforcement to whale baleen α-keratin. *Proc. Roy. Soc. B: Biol. Sci.* **2010**, *277*, 2597.
[24] L. Tombolato, E. E. Novitskaya, P.-Y. Chen, F. A. Sheppard, J. McKittrick. Microstructure, elastic properties and deformation mechanisms of horn keratin. *Acta Biomater.* **2010**, *6*, 319.
[25] S. F. Chou, R. A. Overfelt. Tensile deformation and failure of North American porcupine quills. *Mater. Sci. Eng. C* **2011**, *31*, 1729.
[26] R. Moll, M. Divo, L. Langbein. The human keratins: biology and pathology. *Histochem. Cell Biol.* **2008**, *129*, 705.
[27] P. Strnad, V. Usachov, C. Debes, F. Gräter, D. A. D. Parry, M. B. Omary. Unique amino acid signatures that are evolutionarily conserved distinguish simple-type, epidermal and hair keratins. *J. Cell Sci.* **2011**, *124*, 4221.
[28] R. C Marshall, J. Gillespie. The keratin proteins of wool, horn and hoof from sheep. *Aust. J. Biol. Sci.* **1977**, *30*, 389.
[29] S. Clerens, C. D. Cornellison, S. Deb-Choudhury, A. Thomas, J. E. Plowman, J. M. Dyer. Developing the wool proteome. *J. Proteomics* **2010**, *73*, 1722.
[30] Z. Yu, J. E. Wildermoth, O. A. M. Wallace, S. W. Gordon, N. J. Maqbool, P. H. Maclean, A. J. Nixon, A. J. Pearson. Annotation of sheep keratin intermediate filament genes and their patterns of expression. *Exp. Dermatol.* **2011**, *20*, 582.
[31] C. Solazzo, J. M. Dyer, S. Clerens, J. E. Plowman, E. E. Peacock, M. J. Collins. Proteomic evaluation of the biodegradation of wool fabrics in experimental burials. *Int. Biodeter. Biodegr.* **2013**, *80*, 48.
[32] W. J. Henzel, C. Watanabe, J. T. Stults. Protein identification: the origins of peptide mass fingerprinting. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 931.
[33] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551.
[34] D. P. Kirby, M. Buckley, E. Promise, S. A. Trauger, T. R. Holdcraft. Identification of collagen-based materials in cultural heritage. *Analyst* **2013**, *138*, 4849.
[35] http://www.mmass.org.
[36] B. Y. Renard, B. Xu, M. Kirchner, F. Zickmann, D. Winter, S. Korten, N. W. Brattig, A. Tzur, F. A. Hamprecht, H. Steen. Overcoming species boundaries in peptide identification with Bayesian Information Criterion-driven Error-tolerant Peptide Search (BICEPS). *Mol. Cell. Proteomics* **2012**, *11*, M111.014167.
[37] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, D. G. Higgins. The CLUSTAL_X Windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **1997**, *25*, 4876.
[38] M. Buckley, M. J. Collins. Collagen survival and its use for species identification in Holocene-lower Pleistocene bone fragments from British archaeological and paleontological sites. *Antiqua* **2011**, *1*, 1.
[39] F. Bibi, E. Vrba, F. Fack. A new African fossil caprin and a combined molecular and morphological bayesian phylogenetic analysis of caprini (Mammalia: Bovidae). *J. Evolut. Biol.* **2012**, *25*, 1843.
[40] A. Ropiquet, A. Hassanin. Molecular evidence for the polyphyly of the genus Hemitragus (Mammalia, Bovidae). *Mol. Phylogenet. Evolut.* **2005**, *36*, 154.
[41] C. Lalueza-Fox, J. Castresana, L. Sampietro, T. Marques-Bonet, J. A. Alcover, J. Bertranpetit. Molecular dating of caprines using ancient DNA sequences of *Myotragus balearicus*, an extinct endemic Balearic mammal. *BMC Evol. Biol.* **2005**, *5*, 70.
[42] M. Buckley, M. Collins, J. Thomas-Oates, J. C. Wilson. Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 3843.
[43] M. Buckley, S. Whitcher Kansa, S. Howard, S. Campbell, J. Thomas-Oates, M. Collins. Distinguishing between archaeological sheep and goat bones using a single collagen peptide. *J. Archaeol. Sci.* **2010**, *37*, 13.
[44] K. Mikulíková, O. Romanov, I. Miksik, A. Eckhardt, S. Pataridis, P. Sedláková. Study of saiga horn using high-performance liquid chromatography with mass spectrometry. *The Scientific World Journal* **2012**, 8.
[45] M. R. McGowen, M. Spaulding, J. Gatesy. Divergence date estimation and a comprehensive molecular tree of extant cetaceans. *Mol. Phylogenet. Evolut.* **2009**, *53*, 891.
[46] http://blast.ncbi.nlm.nih.gov/Blast.cgi.
[47] http://web.expasy.org/peptide_mass/.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

