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

Ludovic Orlando1*, Aurélien Ginolhac1*, Guojie Zhang2*, Duane Froese3, Anders Albrechtsen4, Mathias Stiller5, Mikkel Schubert1, Enrico Cappellini1, Bent Petersen6, Ida Moltke7,8, Philip L. F. Johnson9, Matteo Fumagalli10, Julia T. Vilstrup1, Maanasa Raghavan1, Thorfinn Korneliusen1, Anna-Sapfo Malaspina1, Josef Vogt11, Danilam Szklarczyk12,13, Christian D. Kelstrup1, Jakob Vintner1,14, Andrei Dolocan15, Jesper Stenderup1,16, Amhed M. V. Velazquez1,17, James Cahill18, Morten Rasmussen1, Xiaoli Wang19, Juemeng Min20, Grant D. Zazula1,19, Andaine Seguin-Orlando1,1, Cecilie Mortensen1,19, Kim Magnussen1,19, John F. Thompson19, John Mundy26, Anders Krogh1,4, M. Thomas P. Gilbert1, Kurt Kjær1, Thomas Sicheritz-Ponten6,23, Lars Juul Jensen10, Michael Hofreiter27, Rasmus Nielsen28, Beth Shapiro5, Jun Wang2,26,29,30 & Eske Willerslev1

MNHN, CP 38, 8, rue Buffon, 75005 Paris, France. Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, SE-751 23 Uppsala, Sweden. Baker Institute 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 Equus3. 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 population1. 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 BP3 (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, during a past interglacial as warm or warmer than present1, 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, age2, and infinite radiocarbon dates2.

Theoretical and empirical evidence 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 BP11. 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 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 ~1.12X genome coverage. We based the size distribution of ancient DNA templates on collapsed Illumina
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 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 or improved 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 and Denisovan specimens, 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 cytosome denaturation 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 (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 (1.8X coverage), a modern donkey (16X; Supplementary Fig. 4.1), 5 modern domestic horses (Arabian, Icelandic, Norwegian fjord, Standardbred and Thoroughbred; 7.9X–21.1X) and one modern Przewalski’s horse (9.6X; 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 (Supplementary Information, section 9.1d), the date of the last common ancestor that 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. We additionally sequenced genomes of a 43-kyr-old (pre-domestication) horse (1.8X coverage), a modern donkey (16X; Supplementary Fig. 4.1), 5 modern domestic horses (Arabian, Icelandic, Norwegian fjord, Standardbred and Thoroughbred; 7.9X–21.1X) and one modern Przewalski’s horse (9.6X; 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 (Supplementary Information, section 9.1d), the date of the last common ancestor that gave rise to extant horses versus donkeys, asses and zebras remains heavily debated. Proposed dates extend as early as 4.2–4.5 Myr BP on the basis of palaeontological estimates to 6.0 Myr BP according to molecular analyses. 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 (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 and the oldest palaeontological records for the monodactyle Plesippus simplicidens, which some 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 (PSCM) approach 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. 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 that climatic changes are major demographic drivers for horse populations. PSCM 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...
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 BP4.

We next investigated whether Przewalski’s horse indeed represents the last survivor of wild horses. Native to the Mongolian steppe, this horse was listed as extinct in the wild (IUCN red list25) but has been reasigned 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 of domestic horses. The MRCA of 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 statistics24 (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 1940s25. Rather, we identified long tracts of variation 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 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).

Figure 3 | Horse phylogenetic relationships and population divergence times. a. Maximum likelihood phylogenetic inference. We performed a supermatrix analysis of 5,359 coding genes (Supplementary Information, section 8.3a, 100 bootstrap pseudo-replicates) and estimated the average age for the main nodes (88s 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.

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 (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,
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 16 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 12, 16. Protein sequencing was performed using nanolow liquid chromatography tandem mass spectrometry 28. DNA sequencing was performed using Illumina and Helicos sequencing platforms 11, 23. Reads were aligned to the horse reference genome 29 and de novo assembled donkey scaffolds using BWA 29. Maximum-likelihood DNA damage rates were estimated performed using nanoflow liquid chromatography tandem mass spectrometry 28.


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. H. C. Gamba 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 Mac-Curie Career Integration grant (FP7 CIG-2938345); the National Science Foundation ARC-0909456; National Science Foundation DBI-0906401; 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-0923906, and a grant (522 HGG00598) from the National Human Genome 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 (SNF). M.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 J.J.O. 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. and M.F.B. provided samples; L.O., J.T.V., M.A.R., T.J. and J.S. did the modern DNA extraction; B.P. and Mi.S. did the Illumina DNA libraries for shotgun sequencing; J.A. did the independent replication in Oxford; M.A.S. did ancient DNA extractions and generated target enrichment sequence data; J.M. and X.W. did Illumina libraries on donkey extracts; K.M., C.M. and A.S.-O. performed population genetics analyses of the Middle Pleistocene and the donkey genomes at BGI; J.F.T. did the Middle Pleistocene and the donkey genomes at BGI; J.F.T. headed true Single DNA Molecule sequencing of the Middle Pleistocene genome; B.P. and M.L.S. did the mapping analyses and generated genome alignments, with input from L.O. and A.K.; J.o.V. and T.S.-P. did the metagenomic analyses, with input from A.G., B.P., S.B. and L.O.; J.o.V. and T.S.-P. did the “ab initio” prediction of the donkey genes and the identification of the Y chromosome scaffold, with input from A.G. and M.S. L.O., A.G. and L.P.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; M.A.S. did the phylogenetic and diversification dating based on modern nuclear data, with input from L.O.; A.A. performed the genome 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; J.A.V. and A.G. 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 C.D.K.; A.M.V.V.; L.O. and T.K. did the amino-acid composition analyses; A.M.V.V. did the population divergence dating based on nuclear data, with input from L.O.; A.G. did the PSMC Bayesian skyline reconstructions on mitochondrial data; Mi.S. did the phylogenetic analyses of the Middle Pleistocene and the donkey genomes at BGI; M.A.R., L.O. and A.G. did the protein prediction of the donkey genes and the identification of the Y chromosome scaffolds; with input from A.G. and M.S. L.O., A.G. and L.P.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; M.A.S. did the phylogenetic and diversification dating based on modern nuclear data, with input from L.O.; A.A. performed the genome 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; J.A.V. and A.G. 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 SR608026 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 declared 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. (Loranbo@snm.ku.dk), Ju.W. (wunjun30@gmail.com) or E.W. (willerslev@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 MSelect 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 Bellat 14C (chronosome 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, Shenzhen (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\(^ {33}\) (Supplementary Information, section 4.2). We generated a draft de novo assembly of the donkey genome using de Bruijn graphs as implemented within SOAPdenovo\(^ {34}\) (Supplementary Information, section 4.1.a), built gene models using Augustus\(^ {35}\) and SpypHy\(^ {36}\) (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^-9, 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\(^ {37}\) 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)\(^ {38}\) 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\(^ {39}\). Genotype and the breed/origin of origin were converted into PLINK map and ped formats\(^ {40}\) and further analysed using the software Smartpa 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.pm 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 5,310 paralogous genes, with matching 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 5.1.c). 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 also performed phylogenetic analyses of whole mitochondrial DNA. The CDS of protein-coding genes were selected from Ensembl (EquCab2.64, 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 Ensembl (EquCab2.64, 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.

**LETTER**
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 makernt program23 (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)24 and excluding sequence data originating from sex chromosomes and scaffolds (Supplementary Information, section 9). For low coverage genomes (<2×), we applied a correction based on an empirical and 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 literature25–27. 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 v1.528 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 hybrid and pure 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 ms29 assuming different divergence times in order to compute the expected relative frequencies of 4 genoe 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 selective pressures 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 sweeps30. We used 2-Mb sliding windows and ignored sites showing average 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, Icelandich, Norwegian fjord and Standardbred), we performed ABBA-BABA tests31,24. 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-M blocks31 (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 12.2), 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 individuals32. Genotypic information from two Mongolian horses was added as out-group. We next selected the best model of nucleotide substitution using modelgenerator 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 (KB3383), Basil (KB7413) and Roland (KB3063), three individuals for which no admixture with domestic horse breeds could be detected in previous studies33 by means of Approximate-Unbiased (AU) and Shimodaira-Hasegawa (SH-) tests, as implemented in CONSEL34.

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 L. laimei, 9 metatarsals of E. cf. scotti of Klonidike, 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 species35. We genome sequenced 8YSCs for 6 genome sequenced in this study. 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).