ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/274048229

Ancient proteins resolve the evolutionary history of Darwin's South American ungulates

Article *in* Nature · March 2015

DOI: 10.1038/nature14249

CITATIONS
43

READS
1,081

31 authors, including:

Matthew James Collins
The University of York
403 PUBLICATIONS **8,344** CITATIONS

SEE PROFILE

Javier N. Gelfo
Facultad de Ciencias Naturales y Museo. Uni…
54 PUBLICATIONS **614** CITATIONS

SEE PROFILE

Roman Fischer
University of Oxford
109 PUBLICATIONS **1,016** CITATIONS

SEE PROFILE

Jesper V Olsen
University of Copenhagen
262 PUBLICATIONS **22,784** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project

Paleo-geographic patterns v/s climate change in South America and the Antarctic Peninsula during the latest Cretaceous: a possible explanation for the origin of the Austral biota? View project

Project

Atomistic modelling of CVD synthesis of carbon nanotubes and graphene View project

All content following this page was uploaded by Ross Macphee on 25 March 2015.

The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.

doi:10.1038/nature14249

Ancient proteins resolve the evolutionary history of Darwin's South American ungulates

Frido Welker[1,2], Matthew J. Collins[1], Jessica A. Thomas[1], Marc Wadsley[1], Selina Brace[3], Enrico Cappellini[4], Samuel T. Turvey[5], Marcelo Reguero[6], Javier N. Gelfo[6], Alejandro Kramarz[7], Joachim Burger[8], Jane Thomas-Oates[9], David A. Ashford[10], Peter D. Ashton[10], Keri Rowsell[1], Duncan M. Porter[11], Benedikt Kessler[12], Roman Fischer[12], Carsten Baessmann[13], Stephanie Kaspar[13], Jesper V. Olsen[14], Patrick Kiley[15], James A. Elliott[15], Christian D. Kelstrup[14], Victoria Mullin[16], Michael Hofreiter[1,17], Eske Willerslev[4], Jean-Jacques Hublin[2], Ludovic Orlando[4], Ian Barnes[3] & Ross D. E. MacPhee[18]

No large group of recently extinct placental mammals remains as evolutionarily cryptic as the approximately 280 genera grouped as 'South American native ungulates'. To Charles Darwin[1,2], who first collected their remains, they included perhaps the 'strangest animal[s] ever discovered'. Today, much like 180 years ago, it is no clearer whether they had one origin or several, arose before or after the Cretaceous/Palaeogene transition 66.2 million years ago[3], or are more likely to belong with the elephants and sirenians of superorder Afrotheria than with the euungulates (cattle, horses, and allies) of superorder Laurasiatheria[4–6]. Morphology-based analyses have proved unconvincing because convergences are pervasive among unrelated ungulate-like placentals. Approaches using ancient DNA have also been unsuccessful, probably because of rapid DNA degradation in semitropical and temperate deposits. Here we apply proteomic analysis to screen bone samples of the Late Quaternary South American native ungulate taxa *Toxodon* (Notoungulata) and *Macrauchenia* (Litopterna) for phylogenetically informative protein sequences. For each ungulate, we obtain approximately 90% direct sequence coverage of type I collagen α1- and α2-chains, representing approximately 900 of 1,140 amino-acid residues for each subunit. A phylogeny is estimated from an alignment of these fossil sequences with collagen (I) gene transcripts from available mammalian genomes or mass spectrometrically derived sequence data obtained for this study. The resulting consensus tree agrees well with recent higher-level mammalian phylogenies[7–9]. *Toxodon* and *Macrauchenia* form a monophyletic group whose sister taxon is not Afrotheria or any of its constituent clades as recently claimed[5,6], but instead crown Perissodactyla (horses, tapirs, and rhinoceroses). These results are consistent with the origin of at least some South American native ungulates[4,6] from 'condylarths', a paraphyletic assembly of archaic placentals. With ongoing improvements in instrumentation and analytical procedures, proteomics may produce a revolution in systematics such as that achieved by genomics, but with the possibility of reaching much further back in time.

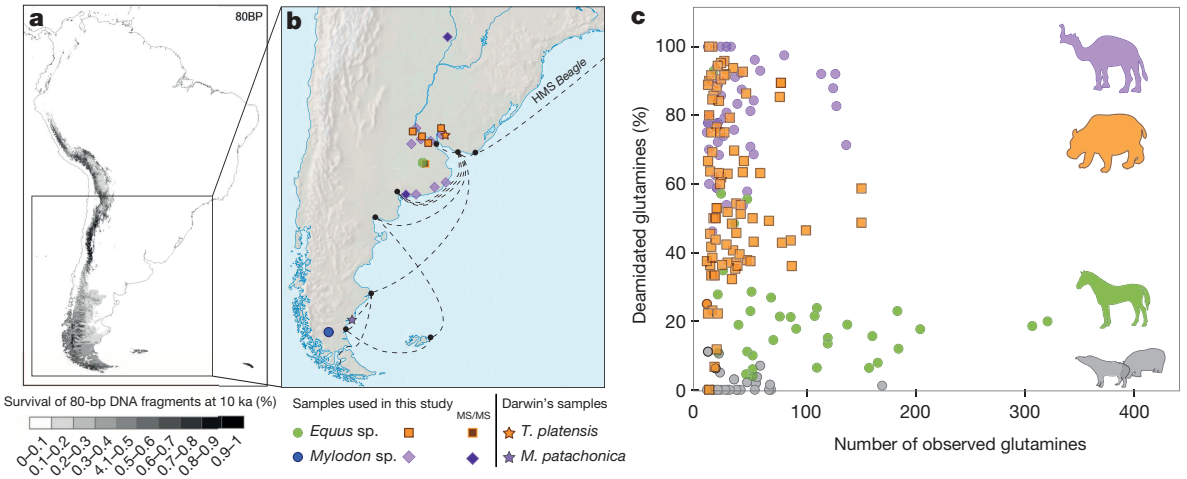


Figure 1 | Samples used in this investigation. **a**, Predicted survival of an 80-base-pair (bp) DNA fragment after 10,000 years (10 ka) modelled using the rate given in ref. 29. **b**, Location of finds by Darwin[1,2] and of samples used in this study (basemap[30]). **c**, Glutamine deamidation ratios for bone samples from the sequenced Pleistocene SANUs are high compared with coeval horse (MACN Pv 5719) as well as modern hippopotamus and tapir, providing support for the authenticity of the ancient sequences (see Supplementary Information).

[1]BioArCh, University of York, York YO10 5DD, UK. [2]Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany. [3]Department of Earth Sciences, Natural History Museum, London SW7 5BD, UK. [4]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen K, Denmark. [5]Institute of Zoology, Zoological Society of London, London NW1 4RY, UK. [6]CONICET- División Paleontología de Vertebrados, Museo de La Plata. Facultad de Ciencias Naturales y Museo de La Plata, Universidad Nacional de La Plata. Paseo del Bosque s/n, B1900FWA, La Plata, Argentina. [7]Sección Paleontología de Vertebrados. Museo Argentino de Ciencias Naturales "Bernardino Rivadavia", 470 Angel Gallardo Av., C1405DJR, Buenos Aires, Argentina. [8]Institute of Anthropology, Johannes Gutenberg-University, Anselm-Franz-von-Bentzel-Weg 7, D-55128 Mainz, Germany. [9]Department of Chemistry, University of York, York YO10 5DD, UK. [10]Bioscience Technology Facility, Department of Biology, University of York, York YO10 5DD, UK. [11]Department of Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA. [12]Target Discovery Institute, Nuffield Department of Medicine, University of Oxford, Roosevelt Drive, Oxford OX3 7FZ, UK. [13]Applications Development, Bruker Daltonik GmbH, 28359 Bremen, Germany. [14]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3b, 2200 Copenhagen, Denmark. [15]Department of Materials Science and Metallurgy, University of Cambridge, Cambridge CB3 0FS, UK. [16]Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland. [17]Institute for Biochemistry and Biology, Karl-Liebknecht-Strasse 24–25, 14476 Potsdam OT Golm, Germany. [18]Department of Mammalogy, American Museum of Natural History, New York, New York 10024, USA.

©2015 Macmillan Publishers Limited. All rights reserved

South American native ungulates (SANUs) are conventionally organized into five orders (Litopterna, Notoungulata, Astrapotheria, Xenungulata, and Pyrotheria) that are sometimes grouped together as a separate placental superorder (Meridiungulata)[10]. They appear very early in the Palaeogene record and evolved thereafter along many divergent lines, as their abundant fossil record attests. Most lineages had become extinct by the end of the Miocene epoch, although a few species of litopterns and notoungulates persisted into the Late Pleistocene epoch. Despite continuing interest in their evolutionary history (for example refs 5, 11–14), phylogenetic relationships of the major SANU clades to one another and to other placentals remain poorly understood (see Supplementary Information). Although some recent investigations (for example refs 4–6) have suggested that basal South American members of Litopterna conclusively group with certain Holarctic condylarths, and are thus best placed in Euungulata (Laurasiatheria), several other studies claim to have identified potential synapomorphies linking various SANU taxa with Afrotheria[5,6,15,16]. This latter view is broadly consistent with such indicators as prolonged late Mesozoic faunal exchange

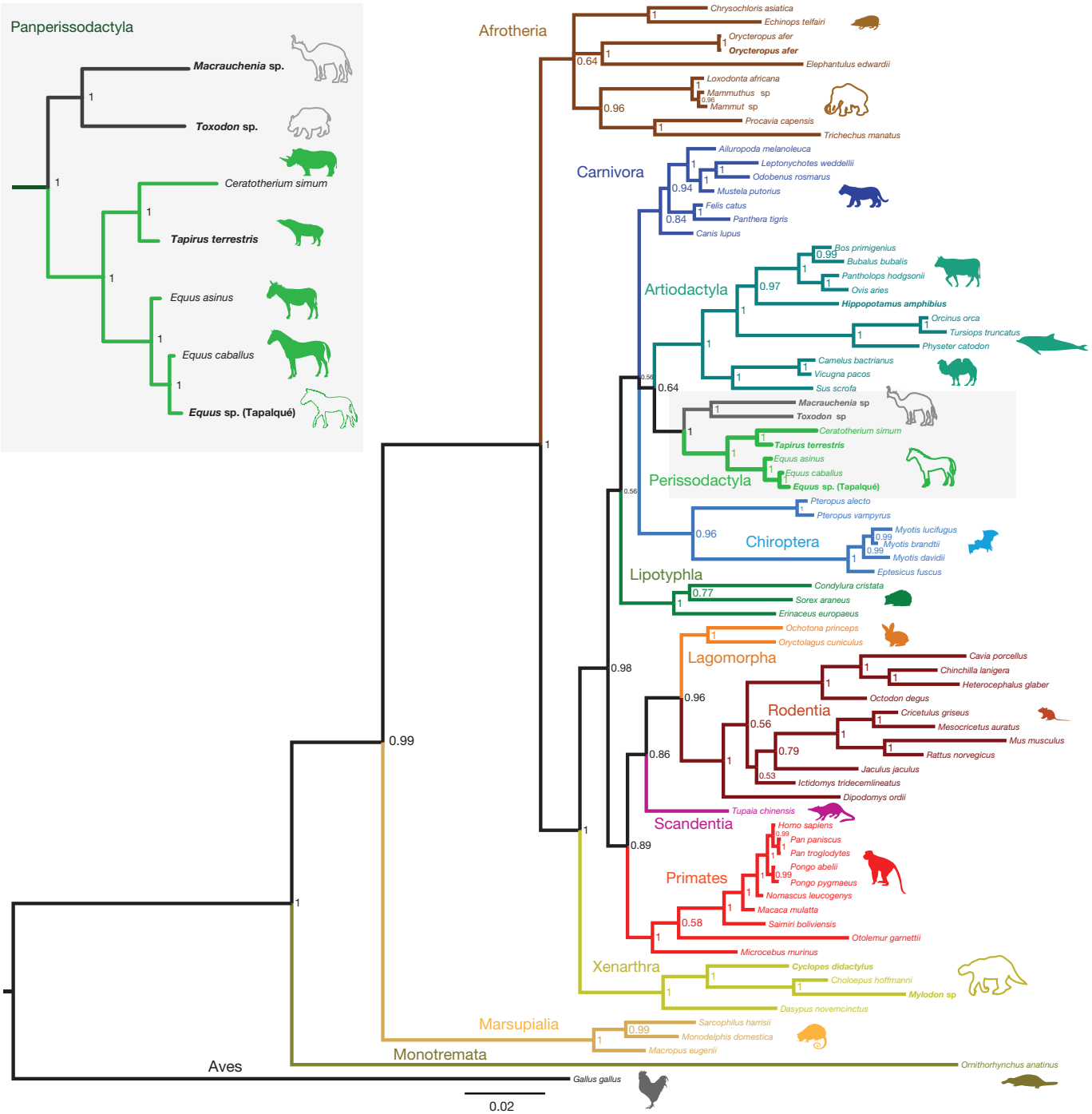


Figure 2 | Relationship of *Toxodon* (Notoungulata) and *Macrauchenia* (Litopterna) to other placental mammals. Fifty per cent majority rule Bayesian consensus tree of COL1 protein sequence data, with chicken (*Gallus*) as outgroup. Scale bar indicates branch length, expressed as the expected number of substitutions per site. Major clades (orders and superorders) are colour coded; species names in bold indicate collagen sequences derived from MS/MS rather than genomic data, fossil taxa depicted in silhouette. Inset: in all tree-reconstructions conducted (see Supplementary Information), *Toxodon* and *Macrauchenia* (dark grey) group monophyletically at the base of crown Perissodactyla (light green) with 100% posterior probability, forming Panperissodactyla.

©2015 Macmillan Publishers Limited. All rights reserved

between Gondwanan landmasses[17] and the possibility that Xenarthra (the other major endemic South American placental clade) is also related to Afrotheria[7–9,18]. However, most of the character evidence on which the SANU–Afrotheria sister-group hypothesis is based is in dispute[19,20]. In principle, a more definitive test of phylogenetic affinities could come from genomic data, but so far the application of ancient DNA techniques has been limited and DNA survival is predicted to be poor (Fig. 1a) (see Supplementary Information).

Type I collagen (COL1), a structural protein comprising two separate chains, COL1α1 and COL1α2 (coded by genes on separate chromosomes), is known to provide useful systematic information ('barcoding')[21], and can be recovered over significantly longer time spans than DNA[22]. Most of the 48 samples of *Toxodon* sp. and *Macrauchenia* sp. we analysed for sequence information came from localities in Buenos Aires province (Supplementary Information and Fig. 1b), especially from areas that experience subtropical to maritime–temperate climates[23]. Peptide mass fingerprinting (ZooMS) (Supplementary Information) of COL1 extracts[24] revealed variable levels of collagen preservation in the sample set (see Supplementary Information and Extended Data Table 1). After screening, two samples each of *Toxodon* and *Macrauchenia* displaying excellent COL1 preservation (see Extended Data Fig. 1) were selected for liquid chromatography–tandem mass spectrometry (LC–MS/MS) sequencing using a variety of LC–MS/MS platforms, and direct radiocarbon dating (Supplementary Information and Extended Data Table 1).

Combining analyses from a total of eight MS/MS runs, we were able to assemble near-complete COL1 sequences for *Macrauchenia* (89.4%) and *Toxodon* (91.0%), similar to levels of sequence coverage for modern samples. Comparative analyses with fossil and modern samples suggest that our SANU COL1 sequences are authentic: COL1 amino-acid sequence variation is located in similar positions along both COL1 chains compared with collagen sequences derived from genomic sources (Extended Data Fig. 2) and deamidation ratios conform to expectations for Pleistocene samples (Extended Data Fig. 3), a criticism of previous pre-Holocene collagen studies[25]. Independent manual *de novo* sequencing of product ion spectra for selected phylogenetically relevant peptides was in full agreement with sequence assignments from database searches. Furthermore, 86.70% and 94.41% of the assembled species consensus sequences for *Macrauchenia* and *Toxodon*, respectively, were covered by a minimum of two independent product ion spectra, with individual positions being covered by an average of 77.1 (for *Macrauchenia*) and 103.9 (for *Toxodon*) product ion spectra (Extended Data Table 2).

Molecular evidence for the phylogenetic placement of the extinct SANUs *Macrauchenia* and *Toxodon* was previously unavailable. To examine the phylogenetic position of these taxa, an alignment of 76 mammalian COL1 sequences and one outgroup (*Gallus*) was constructed from available mammalian genomic COL1 sequences in GenBank, as well as several MS/MS-derived protein sequences obtained for this study. A Bayesian phylogenetic tree was estimated from the data, with separate models of substitution applied to two partitions (COL1α1 and COL1α2). The resulting consensus tree (Fig. 2) is based solely on protein sequence data, but its topology corresponds closely to branching relationships in Placentalia recovered in recent molecular studies[7–9]. Furthermore, nodes poorly supported in this study (for example, those within Laurasiatheria) involve the same series of phylogenetic relationships that have proved difficult to resolve in other studies[5,7–9]. To examine how alternative topologies could affect the position of our target taxa we ran additional Bayesian analyses, using constraints mirroring differing mammal phylogenies (Extended Data Fig. 4 and Supplementary Information).

In all phylogenetic analyses performed with our data (including the use of unconstrained parsimony and probabilistic tree reconstruction methods), *Macrauchenia* and *Toxodon* formed a strongly supported monophyletic pair that grouped exclusively with Perissodactyla (as represented by extant *Equus*, *Tapirus*, and *Ceratotherium*). Neither showed any association with the clades conventionally contained in Afrotheria (see Supplementary Information). In future, and with more evidence, it may be appropriate to include these SANUs within an augmented definition of Perissodactyla. At present, we prefer to recognize Litopterna and Notoungulata as part of a branch-based rankless taxon Panperissodactyla, uniting all taxa more closely related to crown Perissodactyla than to any other extant taxon of placentals (see Supplementary Information).

Despite poor resolution at the base of Laurasiatheria, the fact that *Macrauchenia* and *Toxodon* were not recovered at a basal position within Euungulata would imply that the initial split between Perissodactyla and Artiodactyla occurred earlier than the origin of the SANU clades. Since fossil evidence indicates that both litopterns and notoungulates were already present in South America by the Early Palaeocene epoch[4,26], this would suggest that the divergence events leading to the modern orders must have occurred at, if not before, the Cretaceous/Palaeogene boundary (Supplementary Information and Extended Data Fig. 5).

These observations do not constitute a full molecular test of SANU monophyly, as there is no proteomic evidence available for members of the remaining orders (Astrapotheria, Xenungulata, Pyrotheria). As far as it is now known, Xenungulata and Pyrotheria became extinct in the Late Palaeogene, but some members of Astrapotheria (sometimes considered the sister group of Notoungulata[27]) persisted until the Middle Miocene (16.0–11.6 million years ago (Ma) (ref. 28)). This is well beyond the extrapolated estimate of less than 4.0 Ma for good collagen survival in an optimal (cool) burial environment[22], although the empirical limits on collagen survival under differing environmental conditions are poorly understood at present (see Supplementary Information).

The results presented here establish that, in principle, the approximately 2,100 residues (that is, one-fifth of the amino-acid residues analysed in ref. 9) comprising bone COL1 in placental mammals are sufficiently variable to provide reliable systematic information. Of course, a phylogeny based on two genes may be sensitive to factors affecting phylogenetic resolution such as gene lineage sorting, missing taxa, aberrant molecular rates, and selection acting on protein coding sequences. Despite this, the topology derived from the collagen sequences in this study is in broad agreement with other mammalian trees, and supports monophyletic placement of two Late Quaternary SANUs with a high degree of confidence. Reliable systematic information is an essential foundation for many other enquiries in evolutionary biology, including patterns of early Cenozoic mammalian divergence, radiation, extinction, and palaeobiogeography. With further development, molecular sequencing of degradation-resistant proteins such as bone COL1 is sure to open new vistas in the study of vertebrate evolution.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 September 2014; accepted 22 January 2015.
Published online 18 March 2015.

1. Owen, R. in *The Zoology of the Voyage of H.M.S. Beagle, under the Command of Captain Fitzroy, during the Years 1832 to 1836* (ed. Darwin, C.) Part I, Numbers I–IV (Smith Elder 1838–40).
2. Darwin, C. *Journal of Researches into the Geology and Natural History of the Various Countries Visited by H.M.S. Beagle: Under the Command of Captain FitzRoy, R.N. from 1832 to 1836* (Henry Colburn, 1839).
3. Husson, D. *et al.* Astronomical calibration of the Maastrichtian (Late Cretaceous). *Earth Planet. Sci. Lett.* **305,** 328–340 (2011).
4. De Muizon, C. & Cifelli, R. L. The "condylarths" (archaic Ungulata, Mammalia) from the early Palaeocene of Tiupampa (Bolivia): implications on the origin of the South American ungulates. *Geodiversitas* **22,** 1–150 (2000).
5. Agnolin, F. L. & Chimento, N. R. Afrotherian affinities for endemic South American "ungulates". *Mamm. Biol.* **76,** 101–108 (2011).
6. O'Leary, M. A. *et al.* The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science* **339,** 662–667 (2013).
7. Dos Reis, M. *et al.* Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. R. Soc. B* **279,** 3491–3500 (2012).

©2015 Macmillan Publishers Limited. All rights reserved

8. Song, S., Liu, L., Edwards, S. V. & Wu, S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl Acad. Sci. USA* **109,** 14942–14947 (2012).
9. Meredith, R. W. *et al.* Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334,** 521–524 (2011).
10. McKenna, M. C., Bell, S. K. & Simpson, G. G. *Classification of Mammals above the Species Level* (Columbia Univ. Press, 1997).
11. Simpson, G. G. The beginning of the age of mammals in South America. Part 2. *Bull. Am. Mus. Nat. Hist.* **137,** 1–259 (1967).
12. Patterson, B. & Pascual, R. The fossil mammal fauna of South America. *Q. Rev. Biol.* **43,** 409–451 (1968).
13. Cifelli, R. L. in *Mammal Phylogeny* (eds Szalay, F. S., Novacek, M. J. & McKenna, M. C.) 195–216 (Springer, 1993).
14. Horovitz, I. Eutherian mammal systematics and the origins of South American ungulates as based on postcranial osteology. *Bull. Carnegie Mus. Nat. Hist.* 63–79 (2004).
15. Asher, R. J. & Lehmann, T. Dental eruption in afrotherian mammals. *BMC Biol.* **6,** 14 (2008).
16. Sánchez-Villagra, M. R., Narita, Y. & Kuratani, S. Thoracolumbar vertebral number: the first skeletal synapomorphy for afrotherian mammals. *Syst. Biodivers.* **5,** 1–7 (2007).
17. Van Bocxlaer, I., Roelants, K., Biju, S. D., Nagaraju, J. & Bossuyt, F. Late Cretaceous vicariance in Gondwanan amphibians. *PLoS ONE* **1,** e74 (2006).
18. Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S. & Miller, W. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* **17,** 413–421 (2007).
19. Billet, G. & Martin, T. No evidence for an afrotherian-like delayed dental eruption in South American notoungulates. *Naturwissenschaften* **98,** 509–517 (2011).
20. Kramarz, A. & Bond, M. Critical revision of the alleged delayed dental eruption in South American "ungulates". *Mamm. Biol.* **79,** 170–175 (2014).
21. Van Doorn, N. L. in *Encyclopedia of Global Archaeology* 7998–8000 (Springer, 2014).
22. Buckley, M. & Collins, M. J. Collagen survival and its use for species identification in Holocene-lower Pleistocene bone fragments from British archaeological and paleontological sites. *Antiqua* **1,** e1 (2011).
23. Hamza, V. M. & Vieira, F. P. in *Climate Change - Geophysical Foundations and Ecological Effects* (eds Blanco, J. & Kheradmand, H.) Ch. 6, 113–136 (Intech, 2011).
24. Buckley, M., Collins, M., Thomas-Oates, J. & Wilson, J. C. Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **23,** 3843–3854 (2009).
25. Asara, J. M., Schweitzer, M. H., Freimark, L. M., Phillips, M. & Cantley, L. C. Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. *Science* **316,** 280–285 (2007).
26. Wilf, P., Rubén Cúneo, N., Escapa, I. H., Pol, D. & Woodburne, M. O. Splendid and seldom isolated: the paleobiogeography of Patagonia. *Annu. Rev. Earth Planet. Sci.* **41,** 561–603 (2013).
27. Van Valen, L. M. Paleocene dinosaurs or Cretaceous ungulates in South America? *Evol. Monogr.* **10,** 1–79 (1988).
28. Vizcaino, M., Mikolajewicz, U., Jungclaus, J. & Schurgers, G. Climate modification by future ice sheet changes and consequences for ice sheet mass balance. *Clim. Dyn.* **34,** 301–324 (2010).
29. Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B* **279,** 4724–4733 (2012).
30. US. Central Intelligence Agency. *The World Factbook 2013–14* (Central Intelligence Agency, 2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the Museo Argentino de Ciencias Naturales "Bernardino Rivadavia", Buenos Aires (MACN), the Museo de La Plata (MLP), and the Natural History Museum of Denmark, Copenhagen (ZMK), for allowing us to sample fossil specimens in their collections for this project. The American Museum of Natural History and the Copenhagen Zoo provided samples of extant mammals suitable for collagen extraction. Mogens Andersen and Kristian Gregersen of ZMK provided information on specimens in their care. This work was partly supported by SYNTAX award "Barcode of Death", European Research Council (ERC) Advanced Award CodeX, ERC Consolidator Award GeneFlow, SYNTHESYS FP7 grant agreement 226506, Engineering and Physical Sciences Research Council NE/G012237/1 and National Science Foundation OPP 1142052. J.T.-O. and D.A.A. are members of the York Centre of Excellence in Mass Spectrometry, created thanks to a major capital investment through Science City York, supported by Yorkshire Forward with funds from the Northern Way Initiative.

Author Contributions R.D.E.M., I.B., and M.J.C. conceived the project and coordinated the writing of the paper with F.W. and J.A.T., with all authors participating. J.N.G., A.K., M.R., E.C., and R.D.E.M. collected fossil and extant mammal samples for protein extraction. M.W., S.B., I.B., J.A.T., J.B., and M.H. conducted DNA analyses. F.W., M.W., P.A., S.K., C.B., C.K., D.A., J.T.-O., R.F., B.K., P.K., J.A.E., E.C., L.O., and M.J.C. performed protein analyses and interpretation of results. J.A.T., I.B., F.W., and M.W. conducted the phylogenetic analyses and constructed trees. S.T.T., J.N.G., M.R., D.M.P., and R.D.E.M. provided the historical, systematic, and palaeontological framework for this study. J.-J.H., E.W., and J.S. provided technical information. Final editing and manuscript preparation was coordinated by M.J.C., R.D.E.M., and I.B.

Author Information Raw MS/MS and PEAKS search files have been deposited to the ProteomeXchange with identifier PXD001411. Generated COL1 species consensus sequences will be available in the UniProt Knowledgebase under the accession numbers C0HJN3–C0HJP8. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.B. (i.barnes@nhm.ac.uk), F.W. (frido.welker@palaeo.eu), M.J.C. (matthew.collins@york.ac.uk), or R.D.E.M. (macphee@amnh.org).

©2015 Macmillan Publishers Limited. All rights reserved

METHODS

No statistical methods were used to predetermine sample size.

Zooarchaeology by MS screening. After zooarchaeology by MS (ZooMS) screening of selected *Macrauchenia* ($n = 26$) and *Toxodon* ($n = 22$), four bone specimens were selected for MS/MS analysis. Using a combination of enzymes, we were able to obtain sequence coverage of around 90% for COL1 for both genera. Subsamples of about 200 mg were taken from each bone or skin sample for COL1 extraction. Bone samples were demineralized in 0.6 M HCl for 8 days at 4 °C. The acid was removed and the samples were washed three times with ultrapure water then heated at 70 °C in 0.6 M HCl for 48 h to gelatinize the COL1. Samples were then ultrafiltered using 30 kilodalton filters and washed through with ultrapure water. Half a millilitre from each sample retentate was taken to dryness overnight in a vacuum centrifuge. One hundred microlitres of 50 mM ammonium bicarbonate solution (pH 8) was added to each sample. The samples were then digested with trypsin (0.5 μg μl^{-1}, for 16 h at 37 °C). After enzyme digestion, samples were acidified with 2 μl of 5% (volume %) trifluoroacetic acid (TFA). Samples were then concentrated using C18 ZipTips: the ZipTips were prepared using a conditioning solution of 50% acetonitrile, 49.9% water, 0.1% TFA; the tips were then washed with a washing solution of 0.1% TFA; the sample was then transferred over the column ten times; the tips were then washed again using 0.1% TFA solution; finally the sample was eluted using the conditioning solution. For ZooMS analysis, 1 μl of each sample was spotted in triplicate onto a ground steel plate with 1 μl of CHCA matrix solution (1% in 50% ACN/0.1% TFA (v/v/v)). MS analysis was on a Bruker ultraflex matrix-assisted laser desorption/ionization–tandem time of flight (MALDI–TOF/TOF) mass spectrometer over the *m*/*z* range 800–4,000 (Extended Data Fig. 1). Screening revealed large differences in COL1 spectral quality between samples. Of 46 SANU samples, only five (3 out of 20 from *Toxodon*, 2 out of 25 from *Macrauchenia*) yielded good ZooMS spectra. One of the three *Toxodon* samples (ZMK 22/1889) produced a few strong MS/MS spectra and only four samples (two each from *Macrauchenia* and *Toxodon*) were used in the main study.

MS/MS sequence analysis. Selected collagen extracts from pooled trypsin (0.4 μg μl^{-1}, 16 h, 37 °C) and elastase digests (0.8 μg μl^{-1}, 16 h, 37 °C) of two specimens of each SANU sample were analysed on both Thermo Scientific Orbitrap and Bruker maXis HD LC–MS/MS platforms. Additionally, Orbitrap and maXis HD instruments were also used for sequencing collagen from modern aardvark (*Orycteropus afer*), silky anteater (*Cyclopes didactylus*), hippopotamus (*Hippopotamus amphibius*), and South American tapir (*Tapirus terrestris*), as well as Pleistocene *Mylodon darwinii* and *Equus* sp. samples from South America.

Hybrid Quadrupole-Orbitrap. Sample separation was performed on an Ultimate 3000 RSLCnano LC system (Thermo Scientific). Peptides were first trapped on a Pepmap μ-pre-column (0.5 cm × 300 μm; Thermo Scientific) and separated on an EASY Spray PepMap UHPLC column (50 cm × 75 μm, 2 μm particles, 40 °C; Thermo Scientific) with a 60 min multi-step acetonitrile gradient ranging from 2% to 35% mobile phase B (mobile phase A: 0.1% formic acid/5% dimethylsulfoxide (DMSO) in water; mobile phase B: 0.1% formic acid/5% DMSO in acetonitrile) at a flow rate of 250 nl min^{-1}. Mass spectra were acquired on a Q Exactive Hybrid Quadrupole-Orbitrap mass spectrometer at a resolution of 70,000 at *m*/*z* 200 using an ion target of 3×10^6 and maximal injection time of 100 ms between *m*/*z* 380 and 1,800. Product ion spectra of up to 15 precursor masses at a signal threshold of 4.7×10^4 counts and a dynamic exclusion for 27 s were acquired at a resolution of 17,500 using an ion target of 10^5 and a maximal injection time of 128 ms. Precursor masses were isolated with an isolation window of 1.6 Da and fragmented with 28% normalized collision energy.

Bruker maXis HD. Sample separation was performed on an Ultimate 3000 RSLCnano LC system (Thermo Scientific). Peptides were first trapped on a Pepmap pre-column (2 cm × 100 μm; Thermo Scientific) and separated on a PepMap UHPLC column (50 cm × 75 μm, 2 μm particles; Thermo Scientific) with a 120 min multi-step acetonitrile gradient ranging from 5 to 35% mobile phase B (mobile phase A: 0.1% formic acid in water; mobile phase B: 0.1% formic acid in acetonitrile) at a flow rate of 400 nl min^{-1}. A CaptiveSpray nanoBooster source (Bruker Daltonik), with acetonitrile as a dopant, was used to interface the LC system to the maXis HD UHR-Q-TOF system (Bruker Daltonik). Source parameters were set to 3 l min^{-1} dry gas and 150 °C dry heater; nitrogen 'flow' setting for the nanoBooster was set to 0.2 bar. Mass spectra were acquired in the *m*/*z* range 150–2,000 at a spectral acquisition rate of 2 Hz. Precursors were fragmented with a fixed cycle time of 4 s using a dynamic method adapting spectra rates between 2 and 10 Hz based on precursor intensities. Dynamic exclusion was set to 0.4 min combined with reconsideration of an excluded precursor for fragmentation if its intensity rose by a factor of 3.

Collagen type I sequence assembly. Product ion data from the maXis HD and Orbitrap platforms were analysed in three stages. Initially MASCOT (Matrix Science) was used to search against the UniColl database, a database of non-redundant synthetic collagen peptides, to generate a list of ranked peptides for each spectrum. Sequences derived from this exercise were added to a local database of genomic and published collagen sequences and common laboratory contaminants, and the original data were then re-analysed by PEAKS[31] using this new database (for an example of PEAKS output see Extended Data Fig. 1b–d).

As an independent check, a limited number of the product ion spectra of peptides (previously assigned by PEAKS) were also manually *de novo* interpreted (by J.T.-O.) without prior knowledge of the assignment, in all cases with full agreement between the two approaches.

Generation of, and searching against, Unicoll. Publicly available COL1α1 and COL1α2 sequences were concatenated and aligned using Mafft[32] with subsequent manual alignment of misaligned sites using Bioedit and Geneious version 4.6 (ref. 33). A custom Python script was used to digest the COL1 with trypsin *in silico*. For each tryptic fragment, all variable amino-acid positions across the aligned sequences were recorded. A new sequence was created for every permutation of these variable sites. These sequences were concatenated and stored in FASTA format with a header indicating the position in the original alignment. The result was a database with each entry a concatenation of sequences representing every permutation of observed mutations for that particular tryptic fragment. One tryptic fragment of the sequence (COL1α2 positions 870–905) was too variable to include without exceeding available memory. Only the original observed variants were included for this part of the sequence. Using this strategy, it was possible to generate the equivalent of more than 10^{200} alternative collagen 'sequences' (cf. 10^{82}, which is the upper estimate of the number of atoms in the universe).

MS/MS data files were merged and submitted to Mascot with enzyme set to Trypsin/P; variable modifications for deamidated (NQ), Lys→Hyl (K), oxidation (M), and Pro→Hyp (P); peptide mass tolerance ±10 ppm; and fragment mass tolerance ±0.07 Da. The structure of sequence entries in Unicoll meant that it could not accommodate missed cleavages. Select summaries containing matched peptides with a Mascot score greater than 30 were exported into Microsoft Excel for each analysis. Peptides were identified by picking the highest scoring hits for each tryptic fragment, if the score exceeded 40; whereas for matches with scores between 30 and 40, the spectra were inspected manually to choose the best hit among the possibilities given by the search engine.

Searching data using PEAKS. Product ion spectra were searched using PEAKS software against a database comprising genomic COL1 sequences plus fossil consensus sequences, composed of UniColl peptide hits, with missing and low coverage regions filled with conserved mammalian COL1 sequences (see Phylogenetic reconstruction section below). Additionally, common laboratory contaminants were included in database searches. Full PEAKS searches (Peptide *de novo*, PEAKS DB, PEAKS PTM, and SPIDER) were performed with peptide mass tolerance ±10 ppm and fragment mass tolerance ±0.07 Da, in addition to respective platform and enzyme details. Searches were performed allowing for deamidated (NQ), Lys→Hyl (K), oxidation (M), and Pro→Hyp (P). False discovery rate was put at 0.5% and peptide scores were only accepted with $-10\log_{10}(P \text{ value})$ scores of at least 30 and average local confidence (%) at least 50. Where there was ambiguity in interpretation of the spectra, peptides were selected on the basis of knowledge of sequence constraints, post-translational modifications, and fragmentation patterns.

Reference sequence authentication. To check the quality of our MS/MS COL1 sequences, we sampled a modern and a fossil sample for which we had independent genomic data, specifically (1) a modern aardvark sample (*Orycteropus afer*) and (2) a fossil equid bone from a geological formation rich in SANU fossils with their respective genome sequences. The fossil sample had similar collagen yields and ZooMS profile to the SANU samples used for MS/MS sequencing (Pleistocene horse, Tapalqué, South America; Fig. 1b) (MS/MS sequence analysis, above). Our modern aardvark MS/MS sequence was identical to that of the protein product inferred from the released genomic sequence. For the Pleistocene *Equus* sp. sequence, two amino-acid substitutions were detected (*T*>L, COL1α1; H>D, COL1α2), similar to the maximum number of differences observed in a recent study comparing *Equus* genomes with the *Equus ferus caballus* reference genome[34].

***De novo* sequence authentication.** The absence of corresponding genomic data prevented similar comparisons with MS/MS-derived sequences for the SANU species. Instead we assessed amino-acid substitution locations along COL1α1 and COL1α2 chains both in our (and previously published[35]) and in fossil COL1 sequences with genomic data, using the COL1α1 and COL1α2 sequence of the Tasmanian devil (*Sarcophilus harrisii*) as an outgroup to eutherian mammals. Carboxy- (C-) and amino- (N-)terminal telopeptides were removed as they were rarely observed from fossil samples. COL1 position numbers are given as a continuous count with COL1α1 and COL1α2 concatenated, with COL1α1 ranging from position 1 to position 1014, and COL1α2 ranging from 1015 to 2028.

We found that the location of amino-acid variation along the COL1α1 and COL1α2 chains was similar among the different COL1 sequences obtained from genomic sources (Extended Data Fig. 2). We identified several regions, mainly

©2015 Macmillan Publishers Limited. All rights reserved

located in COL1α1, that appeared to lack sequence variation among the four major mammalian superorders. This could be a result of the functional importance of some of these regions during COL1 fibril formation, α1 and α2 chain binding, and COL1 hydroxylation[36–38]. Additionally, we observed a substitution rate in COL1α2 roughly twice that observed in COL1α1.

Comparing COL1 sequences derived from MS/MS data in this and an earlier study[35] with genomic data for laurasiatheres revealed good correspondence in the location of substitutions along the COL1α1 and COL1α2 chains between our results and genomic data (Extended Data Fig. 2). The MS/MS data in ref. 35 for laurasiatheres were derived from a single species (*Manis tetradactyla*). Sequence variation from those data compared with genomic data were similar, although we noted that several regions displaying high rates of amino-acid substitution were missing from the *Manis* consensus sequence provided (notably around positions 726–756, 991–1089, 1306–1364, 1423–1443 and 1899–1977).

Reference 35 provided two xenarthran and five afrotherian COL1 sequences obtained using mass spectrometric sequencing. Regions with high substitution complexity were missing from the consensus sequences provided, for Afrotheria (1024–1089, 1588–1599, 1740–1754) and Xenarthra (1024–1089, 1207–1234, 1348–1364, 1588–1638, 1771–1806, 1921–1947). The absence of such regions prohibited the inclusion of these sequences in our phylogenetic tree-building, as the majority of informative positions were missing from the sequences provided. For substitution locations, our data suggest structural and/or functional organization of these, and their frequency, in specific regions of both chains.

We criticized claims of authentic collagen sequences retrieved from a *Tyrannosaurus rex* sample[39] based in part on the low levels of reported deamidation[40], and more recently have demonstrated an increase in glutamine deamidation in archaeological rather than modern collagen, which correlated with thermal age (Extended Data Fig. 3 and ref. 41); similar levels have been reported for Pleistocene mammoths and equids[34].

Deamidation ratios observed for glutamine here are consistent with ancient collagen of equivalent thermal age (Extended Data Fig. 1). The lowest levels of Gln to Glu deamidation are observed in modern samples from hippopotamus (1.8% ± 3.2) and tapir (5.7% ± 10.9) bone. The highest levels of Gln deamidation occur in the radiocarbon samples from dead *Macrauchenia* (Glu = 82.8% ± 14.3). The *Toxodon* samples are less deamidated (Glu = 59.2% ± 24.5), which is consistent with a Late Pleistocene date (12,000 years ago). However, by contrast, the Pleistocene equid is much better preserved (Glu = 18.9% ± 18.4), despite the fact that it cannot be much younger than *Toxodon* (Fig. 1c).

DNA extraction and sequencing. Approximately 250 mg of the three samples with the highest number of peaks in the mass spectra from each species (see Zooarchaeology by MS screening, above) were used for DNA extraction. DNA extraction was performed as in the method described in ref. 42. PCR primers were designed to target Perissodactyla- and Laurasiatheria-specific regions of the cytb, COX1, 16S, and 12S genes using mitochondrial DNA sequences downloaded from the National Center for Biotechnology Information (NCBI) (Supplementary Table 1). Primer design used the program Primer3. PCR was performed for 60 cycles and samples were visualized on 2.5% agarose gel. Products were successfully amplified from several samples whereas PCR controls showed no amplification products. BLAST searches of the sequences obtained revealed no homology to any previously derived sequence for several of the products, whereas sequences from two Macrauchenia samples showed high similarity (98% and 99%, respectively) to domestic pig sequences, a common contaminant in ancient DNA analyses[43]. A Pleistocene horse bone from the same depositional context as some of the SANU specimens yielded a sequence 98% identical to modern horse (*Equus caballus*), suggesting that the failure to amplify putative SANU DNA sequences by PCR was not because of technical problems, but because of a lack of endogenous DNA in the samples investigated.

DNA next-generation sequencing approach. After failing to amplify endogenous DNA through Sanger sequencing of targeted PCR products, we applied a non-targeted, next-generation sequencing (NGS) shotgun approach in a further attempt to identify whether endogenous DNA could be obtained. Based on the collagen sequencing results, *Macrauchenia* sample 12-1641 (metapodial) was selected as the most likely candidate for NGS analyses. DNA extractions of *Macrauchenia* sample 12-1641 followed protocols described in ref. 44 and were performed in the dedicated ancient-DNA laboratory at Royal Holloway, University of London, UK. The library was constructed in a dedicated laboratory for ancient DNA (Johannes Gutenberg University, Mainz, Germany) using a modified version of the protocol in ref. 45. Modifications were as follows: the initial DNA fragmentation step was not required, and all clean-up steps used MinElute PCR purification kits. For the blunt-end repair step, Buffer Tango and ATP were replaced with 0.1 mg ml^{-1} BSA and 1× T4DNA ligase buffer. The proceeding clean-up step was replaced by an inactivation step, heating to 75 °C for 10 min. For the adaptor ligation step, 0.5 mM ATP replaced the T4 DNA Ligase buffer. The index PCR step followed a further protocol[46] using AmpliTaq Gold DNA polymerase and the addition of 0.4 mg ml^{-1} BSA. The index PCR was set for 20 cycles with three PCR reactions conducted per library. The indexed library was sequenced on an Illumina HiSeq platform (Mainz) using a single lane, paired-end read, sequencing run.

Bioinformatics methods and conclusion. Paired-end reads were quality trimmed ($q = 10$) with cut-adapt[47] and then sequences were simultaneously adaptor trimmed and the paired reads joined together with Seq-Prep (available from https://github.com/jstjohn/SeqPrep). Reads shorter than 17 base pairs were discarded. In the absence of any close phylogenetic relative (required for the accurate genomic mapping of reads), processed reads were *de novo* assembled into contigs using clc_denovo_assembler (available in CLC Assembly Cell version 4.2), with contigs shorter than 70 base pairs discarded. Two approaches were then used to investigate the data for mammalian genomic sequences (which had proved successful for other ancient DNA NGS samples).

To examine whether there were any mammalian DNA sequences suitable for phylogenetic analysis in our data set, first, contigs were blasted using blastn to a local nucleotide database, downloaded from NCBI. Custom perl scripts (available on request) were used to assign taxonomic and gene information to BLAST hits. These results were searched for standard orthologous mitochondrial and nuclear phylogenetic sequences. Each of the potential hits blasting to mammalian sequences was inspected; however, all were assignable to bacterial elements, and no blast hit could be attributed to mammalian genes.

Second, two separate BLAST databases were generated: one from the contigs and a second from the processed reads, using the makeblastdb command in BLAST. These databases could then be queried with mammalian (including perissodactyl) mitochondrial and nuclear phylogenetic sequences of interest using blastn. Neither of these searches returned any matching contigs. Thus, the NGS data set yielded nothing of use for phylogenetic analysis, and gave no indication that any *Macrauchenia* DNA had persisted in the sample.

Phylogenetic reconstruction. Before the advent of DNA-based molecular phylogeny, variations in protein structure and sequence had been used to explore evolutionary relationships[48,49]. The comparative data set for this paper was built using consensus amino-acid sequences for COL1α1 and COL1α2 generated by MS/MS analysis for the target taxa *Toxodon* and *Macrauchenia* as well as representatives of all extant major mammalian clades. Leucine (L) and isoleucine (I) were converted into isoleucines as these are isobaric and low-energy MS/MS sequencing is not capable of discriminating between them. Partition Finder[50] was used to select the best-fit partitioning scheme from the amino-acid data. This was identified as two separate partitions, for Col1a1 and Col1a2. Bayesian phylogenies were generated using MrBayes version 3.2.1 (ref. 51) with the amino-acid model estimated from the data (to allow model jumping between fixed-rate amino-acid models, the prior for the amino-acid model was set as prset aamodellpr = mixed). The proportion of invariant sites, and the distribution of rates across sites (approximating to a gamma distribution), were also estimated from the data. Two chains were run for 5 million generations (sampled every 500), with convergence between chains assessed in Tracer version 1.6 (ref. 52). All effective sample sizes of parameters were greater than 100. After burn-in was removed, a majority rule consensus tree was constructed, using the sumt command in MrBayes, from the trees sampled in the posterior distribution.

To test for the robustness of the results of the Bayesian analysis under other methods of tree reconstruction, we also conducted maximum likelihood and maximum parsimony analyses. We performed parsimony analyses running PAUP* version 4.0b10 (ref. 53), using the heuristic search option with a random taxon addition sequence (1,000 repetitions) and TBR branch swapping, and rooting the tree along the branch leading to Aves. A maximum likelihood phylogeny was estimated in RAxML version 7.0.4 (ref. 54). A Dayhoff model of protein sequence evolution with gamma-distributed variation in rates across sites (corresponding to the PROTGAMMADAYHOFFF model in RAxML) was applied to each partition. Twenty separate maximum likelihood analyses were performed (using the '−f d' command in RAxML), and the tree with the highest likelihood was chosen from this set.

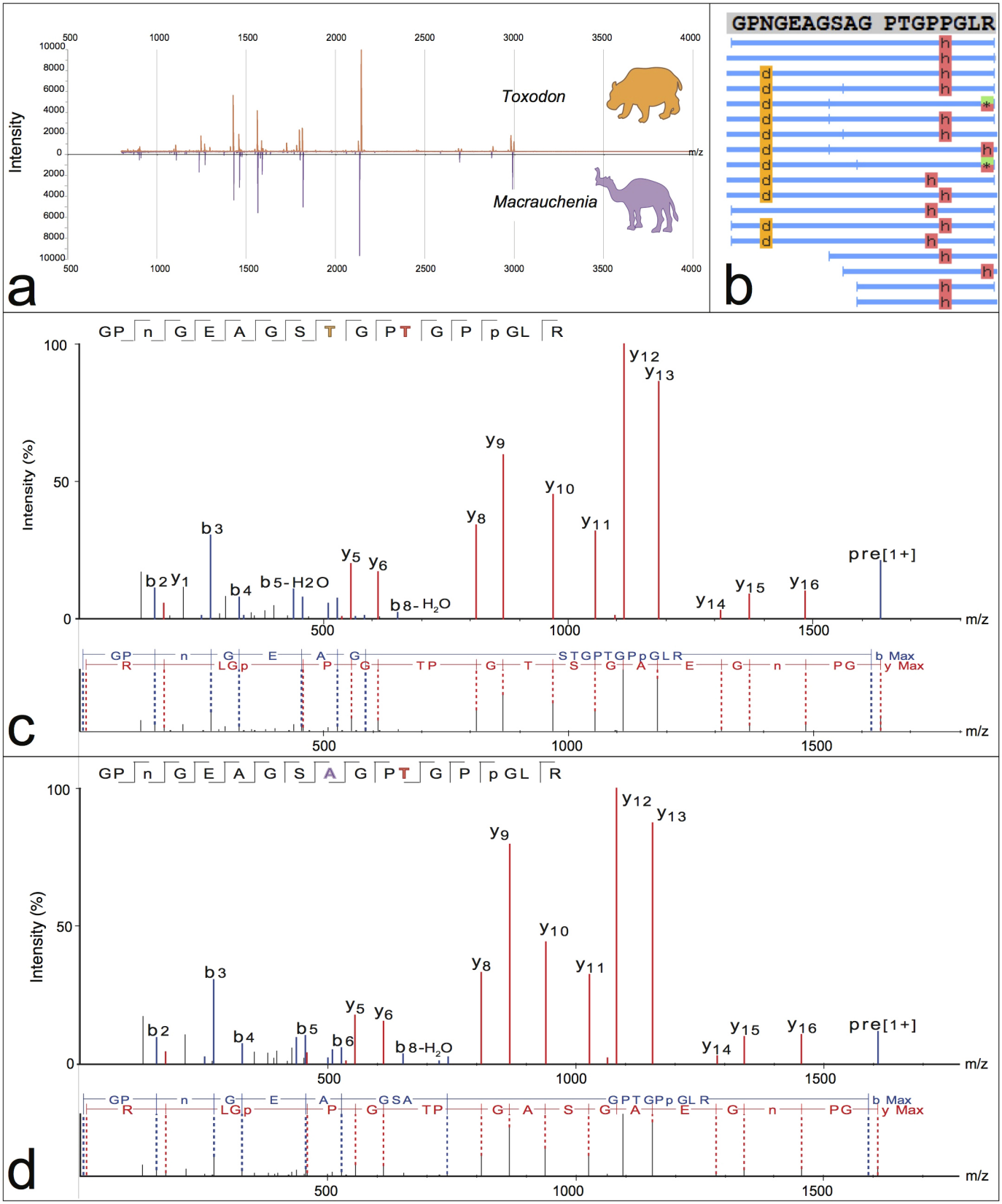
Molecular clock analysis. Fossil-calibrated phylogenies were constructed in BEAST version 1.7 (ref. 55) with the Dayhoff amino-acid model (chosen under the MrBayes mixed model) together with the proportion of invariant sites and the distribution of rates across sites (approximating a gamma distribution) applied to each partition. Analyses were run under a strict clock (estimated from the data), with the Yule model of speciation, for 10 million generations (sampled every 1,000 generations). Clock and tree parameters were linked across partitions. Prior distributions on the root and 33 other nodes were applied based on an interpretation of the mammalian fossil record (see Supplementary Table 3)[7,56]. The clock rate prior was set as an uninformative uniform distribution (upper = 10^{100}, lower = 10^{-12}). All other priors were left as the default values in BEAUti[57]. Full details of all prior distributions for divergence times are presented in Supplementary Table 3. As in the case of the MrBayes analysis, convergence and effective sampling were assessed using Tracer

©2015 Macmillan Publishers Limited. All rights reserved

1.7. A maximum clade credibility tree was constructed using TreeAnnotator (available with BEAST) from the trees sampled in the posterior distribution.

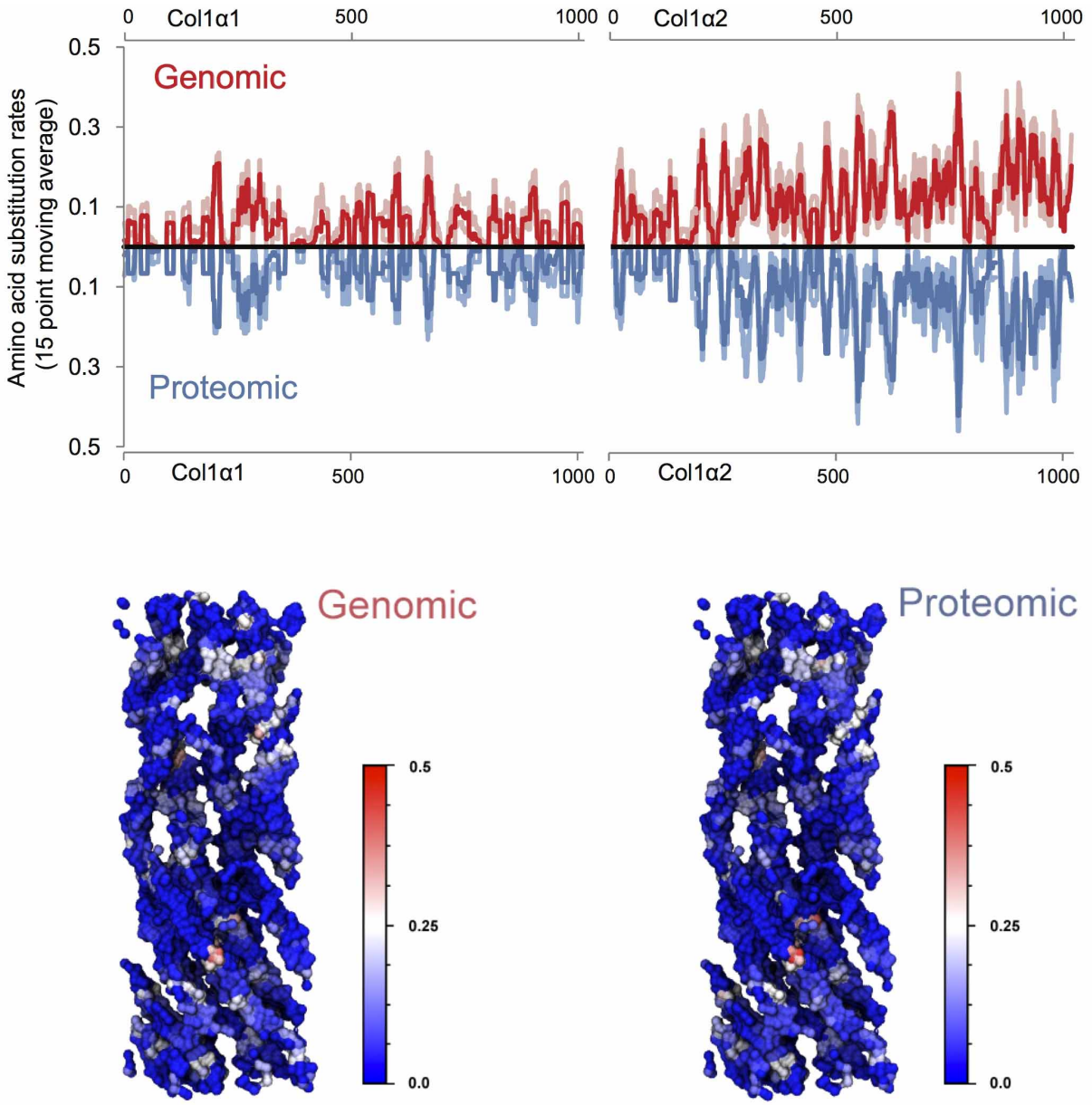
31. Ma, B. *et al.* PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17,** 2337–2342 (2003).
32. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
33. Drummond, A. J. *et al.* Geneious v4.7. (Geneious, 2010).
34. Orlando, L. *et al.* Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499,** 74–78 (2013).
35. Buckley, M. A molecular phylogeny of *Plesiorycteropus* reassigns the extinct mammalian order 'Bibymalagasia'. *PLoS ONE* **8,** e59614 (2013).
36. Terajima, M. *et al.* Glycosylation and cross-linking in bone type I collagen. *J. Biol. Chem.* http://dx.doi.org/10.1074/jbc.M113.528513 (2014).
37. Hudson, D. M., Weis, M. & Eyre, D. R. Insights on the evolution of prolyl 3-hydroxylation sites from comparative analysis of chicken and *Xenopus* fibrillar collagens. *PLoS ONE* **6,** e19336 (2011).
38. Hudson, D. M., Werther, R., Weis, M., Wu, J.-J. & Eyre, D. R. Evolutionary origins of C-terminal (GPP)n 3-hydroxyproline formation in vertebrate tendon collagen. *PLoS ONE* **9,** e93467 (2014).
39. Schweitzer, M. H. *et al.* Analyses of soft tissue from *Tyrannosaurus rex* suggest the presence of protein. *Science* **316,** 277–280 (2007).
40. Buckley, M. *et al.* Comment on ''Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry''. *Science* **319,** 33 (2008).
41. Van Doorn, N. L., Wilson, J., Hollund, H., Soressi, M. & Collins, M. J. Site-specific deamidation of glutamine: a new marker of bone collagen deterioration. *Rapid Commun. Mass Spectrom.* **26,** 2319–2327 (2012).
42. Rohland, N. & Hofreiter, M. Comparison and optimization of ancient DNA extraction. *Biotechniques* **42,** 343–352 (2007).
43. Leonard, J. A. *et al.* Animal DNA in PCR reagents plagues ancient DNA research. *J. Archaeol. Sci.* **34,** 1361–1366 (2007).
44. Brace, S. *et al.* Population history of the Hispaniolan hutia *Plagiodontia aedium* (Rodentia: Capromyidae): testing the model of ancient differentiation on a geotectonically complex Caribbean island. *Mol. Ecol.* **21,** 2239–2253 (2012).
45. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010,** http://dx.doi.org/10.1101/pdb.prot5448 (2010).
46. Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52,** 87–94 (2012).
47. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
48. Zuckerkandl, E., Jones, R. T. & Pauling, L. A comparison of animal hemoglobins by tryptic peptide pattern analysis. *Proc. Natl Acad. Sci. USA* **46,** 1349–1360 (1960).
49. Sarich, V. M. & Wilson, A. C. Rates of albumin evolution in primates. *Proc. Natl Acad. Sci. USA* **58,** 142–148 (1967).
50. Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29,** 1695–1701 (2012).
51. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61,** 539–542 (2012).
52. Rambaut, A., Drummond, A. J. & Suchard, M. Tracer v.1. 6. (2013).
53. Swofford, D. L. *PAUP**. Phylogenetic Analysis Using Parsimony (*and Other Methods) v.4.0b10 (Sinauer Associates, 2003).
54. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
55. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–1973 (2012).
56. Benton, M. J., Donoghue, P. C. J. & Asher, R. J. in *The Timetree of Life* (eds Hedges, B. S. & Kumar, S.) 35–86 (Oxford Univ. Press, 2009).
57. Rambaut, A. & Drummond, A. BEAUTi v.1. 4.2. Bayesian evolutionary analysis utility (2007).
58. Humphrey, W., Dalke, A. & Schulten, K. VMD – Visual Molecular Dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).

©2015 Macmillan Publishers Limited. All rights reserved



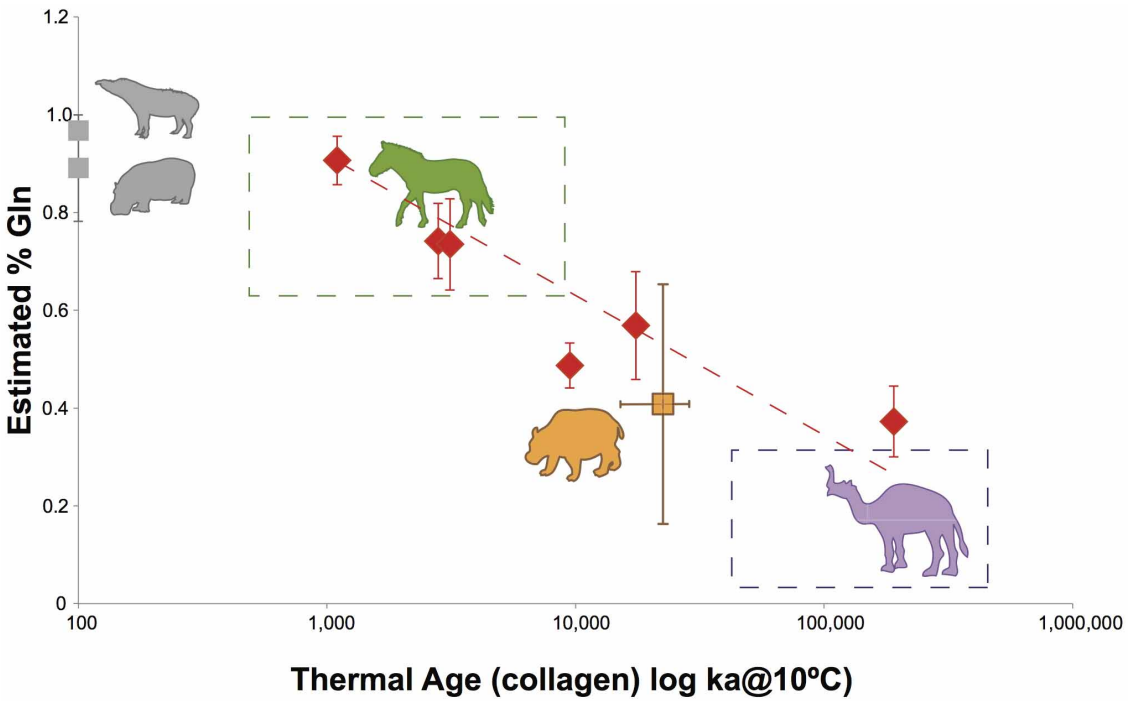
Extended Data Figure 1 | Examples of MALDI–TOF–MS and MS/MS product ion spectra. **a**, MALDI–TOF–MS ZooMS spectra for *Toxodon* (upper) and *Macrauchenia* (lower) were used to screen for samples for the best collagen preservation. **b**, PEAKS alignment of matching product ion spectra for *Macrauchenia* MLP 96-V-10-19 (specimen sample number MLP2012.12) highlighting peptides aligning to the sequence GPNGEAGSAGPTGPPGLR. **c**, **d**, Annotated PEAKS report of product ion spectra for the same peptide sequence detailed in **b** for *Toxodon* (**c**) and *Macrauchenia* (**d**), detailing differences between both genera (gsT and gsA, highlighted) and shared substitutions compared with *Equus* (gpA for *Equus*, gpT for *Toxodon* and *Macrauchenia*). Note in **b** that both deamidation (N→D) and variable hydroxylation (P→h) were detected in different peptides covering this region of the sequence.

©2015 Macmillan Publishers Limited. All rights reserved



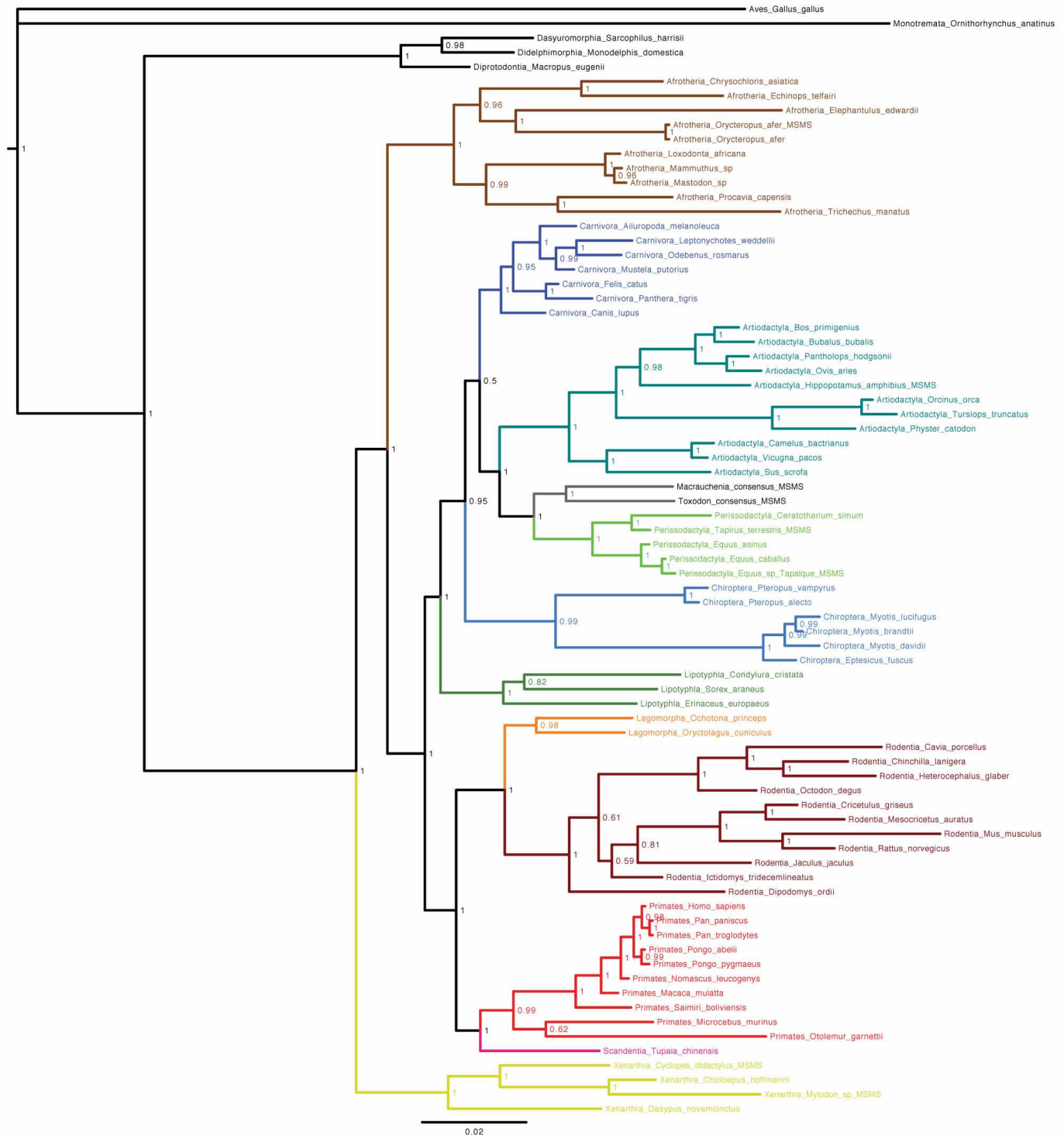
Extended Data Figure 2 | Collagen type I substitution variability for placental mammals (genomic and proteomic data) compared with the dasyurid marsupial *Sarcophilus harrisii* (Tasmanian devil) as outgroup. Substitution variability scores range between 0 and 1 and incorporate sequence coverage for a given number of species over a 15-amino-acid moving average (95% standard deviation in lighter tone). Top, along-chain variation in genomic sequence variability (upper red) is similar to proteomic sequence variability (lower blue) both for COL1α1 and for COL1α2 chains. Bottom, molecular surface rendering (via VMD[58]) of the collagen unit cell taken from coordinates given in Protein Data Bank accession number 3HR2. Colours represent genomic (left) and proteomic (right) sequence variability throughout the structure.

©2015 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 3 | Comparison of levels of deamidation for samples in this study with ref. 22 (diamonds). The *Macrauchenia* sample was ^{14}C dead, consistent with observed levels of deamidation, which are lower than either *Toxodon* dated to 12,000 years ago or *Equus* sp. (Tapalqué; not dated). Dotted lines indicate error ranges on Gln estimation for samples that were not dated or were undateable. The measurement approach used in this study—frequency of deamidation in positions represented in at least seven MS/MS spectra—is different from the approach used in ref. 22, so the absolute values may not be directly comparable.

©2015 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 4 | Bayesian constraint tree based on phylogeny published in figure 1 in ref. 6. See Methods and Supplementary Information section 3.2 for further details and discussion.

©2015 Macmillan Publishers Limited. All rights reserved

Extended Data Figure 5 | Maximum clade credibility phylogeny from BEAST molecular dating analysis. Branch lengths are measured in millions of years; scale axis indicates intervals of 100 Ma. Node labels show 95% highest probability densities for molecular dates (in millions of years). Fossil constraints are provided in Supplementary Table 3. Vertical dashed line indicates Cretaceous/Palaeogene boundary.

©2015 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | *Toxodon* and *Macrauchenia* specimens used in this study

Genus	Species	Museum collection	Specimen number	MS/MS code	Location	Longitude	Latitude	Province	Age as recorded	Element	Age (14C cal yrbp)	±	Radiocarbon Laboratory number	Effective Burial Temperature T_{eff} (°C)	Thermal age[1] ka@ 10°C
Toxodon	sp.	MLP	12-1132		Arrecifes	-60.1	-34.1	Buenos Aires	Pampean Fm	Axis				13	93
Toxodon	sp.	MLP	12-1159		Olivera	-59.2	-34.6	Buenos Aires	Pampean Fm	Skull cap, juvenile				13	94
Toxodon	sp.	MLP	12-2432		no location	--	--	Buenos Aires	Pampean Fm	Jaw, juvenile					
Toxodon	sp.	MLP	M-218		no location	--	--	Buenos Aires	Pampean Fm	Mandible					
Toxodon	sp.	MLP	12-1169		Arrecifes	-59.1	-34.4	Buenos Aires	Pampean Fm	Mandible				13	93
Toxodon	sp.	MLP	12-1224		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Upper incisor					
Toxodon	sp.	MLP	12-1125		Arrecifes	-60.1	-34.1	Buenos Aires	Pampean Fm	3rd lumbar				13	94
Toxodon	*platensis*	MLP	12-1190		Chelforó	-39.0	-66.5	Buenos Aires	Pampean Fm	Turbinals					
Toxodon	sp.	MLP	12-1227		Luján	-59.0	-34.5	Buenos Aires	Lujanian	Incisor				17	188
Toxodon	sp.	MLP	12-1160		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Pterygoid				17	188
Toxodon	sp.	MLP	94-II-1-17		Río Quequén Saladc	-60.5	-38.9	Buenos Aires	Pleistocene	Ulna				10	51
Toxodon	sp.	MLP	94-II-1-17		Río Quequén Saladc	-60.5	-38.9	Buenos Aires	Pleistocene	Ulna				10	51
Toxodon	**sp.**	**MLP**	**44-XII-29-5**	**MLP2012.04**	**Tapalqué**	**-60.0**	**-36.3**	**Buenos Aires**	**Pleistocene**	**Mandible**	**11,900**	**35**	**(UCIAMS 143034)**	13	22
Toxodon	sp.	MLP	12-1180		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Maxilla				17	188
Toxodon	sp.	MACN Pv	5287		no location	--	--	--	Fm Pampeana	Metacarpal					
Toxodon	**sp.**	**MACN Pv**	**17710**		**Arroyo Tapalqué**	**-60.0**	**-36.4**	**Buenos Aires**	**L Pleistocene**	**Tibia**	**12,040**	**70**	**(UCIAMS 143035)**	13	22
Toxodon	**sp.**	**MACN Pv**	**17710**	**MACN2012.12**	**Arroyo Tapalqué**	**-60.0**	**-36.4**	**Buenos Aires**	**L Pleistocene**	**Tibia**	**12,040**	**70**	**(UCIAMS 143035)**	13	22
Toxodon	sp.	MACN Pv	5717		Arroyo Tapalqué	-60.0	-36.4	Buenos Aires	Pampean Fm/ Lujanian	Molar				12	72
Toxodon	sp.	MACN Pv	2760		no location	--	--	--	--	Metapodial					
Toxodon	sp.	MACN Pv	5712		Arroyo Tapalqué	-60.0	-36.4	Buenos Aires	Pampeana/ Lujanian	Tibia				12	72
Toxodon	sp.	MACN Pv	9666		Río Quequen Saladc	-60.7	-38.4	Buenos Aires	L Pampeano Fm	Cervical				10	52
Toxodon	*platensis*	Z.M.K.	16/1887		Arroyo del Medio	-60.8	-33.6	Buenos Aires/ Santa Fe	Pleistocene	Jaw				14	102
Macrauchenia	*patachonica*	MLP	12-1641		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Metapodial	12,185	55	(OxA-25840)	14	28
Macrauchenia	*patachonica*	MLP	12-1648		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Metapodial				17	188
Macrauchenia	*patachonica*	MLP	12-1659		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Metapodial				17	188
Macrauchenia	*patachonica*	MLP	12-1488		no location	--	--	Buenos Aires	--	Phalanx					
Macrauchenia	**sp.**	**MLP**	**96-V-10-19**		**Rio Pilcomayo**	**-57.7**	**-25.2**	**Formosa**	**Pleistocene**	**Thoracic**	**no date**			15	0
Macrauchenia	**sp.**	**MLP**	**96-V-10-19**	**MLP2012.12**	**Rio Pilcomayo**	**-57.7**	**-25.2**	**Formosa**	**Pleistocene**	**Thoracic**	**no date**			15	0
Macrauchenia	sp.	MLP	50-X-5-5		Mar del Plata	-57.6	-38.1	Buenos Aires	--	Jaw				11	59
Macrauchenia	sp.	MLP	71-III-6-1		Río Salado	-61.0	-34.6	Buenos Aires	--	Tibia				13	89
Macrauchenia	*patachonica*	MLP	12-1434		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Mandible				17	188
Macrauchenia	*patachonica*	MLP	12-1458		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Mandible				17	188
Macrauchenia	*patachonica*	MLP	12-2326		no location	--	--	Buenos Aires	Pampean Fm	Skull, tooth					
Macrauchenia	*patachonica*	MLP	12-1660		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Tarsal				17	188
Macrauchenia	sp.	MLP	80-IX-5-1		Laguna de la Bombilla	-69.3	-44.1	Chubut	Pleistocene	Cervical					
Macrauchenia	sp.	MLP	12-1661		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Pedal sesamoid				17	188
Macrauchenia	sp.	MLP	12-1660		Luján	-59.0	-34.5	Buenos Aires	Pampean Fm	Phalanx				17	188
Macrauchenia	sp.	MLP	12-1457		no location	--	--	--	Pampean Fm	Jaw					
Macrauchenia	patachonica	MLP	80-II-10-2		Rio Quequén Saladc	-60.7	-38.4	Buenos Aires	Lujanian	Jaw				10	52
Macrauchenia	sp.	MACN Pv	6708		Rio Quequén	-58.8	-38.2	Buenos Aires	--	Tooth				11	56
Macrauchenia	**sp.**	**MACN Pv**	**18952**	**MACN2012.02**	**ıear Monte Hermosa**	**-61.3**	**-39.0**	**Buenos Aires**	**Lujanian**	**Cervical**	**no date**			11	61
Macrauchenia	sp.	MACN Pv	7107		Arroyo Seco, Mirama	-60.5	-33.1	Buenos Aires	Pampeana Fm/ Lujanian?	Pedal sesamoid				15	117
Macrauchenia	sp.	MACN Pv	3		Salto	-60.3	-34.3	Buenos Aires	Pampean Fm	Tibia				13	91
Macrauchenia	sp.	MACN Pv	2 (05)		Salto	-60.3	-34.3	Buenos Aires	Pampean Fm	Humerus				13	91
Macrauchenia	sp.	MACN Pv	2 (06)		Salto	-60.3	-34.3	Buenos Aires	Pampean Fm	Tibiofibula				13	91
Macrauchenia	sp.	MACN Pv	2 (07)		Salto	-60.3	-34.3	Buenos Aires	Pampean Fm	Femur				13	91
Macrauchenia	patachonica	MACN Pv	6708		Río Quequén (Grande	-59.1	-34.6	Buenos Aires	Pampeana Fm/ Lujanian	Mandible				13	94
Macrauchenia	sp.	MACN Pv	10530		Río Quequén (Grande	-58.7	-38.6	Buenos Aires	Pampeana Fm/ Lujanian	Metapodial				11	56

Specimens highlighted in bold produced high-quality collagen and were sequenced. Specimens that appear twice were re-sampled. MACN Pv, Museo Argentino de Ciencias Naturales (vertebrate palaeontology collection), Buenos Aires, Argentina; MLP, Museo de La Plata; UCIAMS, Keck Carbon Cycle AMS Spectrometer facility, University of California, Irvine, USA; ZMUC, Natural History Museum of Denmark and Zoological Museum, Copenhagen, Denmark. [1]Thermal age of samples with a location, but without a radiocarbon date, are calculated at 50,000 years ago.

©2015 Macmillan Publishers Limited. All rights reserved

Extended Data Table 2 | Comparative run statistics combining multiple runs

Species	Museum Collection Number	Laboratory Specimen Number	Platform	Enzymatic digestion	Runs	#Measured spectra (# of runs)	#Matched COL1 spectra (%)	Triple-helical sequence coverage (%)	Fold coverage	% Unique coverage
Equus sp. (Tapalqué)		**MACN2010.03**	**Orbitrap**	**Trypsin/P**	**Consensus**	**31,309 (5)**	**9,530 (30.4)**	**87.3**	**85.9**	
Macrauchenia sp.		**-**	**-**	**-**	**Consensus**	**78,515 (4)**	**9,400 (12.0)**	**89.4**	**77.1**	
	18952	MACN2012.02	Bruker maXis HD	Trypsin/P	MACN201202	17,334	1,485 (8.6)	63.3	16.6	0.19
	18952	MACN2012.02	Orbitrap	Trypsin/P+elastase	York14	6,625	769 (11.6)	59.0	9.6	0
	96-V-10-19	MLP2012.12	Bruker maXis HD	Trypsin/P	MLP2012.12	34,525	3,410 (9,9)	77.3	35.4	2
	96-V-10-19	MLP2012.12	Orbitrap	Trypsin/P+elastase	York15	20,031	3,736 (18.7)	88.5	32.8	11.11
Toxodon sp.		**-**	**-**	**-**	**Consensus**	**82,448 (4)**	**12,028 (14.6)**	**91.0**	**103.9**	
	44-XII-29-5	MLP2012.04	Bruker maXis HD	Trypsin/P	MLP2012.04	20,499	2,720 (13.3)	80.7	28.1	1.29
	44-XII-29-5	MLP2012.04	Orbitrap	Trypsin/P+elastase	York13	20,706	3,610 (17.4)	84.0	35.8	1.81
	17710	MACN2012.12	Bruker maXis HD	Trypsin/P	MACN201212	20,134	2,188 (10.9)	76.7	21.0	2.1
	17710	MACN2012.12	Orbitrap	Trypsin/P+elastase	York12	21,109	3,510 (16.6)	81.7	35.5	0.38
Mylodon darwinii	**MLP 94-VIII-10-32**		**Orbitrap**	**Trypsin/P**	**Consensus**	**16,592 (1)**	**1,230 (7.4)**	**67.8**	**14.3**	
****Tapirus terrestris***		**-**	**Orbitrap**	**Trypsin/P+elastase**	**Consensus**	**17,459 (1)**	**1,111 (6.4)**	**92.0**	**9.8**	
****Hippopotamus amphibius***		**-**	**Orbitrap**	**Trypsin/P+elastase**	**Consensus**	**22,450 (1)**	**3,080 (13.7)**	**89.6**	**26.1**	
****Orycteropus afer***	**AMNH 51910**		**Orbitrap**	**Trypsin/P**	**Consensus**	**20,481 (1)**	**3,673 (17.9)**	**93.8**	**33.1**	
****Cyclopes didactylus***	**AMNH 99199**		**Orbitrap**	**Trypsin/P+elastase**	**Consensus**	**41,046 (2)**	**3,230 (7.9)**	**83.1**	**26.5**	

Taxa with asterisks are modern; others are fossil. Spectra were acquired on two platforms: Orbi-trap for *Macrauchenia*, *Toxodon*, *Tapirus*, *Hippopotamus*, *Orycteropus*, *Mylodon*, and *Cyclopes*; and maXis HD for Tapalqué *Equus*, *Macrauchenia*, and *Toxodon*. Individual samples were digested using either trypsin (Tapalqué *Equus*, *Macrauchenia*, *Toxodon*, *Orycteropus*, and *Mylodon*) or trypsin pooled with elastase digests (*Macrauchenia*, *Toxodon*, *Tapirus*, *Hippopotamus*, and *Cyclopes*).

View publication stats

©2015 Macmillan Publishers Limited. All rights reserved