LETTER

doi:10.1038/nature12323

Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse

Ludovic Orlando[1]*, Aurélien Ginolhac[1]*, Guojie Zhang[2]*, Duane Froese[3], Anders Albrechtsen[4], Mathias Stiller[5], Mikkel Schubert[1], Enrico Cappellini[1], Bent Petersen[6], Ida Moltke[4,7], Philip L. F. Johnson[8], Matteo Fumagalli[9], Julia T. Vilstrup[1], Maanasa Raghavan[1], Thorfinn Korneliussen[1], Anna-Sapfo Malaspinas[1], Josef Vogt[6], Damian Szklarczyk[10]†, Christian D. Kelstrup[10], Jakob Vinther[11]†, Andrei Dolocan[12], Jesper Stenderup[1], Amhed M. V. Velazquez[1], James Cahill[5], Morten Rasmussen[1], Xiaoli Wang[2], Jiumeng Min[2], Grant D. Zazula[13], Andaine Seguin-Orlando[1,14], Cecilie Mortensen[1,14], Kim Magnussen[1,14], John F. Thompson[15], Jacobo Weinstock[16], Kristian Gregersen[1,17], Knut H. Røed[18], Véra Eisenmann[19], Carl J. Rubin[20], Donald C. Miller[21], Douglas F. Antczak[21], Mads F. Bertelsen[22], Søren Brunak[6,23], Khaled A. S. Al-Rasheid[24], Oliver Ryder[25], Leif Andersson[20], John Mundy[26], Anders Krogh[1,4], M. Thomas P. Gilbert[1], Kurt Kjær[1], Thomas Sicheritz-Ponten[6,23], Lars Juhl Jensen[10], Jesper V. Olsen[10], Michael Hofreiter[27], Rasmus Nielsen[28], Beth Shapiro[5], Jun Wang[2,26,29,30] & Eske Willerslev[1]

The rich fossil record of equids has made them a model for evolutionary processes[1]. Here we present a 1.12-times coverage draft genome from a horse bone recovered from permafrost dated to approximately 560–780 thousand years before present (kyr BP)[2,3]. Our data represent the oldest full genome sequence determined so far by almost an order of magnitude. For comparison, we sequenced the genome of a Late Pleistocene horse (43 kyr BP), and modern genomes of five domestic horse breeds (*Equus ferus caballus*), a Przewalski's horse (*E. f. przewalskii*) and a donkey (*E. asinus*). Our analyses suggest that the *Equus* lineage giving rise to all contemporary horses, zebras and donkeys originated 4.0–4.5 million years before present (Myr BP), twice the conventionally accepted time to the most recent common ancestor of the genus *Equus*[4,5]. We also find that horse population size fluctuated multiple times over the past 2 Myr, particularly during periods of severe climatic changes. We estimate that the Przewalski's and domestic horse populations diverged 38–72 kyr BP, and find no evidence of recent admixture between the domestic horse breeds and the Przewalski's horse investigated. This supports the contention that Przewalski's horses represent the last surviving wild horse population[6]. We find similar levels of genetic variation among Przewalski's and domestic populations, indicating that the former are genetically viable and worthy of conservation efforts. We also find evidence for continuous selection on the immune system and olfaction throughout horse evolution. Finally, we identify 29 genomic regions among horse breeds that deviate from neutrality and show low levels of genetic variation compared to the Przewalski's horse. Such regions could correspond to loci selected early during domestication.

In 2003, we recovered a metapodial horse fossil at the Thistle Creek site in west-central Yukon Territory, Canada (Fig. 1a). The fossil was from an interglacial organic unit associated with the Gold Run volcanic ash, dated to 735 ± 88 kyr BP[2,3] (Fig. 1b). Relict ice wedges below the unit indicate persistent permafrost since deposition (Supplementary Information, section 1.1), whereas the organic unit, hosting the fossil, indicates a period of permafrost degradation, or a thaw unconformity[7], during a past interglacial as warm or warmer than present[3], and rapid deposition during either marine isotope stage 19, 17 or 15. This indicates that the fossil dates to approximately 560–780 kyr BP. The metapodial shows typical caballine morphology, consistent with Middle rather than the smaller Late Pleistocene horse fossils from the area (Fig. 1c and Supplementary Information, section 1.2). This age is consistent with small mammal fossils from this unit indicating a Late Irvingtonian, or Middle Pleistocene, age[3], and infinite radiocarbon dates[8].

Theoretical[9] and empirical evidence[10] indicates that this age approaches the upper limit of DNA survival. So far, no genome-wide information has been obtained from fossil remains older than 110–130 kyr BP[11]. Time-of-flight secondary ion mass spectrometry (TOF-SIMS) on the ancient horse bone revealed secondary ion signatures typical of collagen within the bone matrix (Fig. 2a and Supplementary Table 7.1), and high-resolution tandem mass spectrometry sequencing[12] revealed 73 proteins, including blood-derived peptides (Supplementary Information, section 7.4). This is consistent with good biomolecular preservation, suggesting possible DNA survival. Therefore, we conducted larger-scale destructive sampling for genome sequencing.

We used Illumina and Helicos sequencing to generate 12.2 billion DNA reads from the Thistle Creek metapodial. Mapping against the horse reference genome yielded ~1.12× genome coverage. We based the size distribution of ancient DNA templates on collapsed Illumina

[1]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen K, Denmark. [2]Shenzhen Key Laboratory of Transomics Biotechnologies, BGI-Shenzhen, Shenzhen 518083, China. [3]Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Alberta T6G 2E3, Canada. [4]The Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. [5]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, California 95064, USA. [6]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark. [7]Department of Human Genetics, The University of Chicago, Chicago, Illinois 60637, USA. [8]Department of Biology, Emory University, Atlanta, Georgia 30322, USA. [9]Department of Integrative Biology, University of California, Berkeley, California 94720, USA. [10]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3b, 2200 Copenhagen, Denmark. [11]Jackson School of Geosciences, The University of Texas at Austin, 1 University Road, Austin, Texas 78712, USA. [12]Texas Materials Institute, The University of Texas at Austin, Austin, Texas 78712, USA. [13]Government of Yukon, Department of Tourism and Culture, Yukon Palaeontology Program, PO Box 2703 L2A, Whitehorse, Yukon Territory Y1A 2C6, Canada. [14]Danish National High-throughput DNA Sequencing Centre, University of Copenhagen, Øster Farimagsgade 2D, 1353 Copenhagen K, Denmark. [15]NABsys Inc, 60 Clifford Street, Providence, Rhode Island 02903, USA. [16]Archeology, University of Southampton, Avenue Campus, Highfield, Southampton SO17 1BF, UK. [17]Zoological Museum, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark. [18]Department of Basic Sciences and Aquatic Medicine, Norwegian School of Veterinary Science, Box 8146 Dep, N-0033 Oslo, Norway. [19]Département histoire de la Terre, UMR 5143 du CNRS, paléobiodiversité et paléoenvironnements, MNHN, CP 38, 8, rue Buffon, 75005 Paris, France. [20]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, SE-751 23 Uppsala, Sweden. [21]Baker Institute for Animal Health, Cornell University, Ithaca, New York 14853, USA. [22]Center for Zoo and Wild Animal Health, Copenhagen Zoo, 2000 Frederiksberg, Denmark. [23]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2970 Hørsholm, Denmark. [24]Zoology Department, College of Science, King Saud University, Riyadh 11451, Saudi Arabia. [25]San Diego Zoo's Institute for Conservation Research, Escondido, California 92027, USA. [26]Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. [27]Department of Biology, The University of York, Wentworth Way, Heslington, York YO10 5DD, UK. [28]Departments of Integrative Biology and Statistics, University of California, Berkeley, Berkeley, California 94720, USA. [29]King Abdulaziz University, Jeddah 21589, Saudi Arabia. [30]Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China. †Present addresses: Bioinformatics Group, Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland (D.S.); Departments of Earth Sciences and Biological Sciences, University of Bristol BS8 1UG, UK (Ja.V.).
*These authors contributed equally to this work.

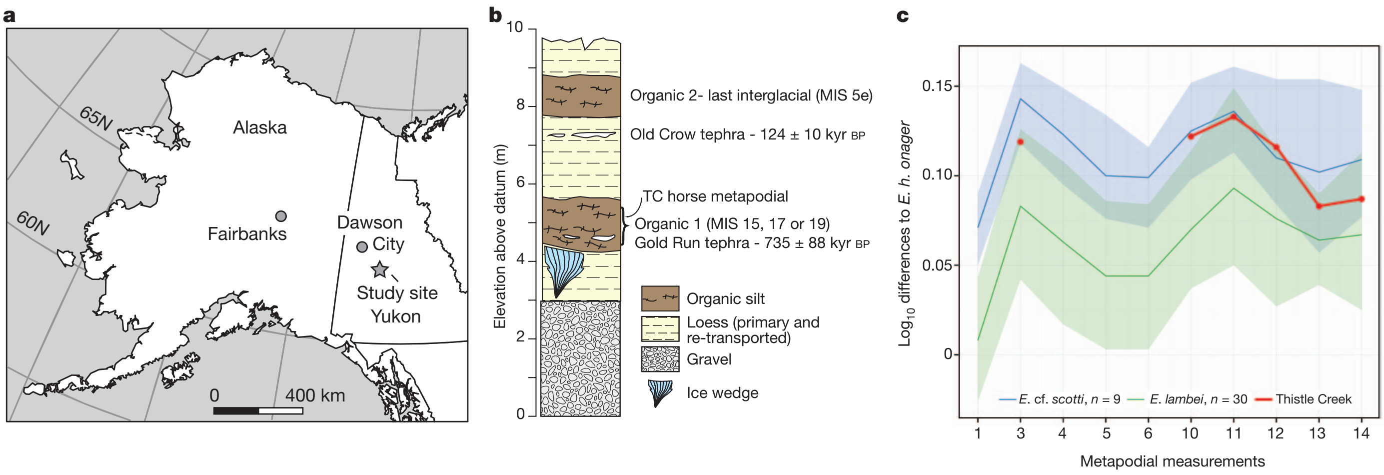

Figure 1 | The early Middle Pleistocene horse metapodial from Thistle Creek (TC). **a**, Geographical localization. **b**, Stratigraphic setting. **c**, Morphological comparison to Middle and Late Pleistocene horses from Beringia. Simpson's ratio diagrams contrasting $\log_{10}$ differences in 10 metapodial measurements between horse fossils and a reference (*E. hemionus onager*) are shown for a series of 9 and 30 horses from the Middle and the Late Pleistocene era, respectively (Supplementary Information, section 1.2). The full distribution range between minimal and maximal values is presented within shaded areas. Numbers reported on the *x* axis refer to the following measurements: 1, maximal length; 3, breadth at the middle of the diaphysis; 4, depth at the middle of the diaphysis; 5, proximal breadth; 6, proximal depth; 10, distal supra-articular breadth; 11, distal articular breadth; 12, depth of the keel; 13, least depth of medial condyle; 14, greatest depth of medial condyle.

read pairs (Supplementary Fig. 4.4), yielding an average length of 77.5 base pairs (bp). The specimen is male based on X to autosomal chromosome coverage (Supplementary Information, section 4.2b) and the presence of Y-chromosome markers (Supplementary Information, section 4.1d). Endogenous read content was lower for Illumina (0.47%) than Helicos (4.21%) using standard[8] or improved[13] single-strand template preparation procedures. This is probably due to 3′ ends available at nicks, resistance of undamaged modern DNA contaminants to denaturation, and Helicos ability to sequence short templates. Despite this, endogenous DNA content was >16.6–20.0-fold lower than for Saqqaq Palaeo-Eskimo[14] and Denisovan specimens[15], both sequenced to high depth.

Several observations support genome sequence authenticity. First, a 348-bp mitochondrial control region segment was replicated independently (Supplementary Fig. 2.2 and Supplementary Information, section 2.4). Second, phylogenetic analyses on data obtained with two sequencing platforms in different laboratories are consistent (Supplementary Fig. 8.4), ruling out post-purification contamination. Third, autosomal, Y-chromosomal and mitochondrial DNA analyses place the Thistle Creek specimen basal to Late Pleistocene and modern horses (Fig. 3a and Supplementary Figs 8.1–8.4). Fourth, we found signs of severe biomolecular degradation, including levels of cytosine deamination at overhangs considerably higher than observed in 28 younger permafrost-preserved fossils from the Late Pleistocene (Fig. 2c, Supplementary Fig. 6.40 and Supplementary Table 6.1) and protein deamidation levels[12,16] (Fig. 2b and Supplementary Information, section 7.5) greater than those reported for younger permafrost-preserved bones.

We additionally sequenced genomes of a 43-kyr-old (pre-domestication) horse (1.8× coverage), a modern donkey (16×; Supplementary Fig. 4.1), 5 modern domestic horses (Arabian, Icelandic, Norwegian fjord, Standardbred and Thoroughbred; 7.9×–21.1×) and one modern Przewalski's horse (9.6×; Supplementary Table 2.1), considered to possibly represent the last surviving wild horse population. We used this data set to address fundamental questions in horse evolution: (1) the timing of the origins of the genus *Equus*; (2) the demographic history of modern horses; (3) the divergence time of horse populations forming the Przewalski's and domestic lineages; (4) the extent to which the Przewalski's horse has remained isolated from domestic relatives; (5) the timing of gene expansions within the horse genome; (6) the identification of genes potentially under selection during horse evolution.

As no accepted *Equus* fossils exist before 2.0 Myr BP[4,5] (Supplementary Information, section 9.1d), the date of the last common ancestor that gave rise to extant horses versus donkeys, asses and zebras[17] remains heavily debated. Proposed dates extend as early as 4.2–4.5 Myr BP on the basis of palaeontological estimates[18] to over 6.0 Myr BP according to molecular analyses[19]. We addressed this issue by taking advantage of the established age for the Thistle Creek horse. As a sample cannot be older than the population it belonged to, we explored a full range of possible calibrations for the *Equus* most recent common ancestor (MRCA) and calculated the divergence time between the populations of the ancient Thistle Creek horse and modern horses[20] (Supplementary Information, section 10.1). Calibrations resulting in divergence times younger than the Thistle Creek bone age were rejected, providing a credible confidence range for the MRCA of *Equus*. We found rates consistent with the *Equus* MRCA living 3.6–5.8 Myr BP to be compatible with our data (Fig. 3b and Supplementary Figs 10.1–10.3). We also found support for slower mutation rates in horse than human (Supplementary Information, section 8.4 and Supplementary Table 8.5), implying a minimal date of 4.07 Myr BP for the MRCA of *Equus* (Supplementary Figs 10.1–10.3). We therefore propose 4.0–4.5 Myr BP for the MRCA of all living *Equus*, in agreement with recent molecular findings[17] and the oldest palaeontological records for the monodactyle *Plesippus simplicidens*, which some[18] consider the earliest fossil of *Equus*. Our result indicates that the evolutionary timescale for the origin of contemporary equid diversity is at least twice that commonly accepted.

Second, we reconstructed horse population demography over the last 2 Myr. The pairwise sequential Markovian coalescent (PSMC) approach[21] shows that horses experienced a population minimum approximately 125 kyr BP, corresponding to the last interglacial when environmental conditions were similar to now throughout their range. The population expanded during the cold stages of marine isotope stage (MIS) 4 and 3 as grasslands expanded. A peak was reached 25–50 kyr BP and was followed by an approximately 100-fold collapse, probably resulting from major climatic changes and related grassland contraction after the Last Glacial Maximum[22] (Fig. 4 and Supplementary Figs 9.4–9.5). A similar demographic history was inferred from Bayesian skyline reconstructions using 23 newly characterized ancient mitochondrial genomes (Supplementary Fig. 9.6). These results support suggestions[22] that climatic changes are major demographic drivers for horse populations. PSMC analyses also revealed two earlier demographic phases (Fig. 4b and Supplementary Figs 9.4–9.5), with population sizes peaking 190–260 kyr BP and 1.2–1.6 Myr BP, respectively, followed by 1.7-fold and 8.1-fold collapses. Extremely low population sizes were inferred approximately 500–800 kyr BP, a time period

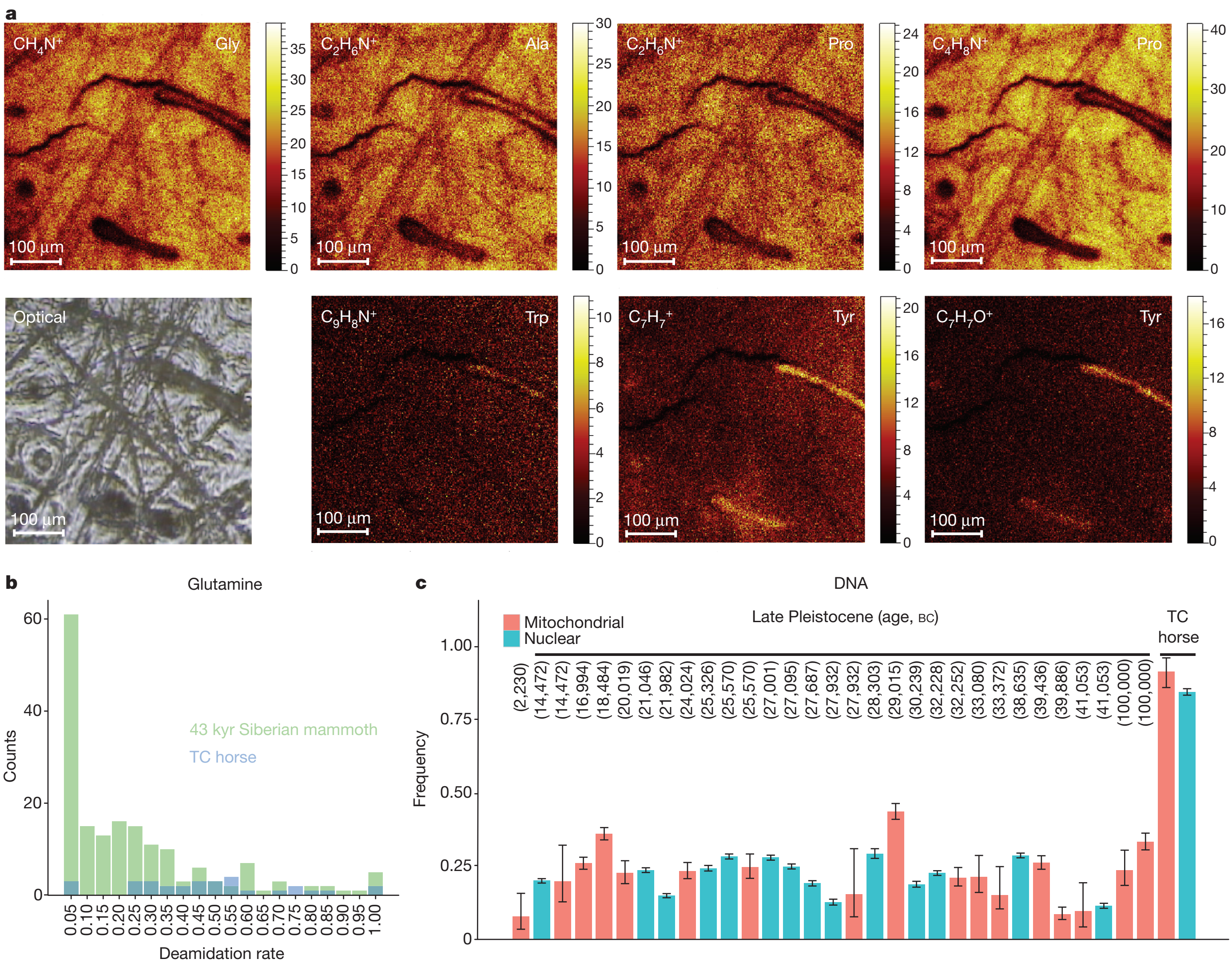

Figure 2 | Amino acid, protein and DNA preservation of the Thistle Creek horse bone. **a**, Amino acid signatures. Secondary ions, characteristic of five amino acids over- or under-represented in collagen, were detected by TOF-SIMS (Supplementary Information, section 7.1). The size of secondary ion maps is 500 × 500 µm^2 with a resolution of 256 × 256 pixels. **b**, Glutamine deamidation. The observed distribution of glutamine deamidation levels (Supplementary Information, section 7.5) is blue for the Thistle Creek (TC) horse bone and green for a 43-kyr-old Siberian mammoth bone. **c**, Post-mortem DNA damage. Maximum likelihood estimates of cytosine deamination at 5′ overhangs were estimated for 29 permafrost-preserved horse bones, including the Thistle Creek bone (Supplementary Information, section 6.3). Mitochondrial and nuclear estimates are provided in red and blue, respectively. Calibrated radiocarbon dates (BC) are provided when available (Supplementary Tables 2.3–4). Error bars refer to 2.5% and 97.5% quantile values, estimated following convergence of the maximum likelihood procedure.

that covers the divergence time of the Thistle Creek and contemporary horse populations. This result may relate to population fragmentation when horses colonized Eurasia from America, in agreement with the earliest presence of horses in Eurasia 750 kyr BP[4].

We next investigated whether Przewalski's horse indeed represents the last survivor of wild horses. Native to the Mongolian steppes, this horse was listed as extinct in the wild (IUCN red list[23]) but has been reassigned to endangered after successful conservation and reintroduction. Using maximum likelihood phylogenetic analyses and topological tests (Supplementary Information, sections 8.2–8.3), we found that the Przewalski's horse genome falls outside a monophyletic group of domestic horses. The MRCA of Przewalski's and domestic horse sequences dates to 341–431 kyr BP (Supplementary Table 8.3), a period consistent with previous estimates[6]. We estimated the divergence time between populations of Przewalski's and domestic horses to approximately 38–72 kyr BP (Supplementary Tables 10.4–10.6). Our 43 kyr BP horse genome branched off before the Przewalski's and domestic horse lineages diverged (Fig. 3a). This specimen belonged to a population that diverged from that leading to modern horses approximately 89–167 kyr BP (Supplementary Figs 10.1–10.3 and Supplementary Table 10.5), providing a maximal boundary for the younger divergence between Przewalski's and domestic horses.

Using quartet alignments and D statistics[24] (Supplementary Information, sections 12.1–12.3) we found no evidence for admixture between the Przewalski's horse and the individual horse breeds investigated in this study using either the donkey or the ancient Thistle Creek genome as out-group (Supplementary Tables 12.1–S12.3). Scanning the Przewalski's horse genome, we also found no long tracts of shared polymorphisms with domestic horses (Supplementary Fig. 12.3), as would be expected if recent admixture occurred after the last wild individual was captured in the 1940s[25]. Rather, we identified long tracts of variation unique to the Przewalski's horse genome, including genes involved in immunity, cytoskeleton, metabolism and the central nervous system that could have been specifically selected in this lineage (Supplementary Information, section 12.6). The average levels of polymorphism present in the Przewalski's horse genome are greater than those observed in the Icelandic, Standardbred and Arabian horse genomes (Supplementary Fig. 5.5 and Supplementary Table 11.10). Thus, unadmixed lineages

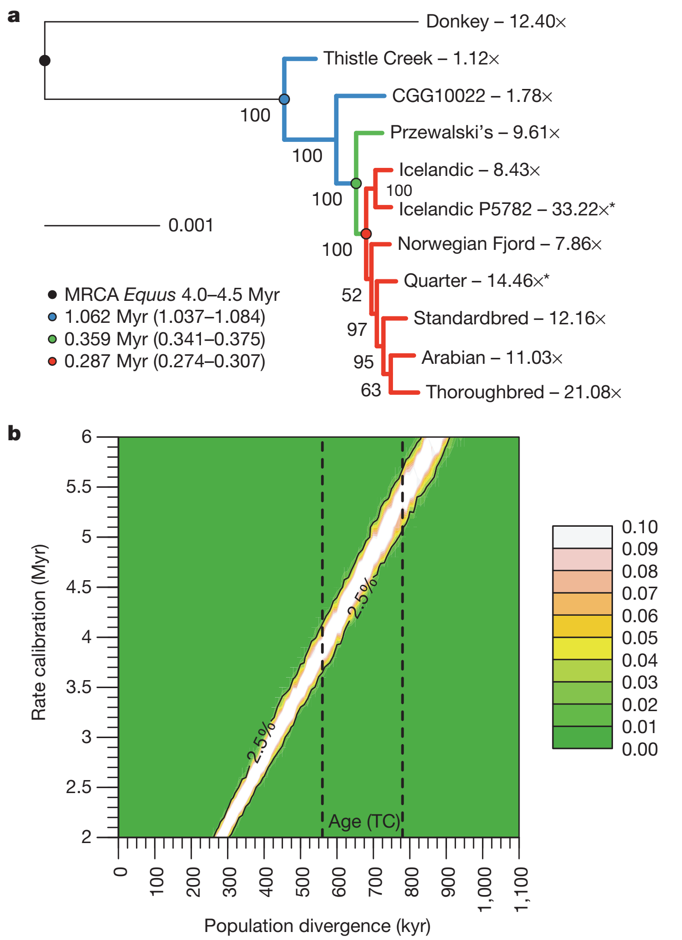

Figure 3 | Horse phylogenetic relationships and population divergence times. a, Maximum likelihood phylogenetic inference. We performed a super-matrix analysis of 5,359 coding genes (Supplementary Information, section 8.3a, 100 bootstrap pseudo-replicates) and estimated the average age for the main nodes (r8s semi-parametric penalized likelihood (PL) method, Supplementary Information, section 8.3c; see Supplementary Table 8.3 for other analyses). Asterisk indicates previously published horse genomes. **b**, Population divergence times. We used ABC to recover a posterior distribution for the time when two horse populations split over a full range of possible mutation rate calibrations (Supplementary Information, section 10.1). The first population included the Thistle Creek horse; the second consisted of modern domestic horses. A conservative age range for the Thistle Creek horse is reported between the dashed lines (560–780 kyr).

are still present in the endangered Przewalski's horse population, with levels of allelic diversity that can support long-term survival of captive breeding stocks despite descending from only 13–14 wild individuals[25].

The sequencing of the horse reference genome showed increased paralogous expansion rates in horses compared to humans and bovines for certain functionally important gene families[26] (Supplementary Information, section 5.1c). Our data set revealed that a limited fraction of horse paralogues (1.7%, representing 258 paralogues) showed no hits among donkey reads, suggesting that most horse paralogues expanded before the origin of the genus *Equus* some 4.0–4.5 Myr BP. Among these 258 regions, 11 L1 retrotransposons and one copy of a keratin gene are absent from the ancient Thistle Creek horse genome but present in the 43 kyr horse and modern horses (Supplementary Table 5.3), suggesting an expansion before their MRCA some 500–626 kyr BP (Supplementary Table 8.3). Similarly, 44 L1-retrotransposon paralogues were found only in modern horse genomes (Supplementary Table 5.4), indicating that expansion of L1 retrotransposons has remained active since then.

Finally, we identified loci potentially selected in modern horses (Supplementary Figs 11.1–11.2), focusing on regions showing unusual densities of derived mutations (Supplementary Information, section 11.1). We caution that local variations in mutation and recombination rates, as well as misalignments, may result in similar signatures at neutrally evolving regions. Functional clustering analyses revealed significant enrichment for immunity-related and olfactory receptor genes (Supplementary Table 11.4), two categories also enriched for non-synonymous single nucleotide polymorphisms (SNPs) (Supplementary Information, section 5.2d). Additionally, we identified 29 regions showing deviation from neutrality and significant reduction in genetic diversity among modern domestic horses compared to Przewalski's horse (Supplementary Tables 11.8–11.9). Such regions could correspond to loci that have been selected and transmitted to all horse breeds investigated here after divergence from the Przewalski's horse population,

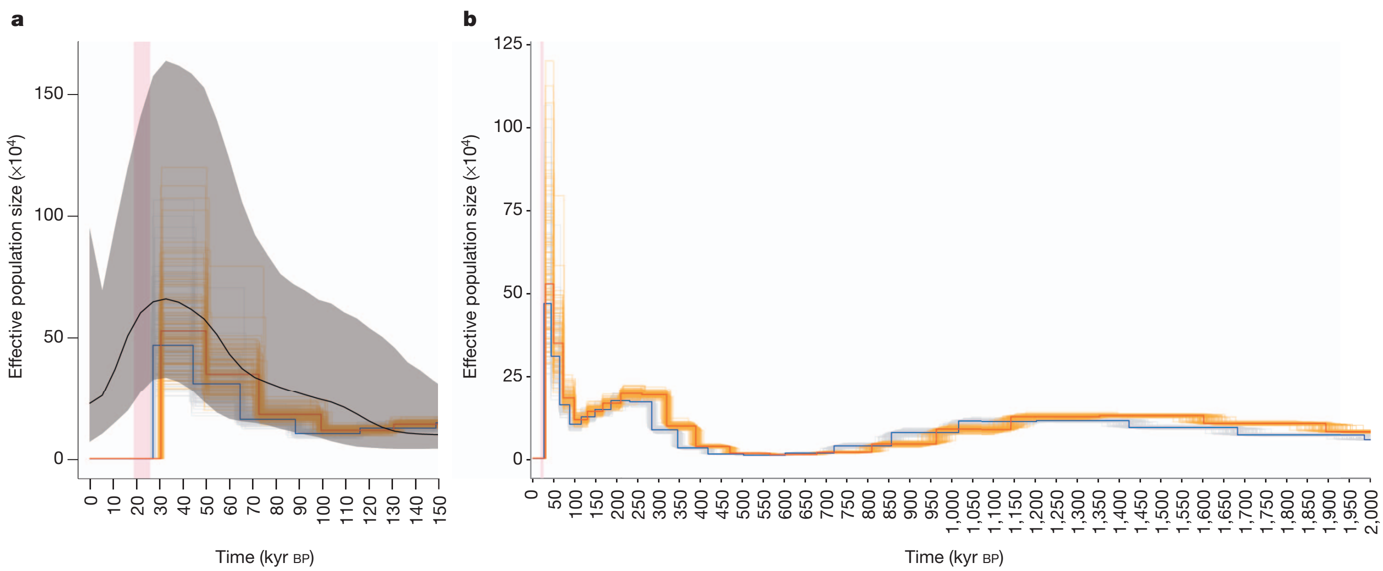

Figure 4 | Horse demographic history. a, Last 150 kyr BP. PSMC based on nuclear data (100 bootstrap pseudo-replicates) and Bayesian skyline inference based on mitochondrial genomes (median, black; 2.5% and 97.5% quantiles, grey) are presented following the methodology described in Supplementary Information, section 9. The Last Glacial Maximum (19–26 kyr BP) is shown in pink. **b**, Last 2 Myr BP. PSMC profiles are scaled using the new calibration values proposed for the MRCA of all living members of the genus *Equus* (4.0 Myr, blue; 4.5 Myr, red), and assuming a generation time of 8 years (for other generation times, see Supplementary Figs 9.4 and 9.5).

possibly related to domestication. These regions include genes for the KIT ligand critical for haematopoiesis, spermatogenesis and melanogenesis, and myopalladin involved in sarcomere organization.

Our study has pushed the timeframe of palaeogenomics back by almost an order of magnitude. This enabled us to readdress a range of questions related to the evolution of *Equus*—a group representing textbook examples of evolutionary processes. The Thistle Creek genome also provided us with direct estimates of the long-term rate of DNA decay[27], revealing that a significant fraction (6.0–13.3%) of short (25-bp) DNA fragments may survive over a million years in the geosphere (Supplementary Fig. 6.42). Thus, procedures maximizing the retrieval of short, but still informative, DNA may provide access to resources previously considered to be much too old. Methods have recently been developed for increasing the sequencing depth of ancient genomes[15] but do not increase the percentage of endogenous sequences retrieved. Overcoming this technical challenge with whole-genome enrichment approaches, and lower sequencing costs, will make retrieval of higher coverage genomes from specimens with low endogenous DNA content practical and economical.

METHODS SUMMARY

Ancient horse extracts and DNA libraries were prepared in facilities designed to analyse ancient DNA following standard procedures[8,12]. Protein sequencing was performed using nanoflow liquid chromatography tandem mass spectrometry[28]. DNA sequencing was performed using Illumina and Helicos sequencing platforms[8,13]. Reads were aligned to the horse reference genome[26] and *de novo* assembled donkey scaffolds using BWA[29]. Maximum-likelihood DNA damage rates were estimated from nucleotide misincorporation patterns. Population divergence times were estimated disregarding transitions to limit the impact of replication of damaged DNA and following ref. 20 with quartet genome alignments instead of trios and implementing approximate Bayesian computation (ABC).

Online Content Any additional Methods, 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 30 October 2012; accepted 30 May 2013.
Published online 26 June 2013.

1. Franzen, J. L. *The Rise of Horses: 55 Million Years of Evolution* (Johns Hopkins Univ. Press, 2010).
2. Froese, D. G., Westgate, J. A., Reyes, A. V., Enkin, R. J. & Preece, S. J. Ancient permafrost and a future, warmer Arctic. *Science* **321,** 1648 (2008).
3. Westgate, J. A. *et al.* Gold Run tephra: A Middle Pleistocene stratigraphic and paleoenvironmental marker across west-central Yukon Territory, Canada. *Can. J. Earth Sci.* **46,** 465–478 (2009).
4. Eisenmann, V. Origins, dispersals, and migrations of *Equus* (Mammalia, Perissofactyla). *Courier Forschungsintitut Senckenberg* **153,** 161–170 (1992).
5. Forsten, A. Mitochondrial-DNA timetable and the evolution of *Equus*: Comparison of molecular and paleontological evidence. *Ann. Zool. Fenn.* **28,** 301–309 (1992).
6. Goto, H. *et al.* A massively parallel sequencing approach uncovers ancient origins and high genetic variability of endangered Przewalski's horses. *Genome Biol. Evol.* **3,** 1096–1106 (2011).
7. Reyes, A. V., Froese, D. G. & Jensen, B. J. Response of permafrost to last interglacial warming: field evidence from non-glaciated Yukon and Alaska. *Quat. Sci. Rev.* **29,** 3256–3274 (2010).
8. Orlando, L. *et al.* True single-molecule DNA sequencing of a Pleistocene horse bone. *Genet. Res.* **21,** 1705–1719 (2011).
9. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362,** 709–715 (1993).
10. Willerslev, E. *et al.* Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science* **317,** 111–114 (2007).
11. Miller, W. *et al.* Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc. Natl Acad. Sci. USA* **109,** E2382–E2390 (2012).
12. Cappellini, E. *et al.* Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins. *J. Proteome Res.* **11,** 917–926 (2012).
13. Ginolhac, A. *et al.* Improving the performance of True Single Molecule Sequencing for ancient DNA. *BMC Genomics* **13,** 177 (2012).
14. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463,** 757–762 (2010).
15. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
16. 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).
17. Vilstrup, J. T. *et al.* Mitochondrial phylogenomics of modern and ancient equids. *PLoS ONE* **8,** e55950 (2013).
18. McFadden, B. J. & Carranza-Castaneda, O. Cranium of *Dinohippus mexicanus* (Mammalia Equidae) from the early Pliocene (latest Hemphillian) of central Mexico and the origin of *Equus. Bull. Florida Museum Nat. History* **43,** 163–185 (2002).
19. Weinstock, J. *et al.* Evolution, systematics, and phylogeography of Pleistocene horses in the new world: a molecular perspective. *PLoS Biol.* **3,** e241 (2005).
20. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
21. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
22. Lorenzen, E. D. *et al.* Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature* **479,** 359–364 (2011).
23. International Union for Conservation of Nature. IUCN Red List of Threatened Species, Version 2010.1, http://www.iucnredlist.org (downloaded 11 March 2010).
24. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060 (2010).
25. Bowling, A. T. *et al.* Genetic variation in Przewalski's horses, with special focus on the last wild caught mare, 231 Orlitza III. *Cytogenet. Genome Res.* **102,** 226–234 (2003).
26. Wade, C. M. *et al.* Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326,** 865–867 (2009).
27. Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. Lond. B* **279,** 4724–4733 (2012).
28. Kelstrup, C. D., Young, C., Lavallee, R., Nielsen, M. L. & Olsen, J. V. Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole orbitrap mass spectrometer. *J. Proteome Res.* **11,** 3487–3497 (2012).
29. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

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

Acknowledgements We thank T. Brand, the laboratory technicians at the Danish National High-throughput DNA Sequencing Centre and the Illumina sequencing platform at SciLifeLab-Uppsala for technical assistance; J. Clausen for help with the donkey samples; S. Rasmussen for computational assistance; J. N. MacLeod and T. Kalbfleisch for discussions involving the re-sequencing of the horse reference genome; and S. Sawyer for providing published ancient horse data. This work was supported by the Danish Council for Independent Research, Natural Sciences (FNU); the Danish National Research Foundation; the Novo Nordisk Foundation; the Lundbeck Foundation (R52-A5062); a Marie-Curie Career Integration grant (FP7 CIG-293845); the National Science Foundation ARC-0909456; National Science Foundation DBI-0906041; the Searle Scholars Program; King Saud University Distinguished Scientist Fellowship Program (DSFP); Natural Science and Engineering Research Council of Canada; the US National Science Foundation DMR-0923096; and a grant RC2 HG005598 from the National Human Genetics Research Institute (NHGRI). A.G. was supported by a Marie-Curie Intra-European Fellowship (FP7 IEF-299176). M.F. was supported by EMBO Long-Term Post-doctoral Fellowship (ALTF 229-2011). A.-S.M. was supported by a fellowship from the Swiss National Science Foundation (SNSF). Mi.S. was supported by the Lundbeck foundation (R82-5062).

Author Contributions L.O. and E.W. initially conceived and headed the project; G.Z. and Ju.W. headed research at BGI; L.O. and E.W. designed the experimental research project set-up, with input from B.S. and R.N.; D.F. and G.D.Z. provided the Thistle Creek specimen, stratigraphic context and morphological information, with input from K.K.; K.H.R., B.S., K.G., D.C.M., D.F.A., K.A.S.A.-R. and M.F.B. provided samples; L.O., J.T.V., Ma.R., M.H., C.M. and J.S. did ancient and modern DNA extractions and constructed Illumina DNA libraries for shotgun sequencing; Ja.W. did the independent replication in Oxford; Ma.S. did ancient DNA extractions and generated target enrichment sequence data; Ji.M. and X.W. did Illumina libraries on donkey extracts; K.M., C.M. and A.S.-O. performed Illumina sequencing for the Middle Pleistocene and the 43-kyr-old horse genomes, the five domestic horse genomes and the Przewalski's horse genome at Copenhagen, with input from Mo.R.; Ji.M. and X.W. performed Illumina sequencing for the Middle Pleistocene and the donkey genomes at BGI; J.F.T. headed true Single DNA Molecule Sequencing of the Middle Pleistocene genome; A.G., B.P. and Mi.S. did the mapping analyses and generated genome alignments, with input from L.O. and A.K.; Jo.V. and T.S.-P. did the metagenomic analyses, with input from A.G., B.P., S.B. and L.O.; Jo.V. and T.S.-P. did the *ab initio* prediction of the donkey genes and the identification of the Y chromosome scaffolds, with input from A.G. and Mi.S.; L.O., A.G. and P.L.F.J. did the damage analyses, with input from I.M.; A.G. did the functional SNP assignment; A.M.V.V. and L.O. did the PCA analyses, with input from O.R.; B.S. did the phylogenetic and Bayesian skyline reconstructions on mitochondrial data; Mi.S. did the phylogenetic and divergence dating based on nuclear data, with input from L.O.; A.G. did the PSMC analyses using data generated by C.J.R. and L.A.; L.O. and A.G. did the population divergence analyses, with input from J.C., R.N. and M.F.; L.O., A.G. and T.K. did the selection scans, with input from A.-S.M. and R.N.; A.A., I.M. and M.F. did the admixture analyses, with input from R.N.; L.O. and A.G. did the analysis of paralogues and structural variation; Ja.V. and A.D. did the amino-acid composition analyses; E.C., C.D.K., D.S., L.J.J. and J.V.O. did the proteomic analyses, with input from M.T.P.G. and A.M.V.V.; L.O. and V.E. performed the morphological analyses, with input from D.F. and G.D.Z.; L.O. and E.W. wrote the manuscript, with critical input from M.H., B.S., Jo.M. and all remaining authors.

Author Information All sequence data have been submitted to Sequence Read Archive under accession number SRA082086 and are available for download, together with final BAM and VCF files, *de novo* donkey scaffolds, and proteomic data at http://geogenetics.ku.dk/publications/middle-pleistocene-omics. 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 L.O. (Lorlando@snm.ku.dk), Ju.W. (wangjun30@gmail.com) or E.W. (ewillerslev@snm.ku.dk).

METHODS

Genome sequencing. All fossil specimens were extracted in facilities designed to analyse ancient DNA using silica-based extraction procedures[30,31] (Supplementary Information, section 2). A total number of 16 ancient horse extracts were built into Illumina libraries (Supplementary Information, section 2) and shotgun-sequenced at the Centre for GeoGenetics (Supplementary Tables 2.3 and 4.9). The full mitochondrial genome of a total number of 16 ancient horse specimens was captured using MYselect in-solution target enrichment kit (Supplementary Information, section 3.3b) following library construction[32], and sequenced at Penn State/UCSC (Supplementary Tables 2.4 and 4.10). The combination of shotgun sequencing and capture-based sequencing performed in those two laboratories resulted in the characterization of 23 novel pseudo-complete ancient horse mitochondrial genomes (Supplementary Table 8.1). Additional sequencing was compatible with the characterization of draft nuclear genomes of two ancient horse specimens (Supplementary Tables 4.9 and 4.11): that of a Middle Pleistocene horse from Thistle Creek (560–780 kyr BP), and that of a Late Pleistocene horse from the Taymyr Peninsula (CGG10022, cal. 42,012–40,094 BC; Supplementary Table 2.3). The Thistle Creek horse draft genome was characterized using Illumina (11,593,288,435 reads, Supplementary Table 3.2; coverage = 0.74×, Supplementary Table 4.11) and Helicos sequence data (654,292,583 reads, Supplementary Table 3.5; coverage = 0.38 ×, Supplementary Table 4.11). Ancient specimens were radiocarbon dated at Belfast 14Chrono facilities (Supplementary Tables 2.3 and 2.4). The Middle Pleistocene Thistle Creek horse bone is associated with infinite radiocarbon dates.

Modern equine genomes from five modern horse breeds (Arabian, Icelandic, Norwegian fjord, Standardbred, Thoroughbred), one Przewalski's horse individual and one domestic donkey were characterized using Illumina paired-end sequencing (Supplementary Information, sections 3.1.b.3–3.1.b.4). DNA was extracted and prepared into libraries (Supplementary Information, section 2.2) in laboratories located in buildings physically separated from ancient DNA laboratory facilities. Modern horse genomes were sequenced at the Danish National High-Throughput DNA Sequencing Centre whereas the donkey genome was characterized at BGI, Shenzen (Supplementary Information, 3.1). Trimmed reads were aligned to the horse reference genome EquCab2.0 (ref. 26), excluding the mitochondrial genome and chrUn, using BWA[29] (Supplementary Information, section 4.2). We generated a draft *de novo* assembly of the donkey genome using de Bruijn graphs as implemented within SOAPdenovo[33] (Supplementary Information, section 4.1.a), built gene models using Augustus[34] and SpyPhy[35] (Supplementary Information, section 4.1.b), and identified candidate scaffolds originating from the X and Y chromosomes (Supplementary Information, sections 4.1.c and 4.1.d). Sequence reads were also aligned against *de novo* assembled donkey scaffolds (Supplementary Information, section 4.2). For all genomes characterized in this study, we estimated that overall error rates were low (Supplementary Information, section 4.4.a), with type-specific error rates inferior to 5.3×10^{-4}, except for ancient genomes where post-mortem DNA damage inflated the GC→AT mis-incorporation rates (Supplementary Table 4.12). Metagenomic assignment of all reads generated from the Thistle Creek horse bone was performed using BWA-sw[36] and mapping against a customized database, which included all bacterial, fungal and viral genomes available (Supplementary Information, section 4.3).

Genomic variation. SNPs were called for modern genomes using the mpileup command from SAMtools (0.1.18)[37] and bcftools, and were subsequently filtered using vcfutils varFilter and stringent quality filter criteria (Supplementary Information, section 5.2). We compared overall SNP variation levels (Supplementary Information, sections 5.2b and 11.2; Supplementary Table 11.10) present in modern horse genomes. We also compared genotypic information extracted from the genomes characterized in this study to that of 362 horse individuals belonging to 14 modern domestic breeds and 9 Przewalski's horses[38]. Genotype and the breed/population of origin were converted into PLINK map and ped formats[39] and further analysed using the software Smartpca of EIGENSOFT 4.0 (ref. 40). PCA plots were generated using R 2.12.2 (ref. 41) (Supplementary Figs 5.6–5.14). Filtered SNPs that passed our quality criteria (Supplementary Information, section 5.2.a) were categorized into a series of functional and structural genomic classes using the Perl script variant_effect_predictor.pl version 2.5 (ref. 42) available at Ensembl and the EquCab2.0 annotation database version 65 (Supplementary Information, section 5.2b). We also screened our genome data for a list of 36 loci that have been associated with known phenotypic defects and/or variants (Supplementary Information, section 5.2e and Supplementary Tables 5.19 and 5.20). We systematically looked in the donkey genome for the presence of genes that have been identified in the horse reference genome as paralogues. This was performed by downloading from Ensembl a list of 15,310 paralogues and extracting genomic coordinates of the 15,171 paralogues that were located on the 31 autosomes and the X chromosome. We next calculated the average depth-of-coverage of these regions using the alignment of donkey reads against the horse reference genome. A total number of 258 paralogues exhibited no hit and were putatively missing from the donkey genome. We further tested for the presence of those paralogues in the different ancient horse genomes characterized here, using a model where observed depth-of-coverage in ancient individual (Illumina data) is a function of the depth-of-coverage observed in a modern horse male individual, local %GC and read length (Supplementary Information, section 5.1c). A similar model was used for identifying segmental duplications in modern equid genomes (Supplementary Information, section 5.1b).

DNA damage. We estimated DNA damage levels in the Thistle Creek horse sample and compared these to the DNA damage levels observed among other Pleistocene horse fossil bones, all associated with more recent ages (Supplementary Tables 2.3 and 2.4). All fossil specimens analysed were permafrost-preserved, limiting environmental-dependent variation in DNA damage rates[43]. DNA fragmentation and nucleotide mis-incorporation patterns were plotted using the mapDamage package[44] (Supplementary Information, section 6.2). We then developed a DNA damage likelihood model after the model presented in ref. 45, with slight modifications, where ancient DNA fragments consist of four non-overlapping regions from 5′ to 3′ ends: (1) a single-stranded overhang; (2) a double-stranded region that extends until a single-strand break is encountered; (3) a double-stranded region that extends 3′ of the single strand break previously mentioned, and; (4) a single stranded overhang (Supplementary Information, section 6.3 and Supplementary Fig. 6.39). All model parameters were estimated using maximum likelihood. Confidence intervals were found by taking each parameter in turn and slowly adjusting that parameter while maximizing the likelihood with respect to all other parameters until finding the points above and below with likelihood 1.92 units below the maximum. Finally, we used the model framework presented in ref. 27 to recover direct estimates of DNA survival rates from next-generation sequence data (Supplementary Information, section 6.4). We restricted our analyses (1) to the distribution of templates showing sizes superior to the modal size category; and (2) to collapsed paired-end reads, as the size of the latter corresponds to the exact size of ancient DNA fragments inserted in the DNA library.

Amino acid and proteomic analyses. A sample of the Middle Pleistocene Thistle Creek horse bone was embedded in Epothin resin under sterile conditions, cut and polished until chemical analysis of the sample surface could be performed with a time-of-flight secondary ion mass spectrometer (TOF-SIMS) instrument (Supplementary Information, section 7). We also performed high-resolution mass spectrometry (MS)-based shotgun proteomics analysis using two fragments from the Middle Pleistocene Thistle Creek horse bone (weighing 86 and 78 mg, respectively) in order to retrieve large-scale molecular information. The overall methodological approach follows the procedure that was previously applied to survey the remains of the bone proteome from three mammoth specimens living approximately 11–43 kyr ago[12], although with significant improvements (Supplementary Information, sections 7.2–7.3). Strict measures to avoid contamination and exclude false-positive results were implemented at every step, allowing to confidently profile 73 ancient bone proteins (from the attribution of 659 unique peptides based on 13,030 spectra). Raw spectrum files were searched on a local workstation using the MaxQuant algorithm version 1.2.2.5 (ref. 46) and the Andromeda peptide search engine[47] against the target/reverse list of horse proteins available from Ensembl (EqCab2.64.pep.all), the IPI v.3.37 human protein database and the common contaminants such as wool keratins and porcine trypsin, downloaded from Uniprot. The spectra were also searched against the Uniprot protein database, taxonomically restricted to chordates, and non-horse peptides were identified and eventually removed. Proteomic data were further compared to similar information already generated from fossil specimens collected in Siberian permafrost and temperate environments. Proteome-wide incidence of deamidation was estimated in relation with protein recovery to further assess the molecular state of preservation of ancient proteins.

Phylogenetic analyses. The CDS of protein-coding genes were selected from the Ensembl website, keeping the transcripts with the most exons in cases where multiple records were found for a single gene. We then extracted corresponding genomic coordinates, filtered for DNA damage/sequencing errors, and aligned each gene using MAFFT G-INS-i ('ginsi')[48,49] (Supplementary Information, section 8.3a). Phylogenetic analysis was carried out using a super-matrix approach. First RAxML v7.3.2[50] was run to generate the parsimony starting trees. The final tree inference was performed using RAxML-Light v1.1.1[51] and one GTRGAMMA model of nucleotide substitutions for each gene partition (codon positions 1 and 2, versus 3). Node support was estimated using 100 bootstrap pseudo-replicates. Bootstrap trees were dated using 'r8s', using the PL method and the Truncated Newton (TN) algorithm, with a smoothing value of 1,000 (ref. 52), or using the Langley–Fitch (LF) method (Supplementary Information, section 8.3.c). The date of the root node was constrained to 4.0–4.5 Myr, the date of CGG10022 was fixed to 43 kyr, and the date of the Thistle Creek specimen was constrained to 560–780 kyr BP. We also performed phylogenetic analyses of whole mitochondrial genomes (Supplementary Information, section 8.1), Y chromosome (Supplementary Information, section 8.2) and a series of topological tests using approximately unbiased tests as implemented in the CONSEL makermt program[53] (Supplementary Information, section 8.3b).

Demographic reconstructions. Past population demographic changes were reconstructed from whole diploid genome information using the pairwise sequentially Markovian coalescent model (PSMC)[21] and excluding sequence data originating from sex chromosomes and scaffolds (Supplementary Information, section 9). For low coverage genomes ($<20\times$), we applied a correction based on an empirical uniform false-negative rate. Three different generation times of 5, 8 and 12 years were considered in agreement with the range of generation times reported in the literature[23,54–56]. Mutation rates were estimated using quartet genome alignments where the donkey was used as out-group (Supplementary Information, section 10.1c). We also reconstructed past horse population demographic changes by means of Bayesian skyline plots using the software BEAST v1.7.2 (refs 57, 58) (Supplementary Information, section 9.2). Complete mitochondrial genomes were aligned and partitioned as described in Supplementary Information, section 8.1b, and a strict clock model was selected. We ran two independent MCMC chains of 50 million iterations each, sampling from the posterior every 5,000 iterations. We discarded the first 10% of each chain as burn-in, and after visual inspection in Tracer v1.5[59] to ensure that the replicate chains had converged on similar values, combined the remainder of the two runs.

Population split. We followed the method presented in ref. 20 to estimate the population divergence date of ancient and modern horses (Supplementary Information, section 10.1). This method was also applied to date the population divergence of Przewalski's horses and domestic horses (Supplementary Information, section 10.2), as both our phylogenetic analyses and admixture tests supported those as two independent populations (Supplementary Information, sections 8.3 and 12). In this method, we focus on heterozygous sites in one of the two populations and randomly sample one of the two possible alleles (ancestral or derived) in the individual belonging to the first population. The number of times a derived allele is sampled (F statistics) can be used to recover a full posterior distribution of the population divergence time using (serial) coalescent simulations and approximate Bayesian computation (ABC) (Supplementary Information, section 10.1). For dating the divergence time between the Przewalski's horse population and domestic breeds, we also performed coalescent simulations using ms[60] assuming different divergence times in order to compute the expected relative occurrences of 4 genotype configurations (Supplementary Information, section 10.2b). We assumed that no gene flow occurred after the population split, in agreement with the absence of detectable levels of admixture. The divergence time was then estimated by minimizing the root mean square deviation (r.m.s.d.) between observed and expected genotype configurations. We minimized the r.m.s.d. using a golden search algorithm. We repeated the minimization from different starting values to ensure convergence.

Selection scans. We used quartet alignments including the donkey as out-group, one ancient horse and two modern horses to scan for genomic regions where the two modern horses shared unusual accumulation of derived alleles (Supplementary Information, section 11.1). We used a sliding window approach on the entire genome, with a window size of 200 kb and calculated an unbiased proxy for selection using the 'delta technique' (see for example ref. 61). We then used an outlier approach to identify candidate loci with a conservative false-positive rate of 0.01. We further retrieved transcript IDs from the different genomic regions identified and performed functional clustering analyses in DAVID[62]. We estimated genetic diversity (theta Watterson) within the Przewalski's horse population and among modern horse breeds using sliding windows of 50 kb. For this, we estimated the population scaled mutation rate and used an empirical Bayes method where we took the uncertainty of the data into account by using genotype likelihoods instead of calling genotypes. We computed the genotype likelihoods assuming a model similar to that of SAMtools version 0.1.18 (ref. 37) (Supplementary Information, section 11.2). Genomic windows showing excessive proportions of segregating sites with regards to species divergence ($>5\%$) or coverage $<90\%$ were discarded. We estimated Tajima's D following the same procedure and identified genomic regions showing minimal Tajima's D values and low genetic diversity among breeds but not in the Przewalski's horse population as a conservative set of gene candidates for positive selection among modern horse breeds. Finally, we scanned modern horse genomes for long homozygosity tracts, which could be indicative of selective sweeps[63]. We used 2-Mb sliding windows and ignored sites showing coverage inferior to 8. This resulted in the identification of 456 outlier regions within 8 modern horse genomes.

Admixture analyses. In order to investigate if there was evidence for gene flow between the Przewalski's horse population and four modern horse domestic breeds (Arabian, Icelandic, Norwegian fjord and Standardbred), we performed ABBA-BABA tests[20,24]. To avoid introducing bias due to differences in sequencing depth we based the tests on data achieved by sampling one allele randomly from each horse at each site. First we used the domestic donkey as out-group, then the Middle Pleistocene Thistle Creek horse. When using the Thistle Creek horse as out-group we removed all sites showing transitions to avoid spurious patterns resulting from nucleotide misincorporations related to post-mortem DNA damage. We estimated the standard error of the test statistic using 'delete-m Jackknife for unequal m' with 10-Mb blocks[64] (Supplementary Information, section 12.1). We also scanned genome alignments to record the proportion of shared SNPs between Przewalski's horse and each horse breed (Supplementary Information, section 12.6), a proxy for recent admixture events that are expected to result in the introgression of alleles from the admixer to the admixed genome and long tracts of shared polymorphisms. Finally, we compared our Przewalski's horse individual to other individuals with different levels of admixture in their pedigree. We extracted genotype information from the Przewalski's horse genome for SNP coordinates already genotyped across 9 Przewalski horse individuals[38]. Genotypic information from two Mongolian horses was added as out-group. We next selected the best model of nucleotide substitution using modellgenerator v0.85 (ref. 65) and performed maximum likelihood phylogenetic analyses using PhyML 3.0 (ref. 66) (Supplementary Information, section 12.5). We further confirmed the phylogenetic position of our Przewalski's horse individual together with Rosa (KB3838), Basil (KB7413) and Roland (KB3063), three individuals for which no admixture with domestic horses could be detected in previous studies[25] by means of Approximate-Unbiased (AU) and Shimodeira-Hasegawa (SH-) tests, as implemented in CONSEL[53].

Morphological analyses. We measured the metapodial of Thistle Creek early Middle Pleistocene bone for 6 dimensions, despite incomplete preservation of its distal end (Supplementary Information, section 1.2). These measurements were compared to 30 metatarsals of *E. lambei*, 9 metatarsals of *E.* cf. *scotti* of Klondike, Central Yukon, Canada (Supplementary Information, section 1.2) and to extant horses (Supplementary Information, section 1.3). Comparisons were made using Simpson's ratio diagrams that provide a standard and accurate comparison of both size and shape, for a single bone or a group of bones (Supplementary Figs 1.2 and 1.3). We also measured taxonomically informative morphometric features on the skull and post-cranial complete skeleton of the modern Przewalski's horse specimen that was genome sequenced. We compared those to a collection of horse measurements available for horses, filtering for specimens of similar age and using principal component analyses (Supplementary Information, section 1.4).

30. Orlando, L. *et al.* Revising the recent evolutionary history of equids using ancient DNA. *Proc. Natl Acad. Sci. USA* **106,** 21754–21759 (2009).
31. Rohland, N. & Hofreiter, M. Ancient DNA extraction from bones and teeth. *Nature Protocols* **2,** 1756–1762 (2007).
32. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **6,** http://dx.doi.org/10.1101/pdb.prot5448 (2010).
33. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1,** 18 (2012).
34. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32,** W309–W312 (2004).
35. Carlton, J. M. *et al.* Draft genome sequence of the sexually transmitted pathogen *Trichonomas vaginalis*. *Science* **315,** 207–212 (2007).
36. Li, H. & Durbin, R. R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
37. Li, H. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
38. McCue, M. E. *et al.* A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet.* **8,** e1002451 (2012).
39. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
40. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
41. R Development Core Team. A language and environment for statistical computing, http://www.R-project.org (R Foundation for Statistical Computing, 2011).
42. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).
43. Smith, C. I., Chamberlain, A. T., Riley, M. S., Stringer, C. & Collins, M. J. The thermal history of human fossils and the likelihood of successful DNA amplification. *J. Hum. Evol.* **45,** 203–217 (2003).
44. Ginolhac, A., Rasmussen, M., Gilbert, T. M., Willerslev, E. & Orlando, L. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27,** 2153–2155 (2011).
45. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA* **104,** 14616–14621 (2007).
46. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnol.* **26,** 1367–1372 (2008).
47. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805 (2011).

48. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–3066 (2002).
49. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9,** 286–298 (2008).
50. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
51. Stamatakis, A. *et al.* RAxML-Light: a tool for computing Terabyte phylogenies. *Bioinformatics* **28,** 2064–2066 (2012).
52. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19,** 301–302 (2003).
53. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17,** 1246–1247 (2001).
54. Lippold, S., Matzke, N. J., Reissmann, M. & Hofreiter, M. Whole mitochondrial genome sequencing of domestic horses reveals incorporation of extensive wild horse diversity during domestication. *BMC Evol. Biol.* **11,** 328 (2011).
55. Achilli, A. *et al.* Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proc. Natl Acad. Sci. USA* **109,** 2449–2454 (2012).
56. Warmuth, V. *et al.* Reconstructing the origin and spread of horse domestication in the Eurasian steppe. *Proc. Natl Acad. Sci. USA* **109,** 8202–8206 (2012).
57. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).
58. 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).
59. Rambaut, A. & Drummond, A. J. Tracer v1. 5, http://beast.bio.ed.ac.uk/Tracer (2009).
60. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).
61. Zhang, Z. *Computational Molecular Evolution* (Oxford Univ. Press, 2006).
62. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols* **4,** 44–57 (2009).
63. Nielsen, R. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39,** 197–218 (2005).
64. Busing, F. M. T. A., Meijer, E. & Van Der Leeden, R. Delete-*m* Jackknife for Unequal *m. Stat. Comput.* **9,** 3–8 (1999).
65. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McInerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6,** 29 (2006).
66. Guindon, S. *et al.* New algorithms and methods to estimate Maximum-Likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).