Evolutionary genomics and conservation of the endangered Przewalski's horse

Clio Der Sarkissian[1], Luca Ermini[1], Mikkel Schubert[1], Melinda A. Yang[2], Pablo Librado[1], Matteo Fumagalli[3], Hákon Jónsson[1], Gila Kahila Bar-Gal[4], Anders Albrechtsen[5], Filipe G. Vieira[1], Bent Petersen[6], Aurélien Ginolhac[1], Andaine Seguin-Orlando[7], Kim Magnussen[7], Antoine Fages[1], Cristina Gamba[1], Belen Lorente[8], Sagi Polani[4], Cynthia Steiner[9], Markus Neuditschko[10], Vidhya Jagannathan[11], Claudia Feh[12], Charles L. Greenblatt[13], Arne Ludwig[14], Natalia Abramson[15], Waltraut Zimmermann[16], Renate Schafberg[17], Alexei Tikhonov[15,18], Thomas Sicheritz-Ponten[6], Eske Willerslev[1], Tomas Marques-Bonet[8], Oliver A. Ryder[9], Molly McCue[19], Stefan Rieder[9], Tosso Leeb[10], Montgomery Slatkin[2], Ludovic Orlando[1,20].

[1]Centre for GeoGenetics - Natural History Museum of Denmark; University of Copenhagen; Copenhagen, 1350K; Denmark [2]Department of Integrative Biology; University of California; Berkeley, CA, 94720-3140; USA [3]UCL Genetics Institute Department of Genetics - Evolution and Environment; University College London; London, WC1E 6BT; UK [4]The Koret School of Veterinary Medicine - The Robert H. Smith Faculty of Agriculture, Food and Environment; The Hebrew University; Rehovot, 76100; Israel [5]The Bioinformatics Centre - Department of Biology; University of Copenhagen; Copenhagen, 2200N; Denmark [6]Center for Biological Sequence Analysis - Department of Systems Biology; Technical University of Denmark; Lyngby, 2800; Denmark [7]National High-Throughput DNA Sequencing Centre; University of Copenhagen; Copenhagen, 1353K; Denmark [8]ICREA at the Institut de Biologia Evolutiva (University Pompeu Fabra); Barcelona, 08003; Spain [9]San Diego Zoo Institute for Conservation Research; Escondido, CA 92027; USA [10]Agroscope; Swiss National Stud Farm; Avenches, 1580; Switzerland [11]Institute of Genetics - Vetsuisse Faculty; University of Bern, Bern, 3001; Switzerland [12]Station Biologique de la Tour du Valat; Arles, 13200; France [13]Department of Microbiology and Molecular Genetics; The Hebrew University-Hadassah Medical School; Jerusalem, 91120; Israel [14]Leibniz Institute for Zoo and Wildlife Research; Department of Evolutionary Genetics; Berlin, 10315; Germany [15]Zoological Institute of the Russian Academy of Sciences; Saint-Petersburg, 199034; Russia [16]EAZA Equid TAG; Köln, 50668; Germany [17]Martin-Luther-University Halle-Wittenberg; Museum of domesticated animals "Julius Kühn"; Halle, 06108; Germany [18]Institute of Applied Ecology of the North; North-Eastern Federal University; Yakutsk, 677980; Russia [19]College of Veterinary Medicine; University of Minnesota; St Paul, MN 55108; USA [20] Université de Toulouse, UPS, CNRS UMR 5288, 37 allées Jules Guesde, 31000 Toulouse, France.

Corresponding author: Dr. Ludovic Orlando, Lorlando@snm.ku.dk.

SUMMARY

Przewalski's horses (PHs, *Equus ferus ssp. przewalskii*) were discovered in the Asian steppes in the 1870s and represent the last remaining true wild horses. PHs became extinct in the wild in the 1960s, but survived in captivity thanks to major conservation efforts. The current population is still endangered, counting 2,109 individuals, with a quarter in Chinese and Mongolian reintroduction reserves [1]. These horses descend from a founding population of twelve wild-caught PHs, and possibly up to four domesticated individuals [2–4]. With a stocky build, an erect mane, stripped and short legs, they are phenotypically and behaviorally distinct from domesticated horses (DHs, *Equus caballus*). Here, we sequenced the complete genomes of eleven PHs representing all founding lineages, and five historical specimens dated to 1878-1929 CE, including the Holotype. These were compared to the hitherto most extensive genome dataset characterized for horses, comprising 21 new genomes. We found that loci showing the most genetic differentiation with DHs were enriched in genes involved in metabolism, cardiac disorders, muscle contraction, reproduction, behavior and signaling pathways. We also show that DH and PH populations split ~45,000 years ago and have remained connected by gene-flow thereafter. Finally, we monitor the genomic impact of ~110 years of captivity, revealing reduced heterozygosity, increased inbreeding, and variable introgression of domestic alleles, ranging from non-detectable to as much as 31.1%. This, together with the identification of ancestry informative markers and corrections to the International Studbook, establishes a framework for evaluating the persistence of genetic variation in future reintroduced populations.

RESULTS AND DISCUSSION

Despite their additional chromosomal pair (2n=66 vs 2n=64, [5]), PHs can successfully reproduce with DHs, resulting in fully viable and fertile offspring. The extent to which those horses admixed in the past is, however, contentious. Mitochondrial DNA (mtDNA) [6–8], the X- [9–11] and Y-chromosomes [11–13], and a limited number of autosomal SNPs [14] have supported admixture, with PHs appearing within the genetic variability of DHs, in line with possible domestic contributions listed in the International Studbook [2–4]. Conversely, ~54,000 SNPs [15] and one complete PH genome [16] indicated that PH and DH populations probably diverged ~38,000-72,000 years ago [19], with no sign of admixture [16].

To elucidate the evolutionary history of PHs, we used the Illumina paired-end technology and generated whole-genome sequences of 33 living horses, including eleven PHs representing all pedigree founding lineages (18.0-23.4× average depth-of-coverage), one F1 PH×DH hybrid (22.8×), and 21 horses from five domestic breeds (7.7-22.6×). These complement an existing dataset of seven DH and three PH genomes [16, 17]. We also sequenced the genomes of five historical specimens: two PH pedigree founders from the early 1900s (SB11/Ewld1_Bijsk1 and SB12/Ewld2_Bijsk2, each at 0.9×), one hybrid (SB56/Ewld5_Theodor, 4.3×), as well as the PH Holotype (1.4×) and one Paratype specimen (3.7×), both from the late 19th century (Table S1). DNA fragmentation and nucleotide mis-incorporation patterns, indicative of post-mortem damage [18-19], supported the authenticity of our ancient genomes (Figure S1). We limited the impact of nucleotide mis-incorporations on downstream analyses by enzymatically treating ancient DNA extracts [20], and/or by downscaling base quality scores at damaged positions [14]. The genomes of the Holotype, Paratype and SB11/Ewld1_Bijsk1 specimens showed error rates lower than those of the other historical horses (0.099-0.258% vs 0.463-0.622% errors/base), but similar to those of Late Pleistocene horses (0.103-0.207% [21]). In comparison, error rates in modern horses were 0.018-0.040% in DHs and 0.036-0.051% in PHs (Table S1).

Phylogenetic analyses identified three main mtDNA lineages and two Y-chromosome haplotypes in PHs. PHs represent a paraphyletic assemblage within DHs in the mtDNA tree [6–8] and exhibit Y-chromosome haplotypes not found in any other domesticated stallion, nor in a ~16,000 years-old horse [21] (Figure S2A-B). The maternal and paternal lineages are in agreement with the recorded PH pedigree, except for SB528 and SB285, respectively. An exome-based maximum-likelihood (ML) phylogeny revealed DHs and PHs as two reciprocally monophyletic clades distinct from Late Pleistocene horses (Figure 1A) [21], in accordance with the population tree reconstructed in TreeMix using 198,932 SNPs (Figure S2C). These results are consistent with Pairwise Sequential Markovian Coalescent (PSMC) demographic trajectories ([21]; Figure 1B), indicating an early divergence (~150,000 years ago) for the population of Late Pleistocene horses, and synchronized DH and PH profiles until their most recent history. This also reflects the clusters identified in the STRUCTURE analyses of twelve autosomal STRs (Figure S2D) and principal component analyses based on genome-wide SNP calls or genotype-likelihoods (Figure 1C-S2E-F). The latter show no overlap for historical and modern PHs

along the first principal component, potentially as a consequence of captivity. Altogether, our results support the two Late Pleistocene horses analyzed, DHs and PHs as three populations that diverged ~150,000 years ago [16, 21]. The discrepancy between the phylogenies inferred from autosomes, the Y-chromosome and mtDNA could reflect the recruitment of only few stallions during the domestication process, the co-segregation of mtDNA haplogroups in the ancestral populations of DHs and PHs, and/or maternally-driven gene-flow post-divergence [10–12, 22, 23].

We then aimed at identifying the genetic variants underlying the striking phenotypical differences between DHs and PHs. We found 509 copy-number variants (CNVs) in PHs, encompassing a total number of 1.65 Mb, 219 kb of which previously reported in [24]. These contained 233 genes including genes encoding proteins associated with glycol-sphingolipid metabolic process, scarring, decreased corneal thickness, as well as the sulfuric ester hydrolase activity, which is key for the bone and cartilage matrix composition [25]. We also extracted 101 exonic ancestry informative markers for PHs, of which 53 represent non-synonymous mutations in genes involved in the cardiac activity (*CACNA1D*, *ITGA10*), protein digestion and absorption (*COL18A1*, *COL15A1*); disorders in muscles, ligaments and the tissues surrounding muscles, blood vessels and nerves (*PALMD*, *MCPH1*); diseases of the sebaceous glands (*PALMD*, *MCPH1*); and musculoskeletal and craniofacial abnormalities (*MCPH1*, *SETBP1*) (Table S2). We further sought for regions maximizing the genetic differentiation between DHs and PHs, using the F_{ST} fixation index. These contained 874 genes showing significant enrichment in the kit receptor, the inhibitor of DNA binding and the chemokine signaling pathways, in the glycogene and glycerophospholipid metabolisms, and in pathways involved in striated muscle contraction, heart and metabolic diseases, and mood disorders (Table S2). We finally used SweeD [26] to identify 76 protein-coding candidates for selective sweeps within PHs, including *MCR2*, which encodes one receptor of the ACTH stress hormone. Candidate genes showed enrichment for energetic and metabolic pathways, the cardiovascular system, muscular contraction and immunity. Two genes, *CACNA1D* and *PLA2G1B*, supported a significant enrichment for the gonadotropin-releasing hormone signaling pathway, which is essential in sexual behavior and aggressiveness [27] possibly in relation to the temper of PHs (Table S2). Overall, our analyses unveil some potential drivers of the marked anatomical, physiological and behavioral differences between DHs and PHs, some of which might have been enhanced during domestication.

We next reconstructed the past population history of DHs and PHs. Using dadi [28], our best-fit comparison consistently supported models involving asymmetrical gene-flow after divergence (~35,000-54,000 years ago, in agreement with [16, 21]), almost entirely from PHs into DHs (Figure 2A; Table S3). We also noticed that the fit could be substantially improved when considering that migration stopped within recent times, i.e. as early as the last 200 years, suggesting genetic restocking from wild animals was likely common until recently (Figure 2B; Table S3). Using genome projections [29], the best model was recovered by accounting for a recent 4% migration pulse from DHs and two epochs with asymmetrical gene-flow, first from DHs into PHs, then reversed (Figure 2C; Table S3). The best fit was obtained for migration changing during the Last Glacial Maximum (~23,000 years ago; Figure 2D), when rates decreased ~31-160-fold. This

change in the direction and magnitude of migration could reflect a combination of sexual behavior and demographic factors, as the detected contribution of DHs was higher when populations were much larger (Figure 1B). Altogether, our analyses suggest connection by gene-flow after divergence and prior to horse domestication.

We then evaluated the genetic impact of the bottleneck associated with ~110 years of captivity in PHs. Their average heterozygosity (0.39-0.59 heterozygous sites per kilobase) was comparable to that in other endangered mammalian species (e.g., 0.12-0.65 in [30, 31]), but lower than in DHs (0.40-0.98) and historical PHs (0.74-2.35; Mann-Whitney test p-values$\leq 1.88 \times 10^{-3}$; Figure 3A). Another consequence of the bottleneck is the higher inbreeding coverage, i.e., the proportion of mostly homozygous blocks [21, 32], in PHs (0.052-0.388) than in modern DHs (0.006-0.285; two-sample t-test, p-value$=6.81 \times 10^{-3}$). Additionally, the inbreeding coefficient of historical PHs, as estimated using a novel method accommodating data uncertainty (FHMM, Supplemental Experimental Procedures), was lower than in modern PHs (p-value$=4.43 \times 10^{-2}$) (Figure 3B; Table S1). Altogether, this demonstrates that captivity has significantly impacted the genetic diversity of PHs, reducing heterozygosity and increasing inbreeding.

We investigated admixture with DHs as another potential consequence of captivity. Using NGSAdmix [33], eight clusters recapitulated DH breeds and their differentiation from PHs (Figures 4A-S3A). Although inconsistent with the pedigree describing the historical specimen SB56/Ewld5_Theodor as the F1 offspring of pure PH (SB11/Ewld1_Bjisk1) and DH (Emgl1) parents, this horse displayed, respectively, 79% and 21% of PH and DH components, providing clear evidence of introgression into the early descent of the PH founders.

We sought further admixture evidence using the f_3- statistics in (PH,DH;PH) tests [34] and found significant evidence of DH ancestry within modern PHs (Table S4). This was confirmed using D-statistics [35] (Figures S3B-D), as topologies in the form of (Outgroup,(DH,(PH,PH))) were rejected in a large proportion of the tests. Taking advantage of the genome sequence of the Paratype specimen and (Outgroup,(DH,(PH,Paratype))) trees, we identified three groups of PHs with increasing levels of domestic ancestry (Figure 4B; Table S4): i) SB274, SB293, SB339 and SB533 were found to be no closer to DHs than the Paratype, ii) SB281, SB524, SB528, SB615 and SB966 showed a significant genetic contribution from all the DHs tested (representing 10.5-13.7% of their genome), iii) whereas SB159, SB285, Prz_D2630, Prz_D2631 and Prz_Przewalski showed a significant genetic contribution (2.5-4.9%) from a minority of the DHs tested.

We next mapped admixture blocks along the genome using LAMP [36] and the whole panel of DH and modern PH genomes (except Prz_Przewalski showing coverage artifacts confounding LAMP). A first group (SB281, SB524, SB528, SB615, and SB966), showed LAMP-admixture proportions of 23.9%-31.1% with introgression tracts unevenly scattered across the genome. The remaining PHs exhibited 10.4-19.2% admixture, corresponding to shorter tracts and earlier admixture events (Figure 4C). Interestingly, we found a derived allele associated with increasing wither height at *ZFAT* in SB524, SB528,

and SB966, in a region likely introgressed from DHs (Table S5). Overall, these results confirm the DH introgression depicted in the International Studbook, with individuals tested for the A-line (SB159, SB274, SB293, and SB533) considered as the pedigree's purest line, virtually devoid of admixture [37]. We should, however, caution that, despite being significantly correlated (Pearson's correlation coefficients≥0.98, p-values≤$7.94x10^{-6}$), admixture coefficients estimated using LAMP and D-statistics were likely over-estimated. The DH ancestry for the KB7903 hybrid is indeed ~62% in LAMP, instead of the expected ~50%, and the higher error rate of the Paratype genome inflates allele sharing, hence, D-statistics, between modern PHs and DHs.

In addition to provide the most extensive genomic resource for horses classified as "endangered" on the IUCN Red List [1], our study fulfills the goal of the IUCN/SSC Equid Specialist Group conservation program [38] by identifying hybrids, AIMs, and candidate alleles maximizing the adaptation of PHs to their native environment. In the future, cost-effective genome-wide analyses targeting these markers should help minimize the contribution of domestic ancestry and limit loss of genetic diversity in a closed breeding program, with the ultimate objective of maintaining a viable population, once reintroduced in the wild.

EXPERIMENTAL PROCEDURES

Samples and methods are described in Supplemental Experimental procedures. The sequencing data are available from the European Nucleotide Archive, project accession no. PRJEB10098.

AUTHOR CONTRIBUTIONS

W.Z., O.A.R., and L.O. conceived the project, and L.O. designed the study. N.A., R.S., A.T., O.A.R., M.M., S.R., and T.L. provided samples. C.D., L.E., G.K.B, S.P., C.S., M.N., V.J., M.M., and L.O. performed laboratory analyses. C.D., L.E., M.Sc., M.A.Y., P.L., M.F., H.J., G.K.B, A.A., F.G.V., B.P., A.G., A.F., B.L., T.M.B., M.Sl. and L.O. analyzed genomic data. G.K.B, A.S.O., K.M., C.F., C.L.G., A.L., T.S.P., E.W., T.M.B., O.A.R., M.M., S.R., T.L. and L.O. provided reagents and material. L.O. wrote the paper with input from all co-authors. C.D., P.L., M.Sc., M.F., M.A.Y. and L.O. wrote the supplementary information, with input from C.G..

ACKNOWLEDGMENTS

We thank the staff of the Danish National High-Throughput DNA Sequencing Center for technical assistance. This work was supported by the Danish Council for Independent Research, Natural Sciences (FNU-4002-00152B); the Danish National Research Foundation (DNFR94); the Villum Fonden Blokstipendium (2014); the Lundbeck Foundation (R52-A5062); the Israel Science Foundation (1365/10); the German Research Council (DFG-LU852/7-4); the NIH (R01-GM40282); the Caesar Kleberg Foundation for Wildlife Conservation; the John and Beverly Stauffer Foundation; FP7 European

Marie-Curie programs (CIG-293845, ITN-290344, IEF-328024, IEF-299176 and IEF-302617); a National Science Foundation Graduate Research Fellowship, and the Human Frontier Science Program (LT000320/2014).

REFERENCES

1. Boyd, L., and King, S. Equus ferus ssp. przewalskii. IUCN 2011 IUCN Red List Threat. Species. Available at: www.iucnreedlist.org.

2. Mohr, E. (1959). Das Urwildpferd (Equus przewalski Poljakoff 1881) Mit 87 Abbildungen Die Neue Brehm-Bücherei 249. (Wittenberg Lutherstadt: A. Ziemsen Verlag).

3. Volf, J., Kus, E., and Prokopova, L. (1991). General studbook of the Przewalski horse. (Prague (Czech Republic): Zoological Garden Prague).

4. Zimmermann, W. (2014). International Przewalski's Horse Studbook (Equus ferus przewalskii). (Cologne Zoo, SPARS version).

5. Benirschke, K., Malouf, N., Low, R. J., and Heck, H. (1965). Chromosome complement: differences between Equus caballus and Equus przewalskii, Poliakoff. Science *148*, 382–383.

6. Ishida, N., Oyunsuren, T., Mashima, S., Mukoyama, H., and Saitou, N. (1995). Mitochondrial DNA sequences of various species of the genus Equus with special reference to the phylogenetic relationship between Przewalskii's wild horse and domestic horse. J. Mol. Evol. *41*, 180–188.

7. Oakenfull, E. A., and Ryder, O. A. (1998). Mitochondrial control region and 12S rRNA variation in Przewalski's horse (Equus przewalskii). Anim. Genet. *29*, 456–459.

8. Goto, H., Ryder, O. A., Fisher, A. R., Schultz, B., Kosakovsky Pond, S. L., Nekrutenko, A., and Makova, K. D. (2011). A Massively Parallel Sequencing Approach Uncovers Ancient Origins and High Genetic Variability of Endangered Przewalski's Horses. Genome Biol. Evol. *3*, 1096–1106.

9. Bowling, A. T., Zimmermann, W., Ryder, O., Penado, C., Peto, S., Chemnick, L., Yasinetskaya, N., and Zharkikh, T. (2003). Genetic variation in Przewalski's horses, with special focus on the last wild caught mare, 231 Orlitza III. Cytogenet. Genome Res. *102*, 226–234.

10. Wallner, B., Brem, G., Müller, M., and Achmann, R. (2003). Fixed nucleotide differences on the *Y* chromosome indicate clear divergence between *Equus przewalskii* and *Equus caballus*. Anim. Genet. *34*, 453–456.

11. Lau, A. N., Peng, L., Goto, H., Chemnick, L., Ryder, O. A., and Makova, K. D. (2008). Horse Domestication and Conservation Genetics of Przewalski's Horse Inferred from Sex Chromosomal and Autosomal Sequences. Mol. Biol. Evol. *26*, 199–208.

12. Lindgren, G., Backström, N., Swinburne, J., Hellborg, L., Einarsson, A., Sandberg, K., Cothran, G., Vilà, C., Binns, M., and Ellegren, H. (2004). Limited number of patrilines in horse domestication. Nat. Genet. *36*, 335–336.

13. Wallner, B., Vogl, C., Shukla, P., Burgstaller, J. P., Druml, T., and Brem, G. (2013). Identification of Genetic Variation on the Horse Y Chromosome and the Tracing of Male Founder Lineages in Modern Breeds. PLoS ONE *8*, e60015.

14. Lander, E. S., and Lindblad-Toh, K. with Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T. L., Adelson, D. L., Bailey, E., Bellone, R. R., et al., Broad Institute Genome Sequencing Platform and Broad Institute Whole Genome Assembly Team (2009). Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. Science *326*, 865–867.

15. McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., Distl, O., Guérin, G., Hasegawa, T., Hill, E. W., et al. (2012). 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.

16. Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., et al. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature *499*, 74–78.

17. Do, K.-T., Kong, H.-S., Lee, J.-H., Lee, H.-K., Cho, B.-W., Kim, H.-S., Ahn, K., and Park, K.-D. (2014). Genomic characterization of the Przewalski′s horse inhabiting Mongolian steppe by whole genome re-sequencing. Livest. Sci. *167*, 86–91.

18. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. PLoS ONE *7*, e34131.

19. Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. U. S. A. *104*, 14616–14621.

20. Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. *38*, e87.

21. Schubert, M., Jónsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., Albrechtsen, A., Dupanloup, I., Foucal, A., Petersen, B., et al. (2014). Prehistoric

genomes reveal the genetic foundation and cost of horse domestication. Proc. Natl. Acad. Sci. *111*, E5661–E5669.

22. Wallner, B. (2004). Isolation of Y Chromosome-specific Microsatellites in the Horse and Cross-species Amplification in the Genus Equus. J. Hered. *95*, 158–164.

23. Lippold, S., Knapp, M., Kuznetsova, T., Leonard, J. A., Benecke, N., Ludwig, A., Rasmussen, M., Cooper, A., Weinstock, J., Willerslev, E., et al. (2011). Discovery of lost diversity of paternal horse lineages using ancient DNA. Nat. Commun. *2*, 450.

24. Ghosh, S., Qu, Z., Das, P. J., Fang, E., Juras, R., Cothran, E. G., McDonell, S., Kenney, D. G., Lear, T. L., Adelson, D. L., et al. (2014). Copy Number Variation in the Horse Genome. PLoS Genet. *10*, e1004712.

25. Soong, B.-W., Casamassima, A. C., Fink, J. K., Constantopoulos, G., and Horwitz, A. L. (1988). Multiple sulfatase deficiency. Neurology *38*, 1273–1273.

26. Pavlidis, P., Zivkovic, D., Stamatakis, A., and Alachiotis, N. (2013). SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. Mol. Biol. Evol. *30*, 2224–2234.

27. Ransom, J. I., Powers, J. G., Garbe, H. M., Oehler, M. W., Nett, T. M., and Baker, D. L. (2014). Behavior of feral horses in response to culling and GnRH immunocontraception. Appl. Anim. Behav. Sci. *157*, 81–92.

28. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet *5*, e1000695.

29. Yang, M. A., Harris, K., and Slatkin, M. (2014). The Projection of a Test Genome onto a Reference Population and Applications to Humans and Archaic Hominins. Genetics *198*, 1655–1670.

30. Cho, Y. S., Hu, L., Hou, H., Lee, H., Xu, J., Kwon, S., Oh, S., Kim, H.-M., Jho, S., Kim, S., et al. (2013). The tiger genome and comparative analysis with lion and snow leopard genomes. Nat. Commun. *4*.

31. Cahill, J. A., Green, R. E., Fulton, T. L., Stiller, M., Jay, F., Ovsyanikov, N., Salamzade, R., St. John, J., Stirling, I., Slatkin, M., et al. (2013). Genomic Evidence for Island Population Conversion Resolves Conflicting Theories of Polar Bear Evolution. PLoS Genet. *9*, e1003345.

32. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., et al. (2013). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.

33. Skotte, L., Korneliussen, T. S., and Albrechtsen, A. (2013). Estimating Individual Admixture Proportions from Next Generation Sequencing Data. Genetics *195*, 693–702.

34. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient Admixture in Human History. Genetics *192*, 1065–1093.

35. Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. *28*, 2239–2252.

36. Pasaniuc, B., Sankararaman, S., Kimmel, G., and Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. Bioinforma. Oxf. Engl. *25*, i213–221.

37. Seal, U., Lacy, R., Zimmermann, W., Ryder, O., and Princee, F. Captive Breeding Specialist Group.SSC/IUCN. (Apple Valley, MN).

38. Moehlman, P. D. R. (2002). Equids: Zebras, Asses, and Horses : Status Survey and Conservation Action Plan (IUCN).

FIGURE LEGENDS

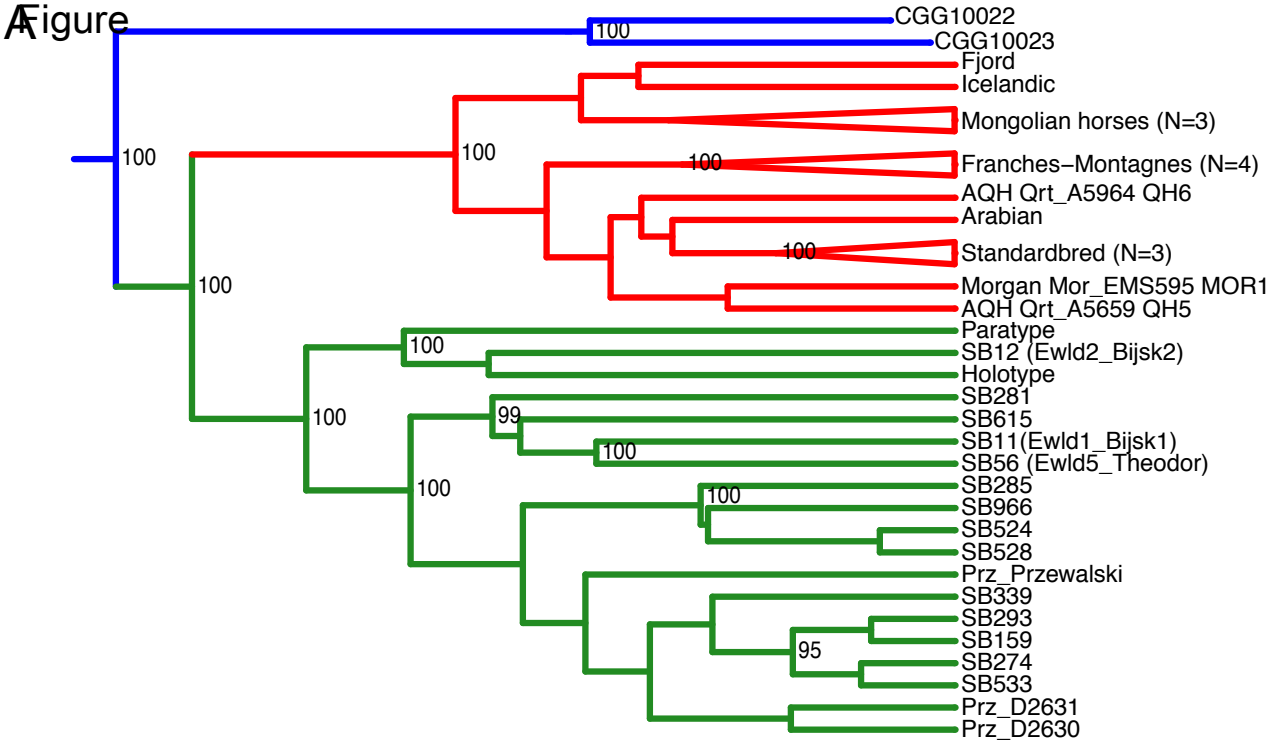
Figure 1. Genomic structure among Late Pleistocene horses (blue), DHs (red) and PHs (green).
(A) Exome-based ML tree. Nodes show bootstrap support≥95%. AQH: American Quarter Horse. (B) PSMC profiles. Thin lines represent 100 bootstrap replicates. (C) PCA based on genotype likelihoods. See also Figures S1-2 and Tables S1-2.

Figure 2. DH and PH demographic models explored in dadi and projection analyses.
(A) Best two-population model supported by dadi. (B) dadi model Log(likelihood) for varying starting dates of isolation between DHs (Franches-Montagnes) and PHs (TEG). (C) Best two-population model supported by projection analyses. (D) Model LSS when considering varying dates for the change in migration dynamics (TMC) for the genome projections based on the DH (Franches-Montagnes, red) and PH (black) reference panels. See also Table S3.

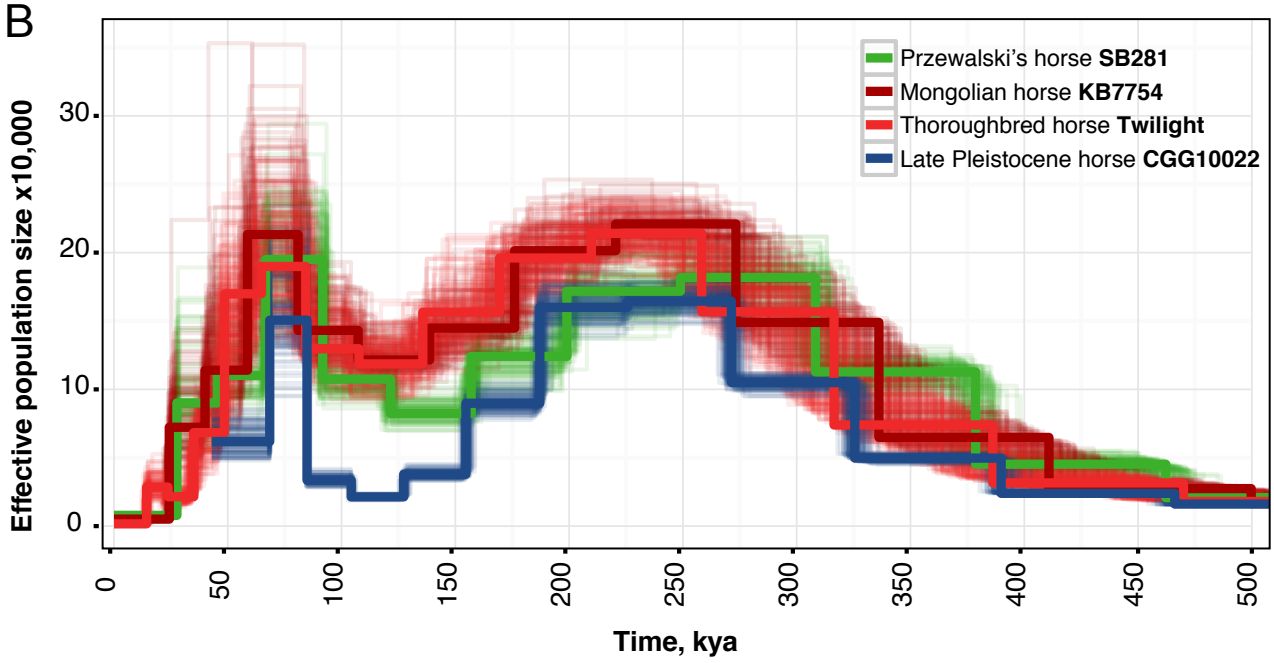
Figure 3. Heterozygosity and inbreeding estimates.
(A) Average genome-wide heterozygosity for autosomes, disregarding transitions. (B) FHMM inbreeding coefficient and inbreeding coverage across modern and historical PHs and DHs. 'x' not estimated.

Figure 4. DH introgression into PHs.
*A-line PHs. (A) Genetic components estimated in NGSAdmix (eight clusters). The grey and red gradients represent the PH and DH components, respectively. (B) D-statistics heatmaps. The tree topology (Outgroup,(PH,(DH,Paratype))) is tested for each DH and PH combination, disregarding transitions. Darker red colors indicate the admixture component in significant tests while darker grey shades indicate non-significant tests. The proportion of DH introgression within each PH genome, estimated by the average of f4-ratios involving Franches-Montagnes horses, is shown between parentheses. (C) Chromosome painting of DH (red) and PH (grey) ancestry blocks identified along chromosome 5. See also Figure S3 and Tables S4-5.

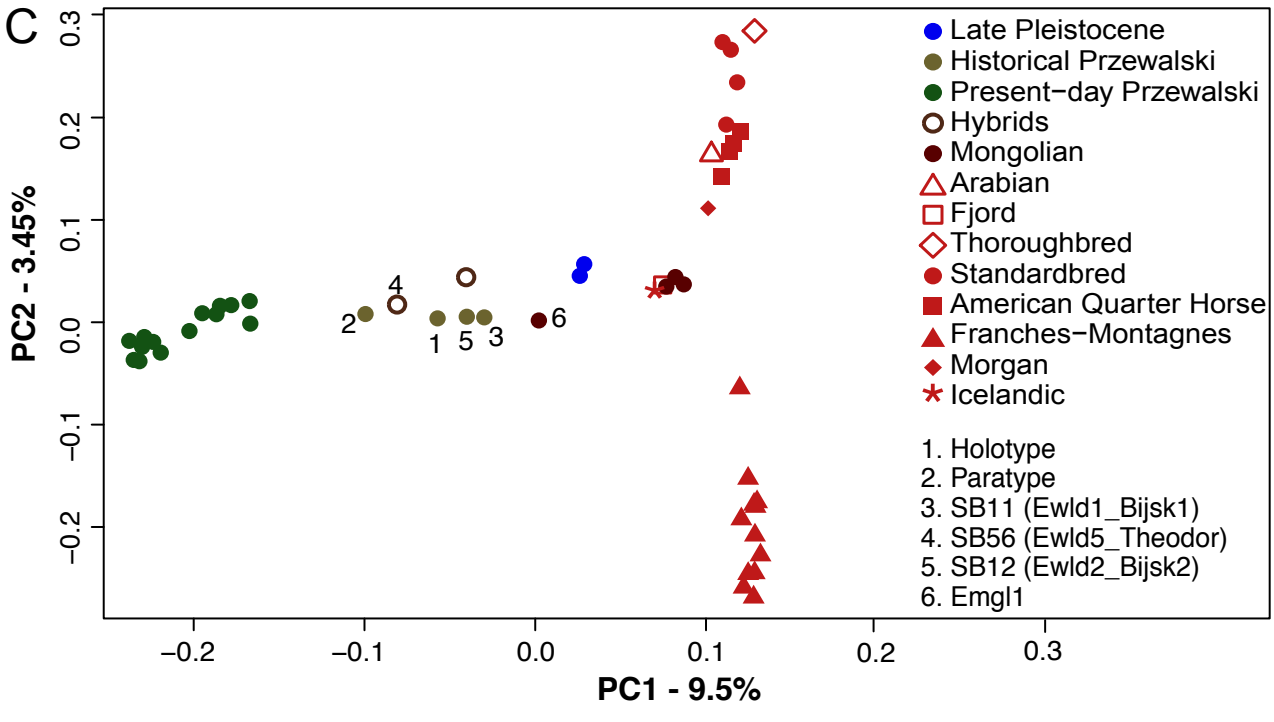
Figure


CGG10022
CGG10023
Fjord
Icelandic
Mongolian horses (N=3)
Franches−Montagnes (N=4)
AQH Qrt_A5964 QH6
Arabian
Standardbred (N=3)
Morgan Mor_EMS595 MOR1
AQH Qrt_A5659 QH5
Paratype
SB12 (Ewld2_Bijsk2)
Holotype
SB281
SB615
SB11(Ewld1_Bijsk1)
SB56 (Ewld5_Theodor)
SB285
SB966
SB524
SB528
Prz_Przewalski
SB339
SB293
SB159
SB274
SB533
Prz_D2631
Prz_D2630
100
99
95

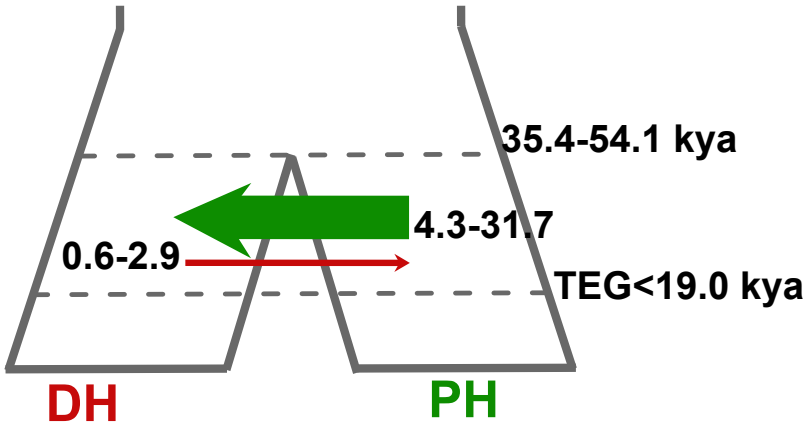
B

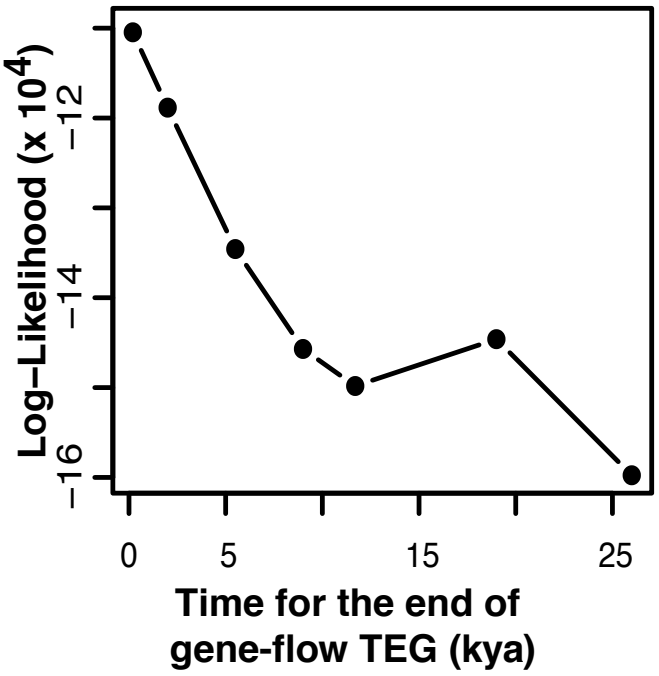

Przewalski's horse SB281
Mongolian horse KB7754
Thoroughbred horse Twilight
Late Pleistocene horse CGG10022
Effective population size x10,000
Time, kya

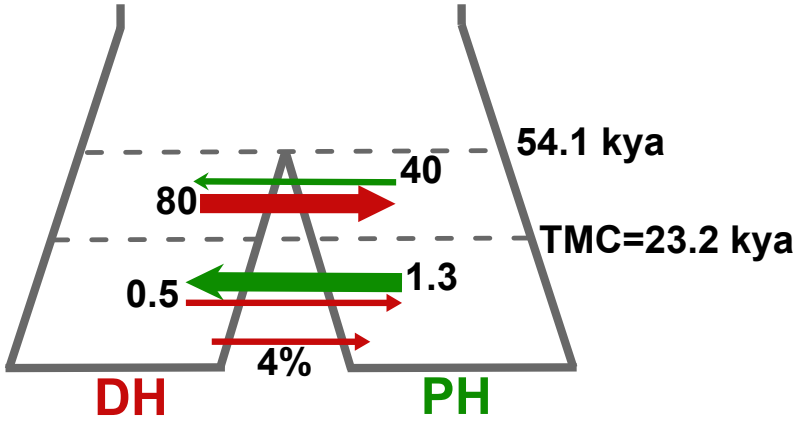
C

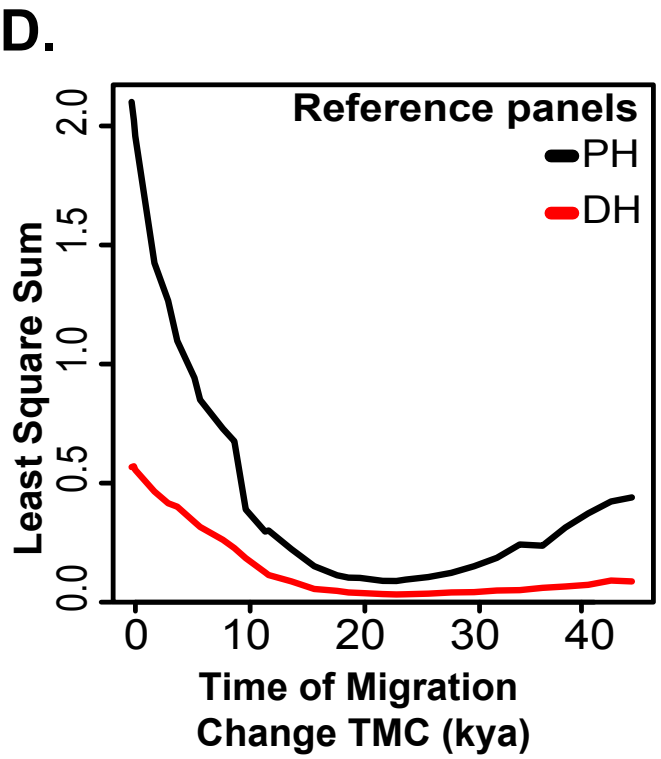

Late Pleistocene
Historical Przewalski
Present−day Przewalski
Hybrids
Mongolian
Arabian
Fjord
Thoroughbred
Standardbred
American Quarter Horse
Franches−Montagnes
Morgan
Icelandic
1. Holotype
2. Paratype
3. SB11 (Ewld1_Bijsk1)
4. SB56 (Ewld5_Theodor)
5. SB12 (Ewld2_Bijsk2)
6. Emgl1
PC2 - 3.45%
PC1 - 9.5%

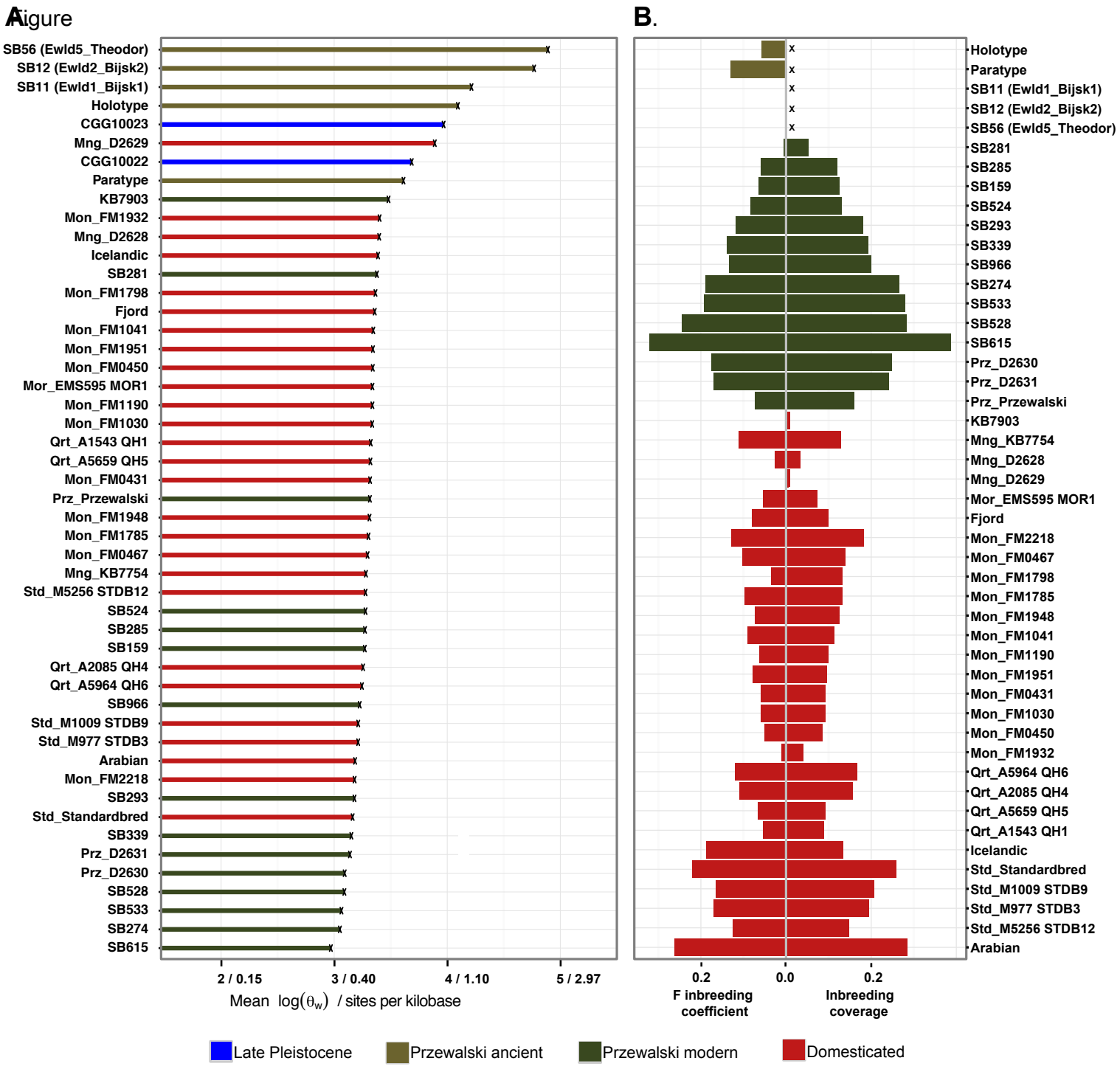
Figure


A.
35.4-54.1 kya
4.3-31.7
0.6-2.9
TEG<19.0 kya
DH
PH

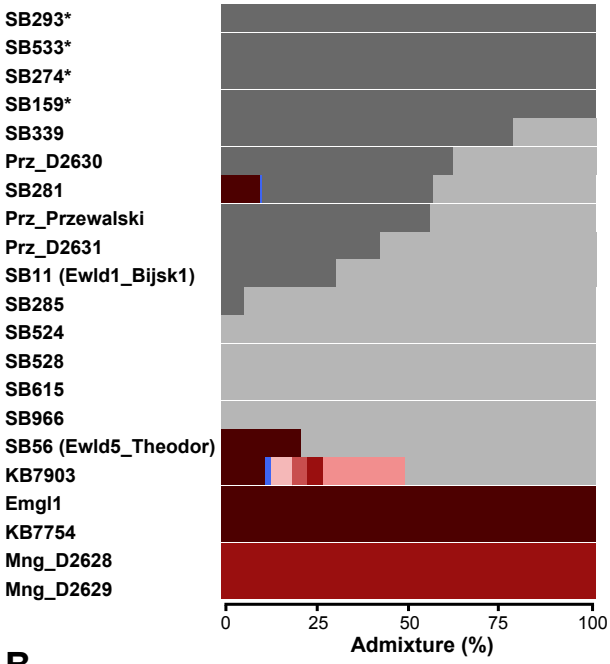

B.
Log–Likelihood (x 10^4)
−12
−14
−16
0
5
15
25
Time for the end of
gene-flow TEG (kya)

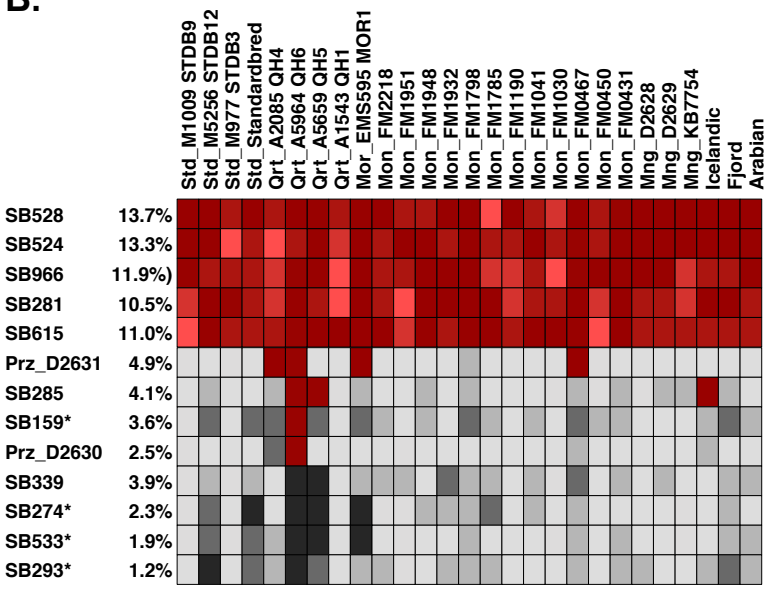

C.
54.1 kya
40
80
TMC=23.2 kya
0.5
1.3
4%
DH
PH

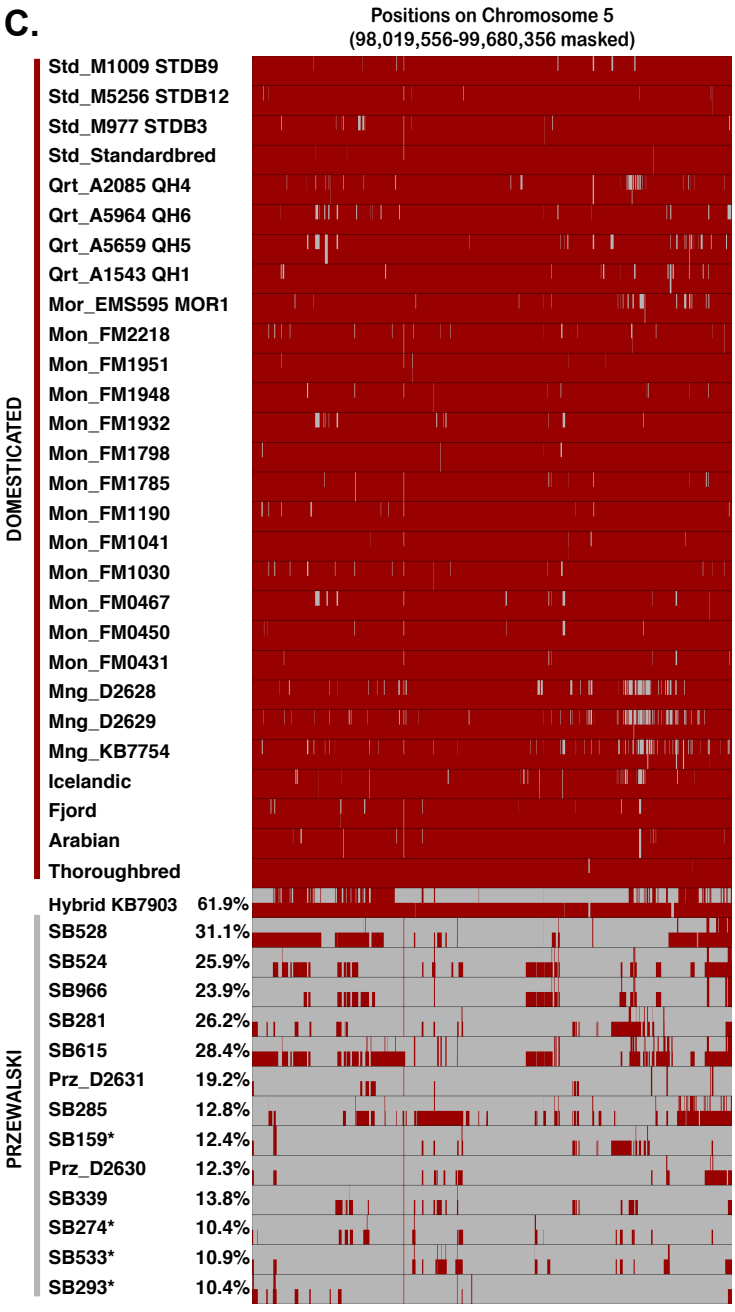

D.
Reference panels
PH
DH
Least Square Sum
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
Time of Migration
Change TMC (kya)


Figure
A.
B.
SB56 (Ewld5_Theodor)
SB12 (Ewld2_Bijsk2)
SB11 (Ewld1_Bijsk1)
Holotype
CGG10023
Mng_D2629
CGG10022
Paratype
KB7903
Mon_FM1932
Mng_D2628
Icelandic
SB281
Mon_FM1798
Fjord
Mon_FM1041
Mon_FM1951
Mon_FM0450
Mor_EMS595 MOR1
Mon_FM1190
Mon_FM1030
Qrt_A1543 QH1
Qrt_A5659 QH5
Mon_FM0431
Prz_Przewalski
Mon_FM1948
Mon_FM1785
Mon_FM0467
Mng_KB7754
Std_M5256 STDB12
SB524
SB285
SB159
Qrt_A2085 QH4
Qrt_A5964 QH6
SB966
Std_M1009 STDB9
Std_M977 STDB3
Arabian
Mon_FM2218
SB293
Std_Standardbred
SB339
Prz_D2631
Prz_D2630
SB528
SB533
SB274
SB615
2 / 0.15
3 / 0.40
4 / 1.10
5 / 2.97
Mean log(θw) / sites per kilobase
Holotype
Paratype
SB11 (Ewld1_Bijsk1)
SB12 (Ewld2_Bijsk2)
SB56 (Ewld5_Theodor)
SB281
SB285
SB159
SB524
SB293
SB339
SB966
SB274
SB533
SB528
SB615
Prz_D2630
Prz_D2631
Prz_Przewalski
KB7903
Mng_KB7754
Mng_D2628
Mng_D2629
Mor_EMS595 MOR1
Fjord
Mon_FM2218
Mon_FM0467
Mon_FM1798
Mon_FM1785
Mon_FM1948
Mon_FM1041
Mon_FM1190
Mon_FM1951
Mon_FM0431
Mon_FM1030
Mon_FM0450
Mon_FM1932
Qrt_A5964 QH6
Qrt_A2085 QH4
Qrt_A5659 QH5
Qrt_A1543 QH1
Icelandic
Std_Standardbred
Std_M1009 STDB9
Std_M977 STDB3
Std_M5256 STDB12
Arabian
0.2
0.0
0.2
F inbreeding coefficient
Inbreeding coverage
Late Pleistocene
Przewalski ancient
Przewalski modern
Domesticated

Figure


A.
SB293*
SB533*
SB274*
SB159*
SB339
Prz_D2630
SB281
Prz_Przewalski
Prz_D2631
SB11 (Ewld1_Bijsk1)
SB285
SB524
SB528
SB615
SB966
SB56 (Ewld5_Theodor)
KB7903
Emgl1
KB7754
Mng_D2628
Mng_D2629
0
25
50
75
100
Admixture (%)


B.
Std_M1009 STDB9
Std_M5256 STDB12
Std_M977 STDB3
Std_Standardbred
Qrt_A2085 QH4
Qrt_A5964 QH6
Qrt_A5659 QH5
Qrt_A1543 QH1
Mor_EMS595 MOR1
Mon_FM2218
Mon_FM1951
Mon_FM1948
Mon_FM1932
Mon_FM1798
Mon_FM1785
Mon_FM1190
Mon_FM1041
Mon_FM1030
Mon_FM0467
Mon_FM0450
Mon_FM0431
Mng_D2628
Mng_D2629
Mng_KB7754
Icelandic
Fjord
Arabian
SB528 13.7%
SB524 13.3%
SB966 11.9%)
SB281 10.5%
SB615 11.0%
Prz_D2631 4.9%
SB285 4.1%
SB159* 3.6%
Prz_D2630 2.5%
SB339 3.9%
SB274* 2.3%
SB533* 1.9%
SB293* 1.2%
PRZEWALSKI


C.
Positions on Chromosome 5
(98,019,556-99,680,356 masked)
DOMESTICATED
Std_M1009 STDB9
Std_M5256 STDB12
Std_M977 STDB3
Std_Standardbred
Qrt_A2085 QH4
Qrt_A5964 QH6
Qrt_A5659 QH5
Qrt_A1543 QH1
Mor_EMS595 MOR1
Mon_FM2218
Mon_FM1951
Mon_FM1948
Mon_FM1932
Mon_FM1798
Mon_FM1785
Mon_FM1190
Mon_FM1041
Mon_FM1030
Mon_FM0467
Mon_FM0450
Mon_FM0431
Mng_D2628
Mng_D2629
Mng_KB7754
Icelandic
Fjord
Arabian
Thoroughbred
Hybrid KB7903 61.9%
SB528 31.1%
SB524 25.9%
SB966 23.9%
SB281 26.2%
SB615 28.4%
Prz_D2631 19.2%
SB285 12.8%
SB159* 12.4%
Prz_D2630 12.3%
SB339 13.8%
SB274* 10.4%
SB533* 10.9%
SB293* 10.4%

Supplemental data

Please follow this link for a detailed description:
http://geogenetics.ku.dk/publications/si-przewalski/SOM_CurrentBiology.pdf
which includes the following sections:

1 FOREWORD
2 GENOME SEQUENCING
3 CHARACTERISATION OF THE HORSE GENOMIC DIVERSITY
4 ADMIXTURE ANALYSES
5 DEMOGRAPHIC RECONSTRUCTION BASED ON NUCLEAR GENOME INFORMATION
6 SELECTION SCANS

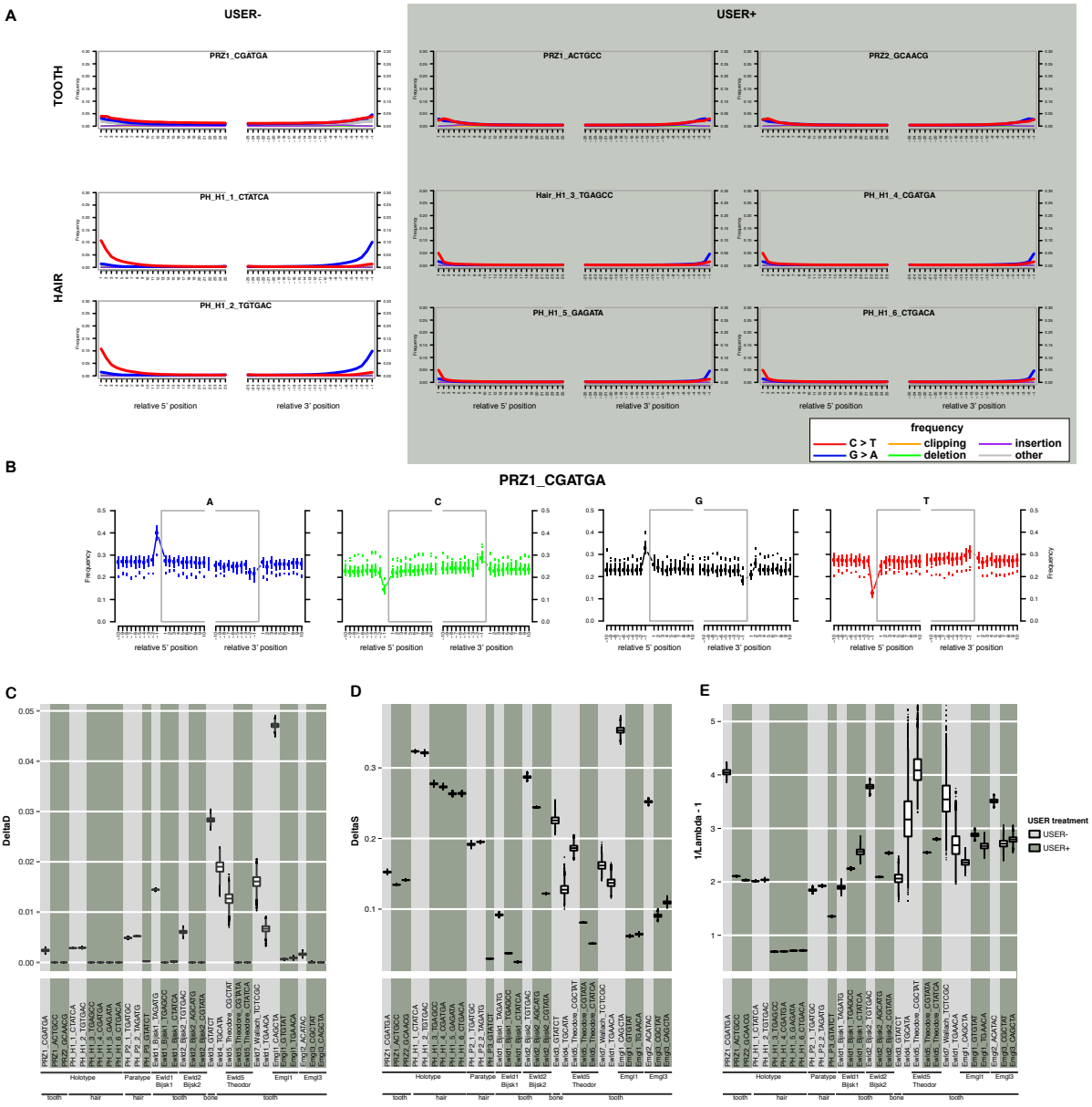


Figure S1. DNA damage parameters in historical horses.
A. Nucleotide misincorporation patterns along DNA reads for each library constructed for the Holotype specimen. DNA was extracted from both tooth and hair samples, and libraries were constructed from both non-USER treated and USER-treated DNA extracts. Misincorporation frequencies are shown for the first and the last 25 nucleotides sequenced. B. DNA fragmentation patterns are shown for one library constructed from the Holotype specimen (PRZ1_CGATGA without prior USER treatment) for the 10 bp preceding read starts (positions -1 to -10 on the left panels of each base composition profile) and the 10 bp following read ends (positions 1 to 10 on the right panels of each base composition profile). D. DeltaD, deamination rate estimates in double-strand DNA contexts. E. DeltaS, deamination rate estimates in single-strand DNA contexts. F. 1/Lambda -1, proxy for the estimate of the length of overhangs, where Lambda is the mean estimate for the probability of reads terminating in an overhang in a template.

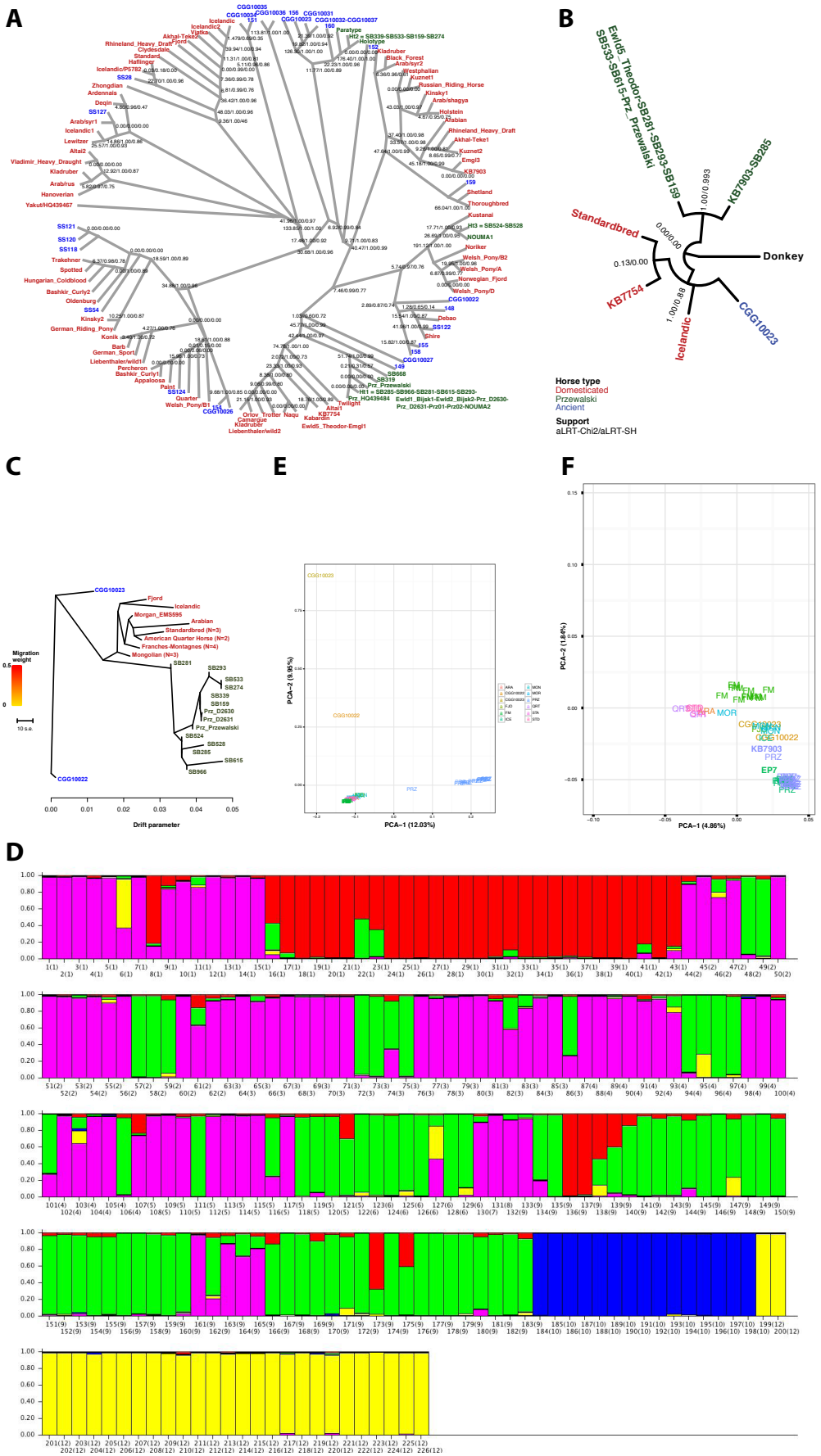

A
B
Ewld5_Theodor-SB281-SB293-SB159
SB533-SB615-Prz_Przewalski
KB7903-SB285
Standardbred
Donkey
KB7754
Icelandic
CGG10023
1.00/0.993
0.00/0.00
0.13/0.00
1.00/0.88
Horse type
Domesticated
Przewalski
Ancient
Support
aLRT-Chi2/aLRT-SH
C
CGG10023
Fjord
Icelandic
Morgan_EMS595
Arabian
Standardbred (N=3)
American Quarter Horse (N=2)
Franches-Montagnes (N=4)
Mongolian (N=3)
SB281
SB293
SB533
SB274
SB339
SB159
Prz_D2630
Prz_D2631
Prz_Przewalski
SB524
SB528
SB285
SB615
SB966
CGG10022
Migration weight
0.5
0
10 s.e.
0.00
0.01
0.02
0.03
0.04
0.05
Drift parameter
E
CGG10023
CGG10022
PRZ
PCA–2 (8.95%)
PCA–1 (12.03%)
F
PCA–2 (14.8%)
PCA–1 (4.86%)
FM
MOR
QRS
GRA
CGG10023
CGG10022
KB7903
PRZ
EP7
D

Figure S2. Genetic structure among ancient and present-day horses.
A. Maximum-likelihood phylogenetic reconstructions of complete mitochondrial genomes for ancient and present-day horses. B. Maximum-likelihood phylogenetic reconstructions of Y-chromosome haplotypes for ancient and present-day horses. Branch supports are given as approximate likelihood-ratio (aLRT), Chi2 and Shimodeira–Hasegawa (SH) tests. (A and B). C. TreeMix population graphs. No migration edge considering each Przewalski's horse as a separate population (95.176 % of the variation is explained by the tree). D. Structure analyses of twelve Single Tandem Repeats in donkeys, domesticated and Przewalski's horses (k=5). 1-43: Arabian (1); 44-62: American Quarter Horse (2); 63-86: Warmblood (3); 87-106: Pony (4); 107-121: Thoroughbred (5); 122-129: Cold Blood (6); 130: Andalusian (7); 131: Tennessee Walker (8); 132-183: Local horses (9); 183-198: Donkeys (10); 199-226: Przewalski's horses (12). E. Principal Component Analyses of full-genome variant genotypes. 198,932 variants. "CGG10022" and "CGG10023": Late Pleistocene horses; "KB7754": Mongolian; "MNG": Mongolian; "FM": Franches-Montagnes; "MOR": Morgan; "QRT": Quarter, "STD": Standardbred; "THO": Thoroughbred; "PRZ": Przewalski's horses. F. Principal Component Analyses of Illumina SNP50 Beadchip variants. 20,949 variants. "AH": Akhal Teke, "AND": Andalusian, "ARA": Arabian, "BEL": Belgian, "CL": Clydesdale, "CS": Caspian, "EX": Exmoor, "FELL": Fell pony, "FINN": FinnHorse, "FJO": Norwegian Fjord, "FM": Franches-Montagnes, "FT": French Trotter, "HAN": Hanoverian, "ICE": Icelandic, "MINI": Miniature, "MNGP": Mangalarga Paulista, "MON": Mongolian, "MOR": Morgan, "NF": New Forest pony, "NORF": Norwegian Fjord, "NSWE": North Swedish, "PERU": Peruvian Paso, "PR": Percheron, "PRZ 1": Przewalski, this study, "PRZ 2": Przewalski, [1], "PT": Paint, "QRT": American Quarter Horse, "RP": Puerto Rican Paso Fino, "SB": Saddlebred, "SHET": Shetland, "SHI": Shire, "STA": Standardbred, "STD": Standardbred, "SZWB": Swiss Warmblood, "TB": Thoroughbred, "TU": Tuva, "EP7": Historical Przewalski.

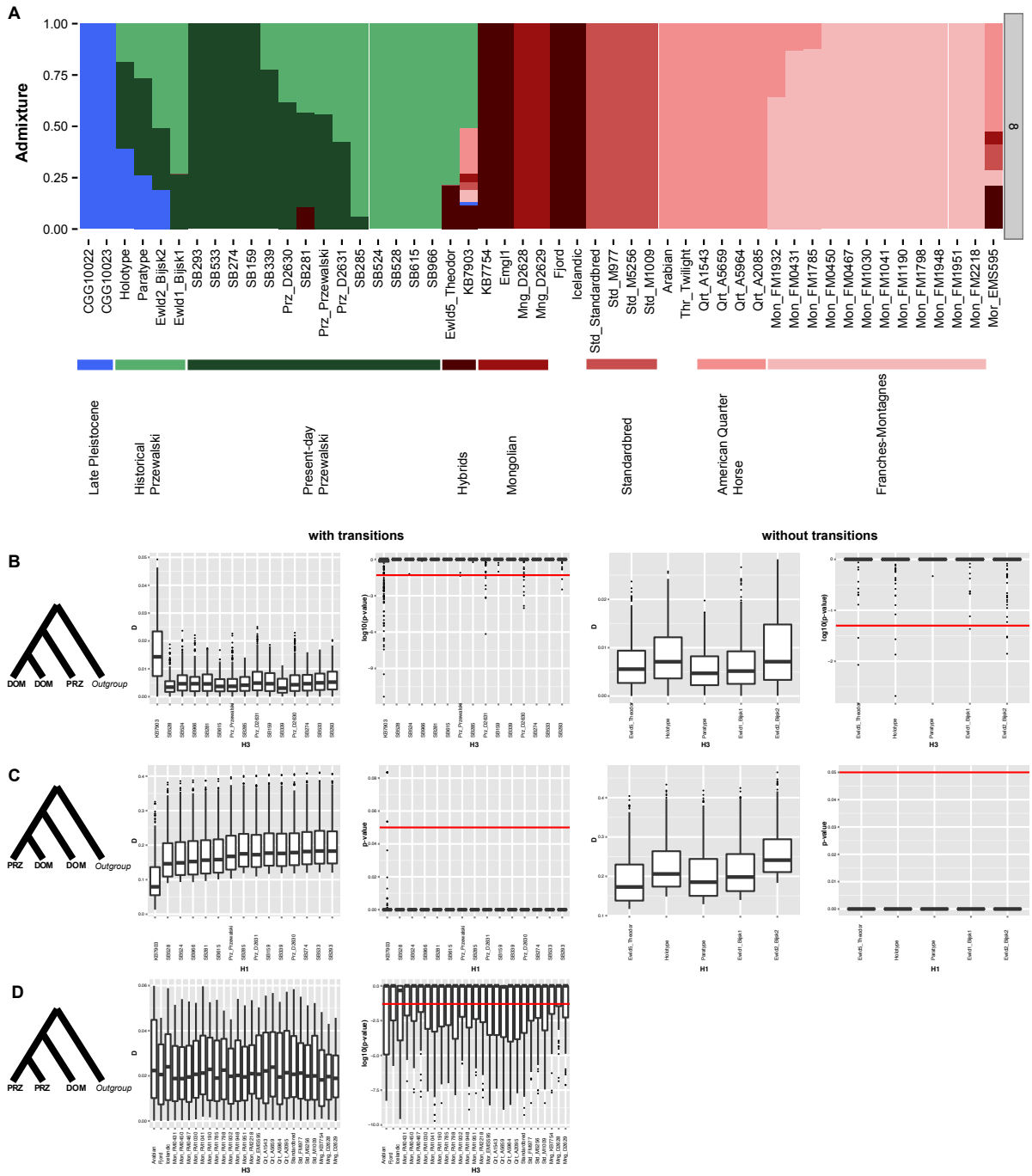


Figure S3. Distributions of the ABBA-BABA D-statistics and of the p-values when testing for admixture between present-day or historical Przewalski's horses and domesticated horses. p-values represent the Z-scores corrected for multiple testing. "DOM": domesticated horses; "PRZ": Przewalski's horses. A. Test of the ((Domesticated1, Domesticated2), Przewalski), Outgroup) tree topologies (351 tests per H3). B. Test of the ((Przewalski1, Domesticated1), Domesticated2), Outgroup) tree topologies (702 tests per H1). C. Test of the ((Przewalski1, Przewalski2), Domesticated), Outgroup) tree topologies (91 tests per H3).

Supplemental Experimental Procedures

In this study we reconstruct the phylogenetic, demographic and evolutionary history of Przewalski's horses, by comparing complete genomes of Przewalski's horses with those from a range of modern domesticated breeds and Late Pleistocene horses (**Table S1**).
To achieve this, we sequenced the complete genomes of:

- present-day Przewalski's horses (12 genomes, sequenced in Copenhagen, Denmark),
- historical Przewalski's horses, Mongolian horses and hybrids (6 samples, sequenced in Copenhagen, Denmark),
- present-day horses belonging to the domesticated breeds of Mongolian (1 genome, sequenced in Copenhagen, Denmark), Franches-Montagnes (12 genomes, sequenced in Bern, Switzerland), Morgan, American Quarter Horse (subsequently referred to as "Quarter") and Standardbred horses (8 genomes, sequenced Minneapolis, Minnesota).

And we compared these genomes to the following genomes, already available from the literature:

- 3 genomes from present-day Przewalski's horses ([2–4]),
- 7 genomes from present-day horses belonging to the domesticated breeds of Mongolian [3], Thoroughbred [5], Arabian, Standardbred, Norwegian Fjord, Icelandic horses [2, 4],
- and the genomes from 2 Late Pleistocene horses [4].

Genome sequencing in Copenhagen: Sample information

Modern sample information and modern DNA extraction

We selected eleven living Przewalski's horses representing the main lineages of the Przewalski's horse pedigree [6–8], as well as one Mongolian horse and one domesticated-Przewalski's horse F1 hybrid (**Table S1**). DNA was extracted from immortalized cell lines at the San Diego Zoo's Institute for Conservation Research using the DNeasy Blood and Tissue kit (Qiagen, CA) according to manufacturer's instructions, before it was shipped to the Centre for GeoGenetics, University of Copenhagen, Denmark, for DNA library construction and high-throughput sequencing. Material was shipped in compliance with CITES regulation for endangered species, following Materials Transfer Agreements between Copenhagen and San Diego (Biomaterial Request # BR2013045) and Certificate of Scientific Exchange (COSE) in place at both institutions 12US50818A/9 San Diego, and #DK003 Natural History Museum of Denmark, Copenhagen).

Ancient sample information and ancient DNA extraction

Our study includes the genomic characterization of the Przewalski's horse holotype, i.e. the individual used to define the taxon in the late 19th century, and a paratype individual, also captured at the same time, and which helped describe the taxon. These individuals do not represent founders of the captive pedigree of Przewalski's horses. Tooth and hair samples for the Przewalski's horse holotype and a hair sample for one Przewalski's horse paratype were obtained from the Zoological Museum of St Petersburg, Russia. The Przewalski's horse holotype (museum ID 512-ZI S-Pt.) was collected in Dzungaria, Guchena, sands Kanabo in the Gobi Desert in 1878 by the Russian explorer Colonel Nikolaus M. Przewalski and was first described by L.S. Poliakov at the Zoological Museum of the Academy of Science in St Petersburg in 1881. The paratype (museum ID 10554-ZI S-Pt.) was collected at Well Ebi between Kobdo and Barkul in the Gobi Desert in 1899 (**Table S1**).

An additional nine bone and tooth samples of historical horses (including Mongolian, Przewalski's horses and alleged hybrids) were obtained from the zoological collections of the Martin-Luther University in Halle, Germany (Museum of domesticated animals "Julius Kühn"). Two of these, Ewld1_Bijsk1 (SB11) and Ewld2_Bijsk2 (SB12), represent historical Przewalski's horses that were captured in 1901 and constitute two of the 12 wild Przewalski founders of the captive Przewalski's horse stock. Individuals Ewld3 (SB55) and Ewld4 (SB57) represent their

direct descent following breeding in Halle. Additionally, cross-breeding between Ewld1_Bijsk1 and the domesticated Mongolian mare Emgl1 (considered as the 13th founder of the Przewalski's horse population) is supposed to have resulted in the individual Ewld5_Theodor, sire of individuals Ewld1 (with Ewld4 as dame) and Ewld7_Wallach (with the domesticated Mongolian mare Emgl3 as dame) (**Table S1**). According to the International Studbook [8], the modern Przewalski's horses SB615, SB528, SB281 and SB966 examined in this study are descendants of the founder Ewld2_Bijsk2 through their maternal line. Descendants of Ewld1_Bijsk1, Ewld2_Bijsk2 and Emgl1 have been previously referred to as the "Prague line", whereas the other line named the "A-line" (previously "Munich line") includes the other horses, mainly descending from SB17, SB18, SB39, and SB40 [9].

DNA from historical horse bones, teeth and hair was extracted at the ancient DNA facilities of the Centre for GeoGenetics, University of Copenhagen, Denmark, and strict procedures were followed to prevent contamination from modern DNA. First, the outer surface of bones and teeth was eroded using a Dremel® tool, and 50-500 mg bone or tooth powder was obtained by drilling inside the bone or tooth root. Hair samples were cut into small pieces and rinsed through two washes of sterile demineralized water. Then, three different ancient DNA (aDNA) extraction methods were applied to the samples: a "silica method", a "silica-hair method", and a "Dabney method" (**Table S1**).

The "silica method" was applied to bone and tooth samples as described in [10], with slight modifications. Double digestion of ancient bone/tooth powder prior to DNA purification was shown to significantly increase the endogenous content of ancient DNA extracts, while allowing the retrieval of molecules less affected by post-mortem damage [11, 12]. Accordingly, we followed a procedure where bone/tooth powder was digested twice overnight at 37°C in 5 mL of a buffer containing EDTA (0.5M), N-laurylsarcosine (0.5%) and proteinase K (0.4 mg/mL), using post-centrifugation pellets of undigested bone/tooth powder for the second digestion step. DNA was then purified using the "silica method" and recovered in a final elution volume of 120 µL EB buffer (QIAGEN).

The "silica-hair method" was applied to hair samples and was similar to the "silica method" except that the hair was digested at 55°C in a different buffer containing Tris (10 mM), NaCl (10 mM), CaCl (5 mM), EDTA (2.5 mM), SDS (1%), DTT (100 mM) and proteinase K (1 mg/mL). We also added 5 mg of proteinase K to the digestion buffer after the first day of digestion before the hair was left to digest one extra day [13].

The "Dabney method" was applied to bone and tooth samples. The protocol described in [14] was followed implementing the double digestion described above. DNA was recovered in a final elution volume of 56 µL EB buffer.

Genome sequencing in Copenhagen: library building and sequencing

We built blunt-ended libraries for shotgun sequencing on the Illumina platforms from both modern Przewalski's and historical horse DNA extracts following the protocols in [2, 15] with few modifications (**Table S1**).

DNA libraries on fresh DNA extracts were constructed in post-PCR facilities at the Centre for GeoGenetics, which are located in a different building than the ancient DNA facilities. For modern DNA extracts, 1 µg of DNA in 100 µL TE buffer was sheared using a Bioruptor NGS device (Diagenode) set to four cycles of 15 seconds ON / 90 seconds OFF in order to obtain a DNA fragment size distribution centered around 500 bp. The sheared DNA solution was then concentrated in 22 µL EB buffer using the MinElute Purification kit by QIAGEN.

Ancient DNA libraries and mixes for PCR amplification were prepared in the ancient DNA facilities of the Centre for GeoGenetics. We built libraries directly (without shearing) on extracts or on extracts after treatment with the USER Enzyme mix by New England Biolabs (**Table S1**). The USER Enzyme contains a mixture of Uracil DNA glycosylase and endonuclease

VIII that removes uracil residues, and was shown to decrease significantly the rate of sequence errors due to cytosine deamination in aDNA molecules [16]. We incubated 16.5 uL of aDNA extract with 5 uL of USER Enzyme mix at 37°C for 3 hours.

Blunt-End libraries were constructed from modern and aDNA extracts using the NEBNext Quick DNA Library Prep Master Mix Set for 454 (New England BioLabs, reference nb. E6070L) following [2, 15], in 25 µL volumes and with 0.5 µM Illumina adapters (final concentration). We used the following incubation times and temperatures: 20 min at 12°C, 15 min at 37°C for end-repair; 20 min at 20°C for ligation; 20 min at 37°C, 20 min at 80°C for fill-in. After the end-repair and ligation, reaction mixes were purified using the MinElute PCR Purification Kit (QIAGEN) with elution volumes of 16 µL and 22 µL EB buffer respectively, while after the fill-in step, the Bst enzyme was heat-inactivated.

Blunt-end libraries were amplified in two independent PCR reactions using indexed Illumina primers (one index per sample and per amplification). Amplification was performed in 50 µL reaction mixes containing: 12.5 µL DNA library, 5 units Taq Gold (Life Technologies), 1X Gold Buffer, 4 mM MgCl2, 1 mg/ml BSA, 0.25 mM of each dNTP, 0.5 µM of Primer inPE1.0 (5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T-3') and 0.5 µM of an Illumina 6 bp-indexed ('I') primer (5'-CAA GCA GAA GAC GGC ATA CGA GAT III III GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG-3'). PCR mixes were prepared in the ancient DNA facilities of the Centre for GeoGenetics, where the DNA libraries were also added. PCR amplifications were performed in the post-PCR facilities at the Centre for GeoGenetics.

Thermal cycler conditions for the amplifications were: activation at 92°C for 10 min; followed by 9-22 cycles (see details below) of denaturation at 92°C for 30 sec, annealing at 60°C for 30 sec, elongation at 72°C for 30 sec; and final elongation at 72°C for 7 min. The number of amplification cycles was nine for the modern Przewalski's horse samples. For the ancient samples, it varied depending on the quality of DNA, and in some instances amplification products were purified and concentrated down to 25 µL using the MinElute PCR Purification Kit (QIAGEN) before being amplified a second time in four independent PCR volumes of 25 µL and in the same conditions as described above. Ultimately, the total number of amplification cycles for the ancient samples ranged from 10 to 22. PCR products were purified using the MinElute PCR Purification kit, with a final elution volume of 25 µL EB buffer. In order to control for contamination, extraction, library and PCR blanks were analysed at the same time as the samples. Amplified libraries and blanks were quantified using the 2100 Bioanalyzer (Agilent) High-Sensitivity DNA Assay. No detectable amount of DNA could be recovered from the blanks, suggesting low amounts of contaminating DNA from the laboratory, if any.

All indexed DNA libraries were pooled and sequenced on MiSeq (modern samples) and HiSeq2000 (modern and ancient samples) Illumina platforms at the Danish National High-Throughput DNA Sequencing Centre. We obtained 51-bp SE-MiSeq sequencing reads and 98-bp pair-end (PE) sequencing reads from modern samples, and 94-bp SE, 76-bp PE and 94-bp PE sequencing reads from ancient samples. The DNA sequence data was deposited on the European Nucleotide Archive (Project number PRJEB10098).

Genome sequencing in Bern

Blood samples were obtained from twelve Franches-Montagnes horses from Agroscope, Swiss National Stud Farm, Avenches, Switzerland. Genomic DNA was extracted at the Institute of Genetics of the University of Bern, Switzerland, using the Nucleon Bacc2 kit (GE Healthcare), and built into Illumina TrueSeq v2 libraries, with 300 bp inserts (**Table S1**). The genome of each horse was sequenced using one lane of 100-bp PE Illumina HiSeq reads with the v3 chemistry.

Genome sequencing in Minneapolis

Blood samples were obtained via jugular veinapuncture from American Quarter Horses (n=4), Standardbred (n=3), and Morgan horses (n=1). Genomic DNA was isolated from whole blood with commercially available kits according to the manufacturers' protocols (Purogene Blood Kit C or Qiagen's DNeasy Blood and Tissue Kit; **Table S1**). Illumina TrueSeq v2 libraries were built, all horses were sequenced with 100-bp, paired end reads, 7 horses were sequenced on one lane (target ≥ 12-fold coverage), 1 was sequenced on one half-lane (target ≥ 6-fold coverage).

Genomes publicly available

In subsequent analyses, we included genomic sequencing data previously published for present-day Przewalski's horses and domesticated horses, as well as for two ancient horses that lived prior to domestication in the Late Pleistocene, and from which complete genomes were reconstructed (**Table S1**).

For complete genome sequences previously published by some of the authors of the present study, i.e., one Przewalski's horse, one Arabian, one Standardbred, one Norwegian Fjord, and one Icelandic horses [2, 4], we directly used complete genome sequence alignments (quality filtered BAM files), as those were produced following the same procedure as the one described below. This was also the case for the Late Pleistocene horses, labeled CGG10022 and CGG10023, excavated in the Taymyr peninsula, Russian Federation, and dated to 42,692 ± 891 (UBA-16478) and 16,099 ± 192 calibrated years before present (yBP; UBA-16479), respectively [2, 4].

For other horses, i.e., one Thoroughbred [5], two Przewalski's horses and two Mongolian horses [3] we retrieved the sequencing reads from FASTAQ files made publicly available and performed read alignment as described below.

Post-sequencing read processing and read alignment against reference genomes

The sequencing reads were processed using the PALEOMIX pipeline and following a procedure similar to that described in [17]. We trimmed adapters using AdapterRemoval v1.2 [18] and the parameters described in [2].

Paired-ended reads showing a sequence overlap of at least 11 bp were collapsed into a single consensus sequence (with a threshold set at 1 for the maximal number of mismatches allowed, or one third of the overlapping region in the case of sequences overlapping for more than 11 bp; [2]).

Sequencing reads were aligned against the reference sequences for the horse nuclear (EquCab2.0, removing the mitochondrial genome; [5]) and mitochondrial genomes (Accession Number NC_001640.1; [19]). Sequencing reads were also mapped against chromosome "Un" (unplaced contigs), so that reads aligning to the genome at several locations could also be taken into account. Mapping was performed using BWA v5.9 [20] in the PALEOMIX pipeline, disabling seeding, and applying a threshold of 25 for the minimum read mapping quality allowed [21]. Duplicates were identified and filtered by the 5'-end mapping coordinate for singleton reads (SE reads and PE reads where one mate was discarded due to low quality), and by both external coordinates for PE and collapsed reads.

Characterizing and accounting for post-mortem damage in ancient DNA extracts

As fragmentation and base misincorporations are the main characteristics of damaged aDNA molecules, we used mapDamage v2.0.1 [22] to assess the extent to which these types of damage affect the DNA molecules in the historical horse extracts. We characterized rates of deamination in double strands (DeltaD) and single strands (DeltaS), as well as the probability of reads not terminating in overhangs (Lambda, transformed into 1/Lambda – 1, a proxy for the overhang length of overhanging regions). Parameters were estimated on the basis of 100,000 randomly selected sequence alignments (option –n 100000) using otherwise default parameters for

each individual DNA library (**Figure S1**). We trimmed 0 to 9 bases at read ends and rescaled the quality score of the bases in the BAM alignment file using mapDamage v.2.0.1. For each library, we used the --rescale option in mapDamage v2.0.1 to rescale quality scores according to the bases' probability of being affected by post-mortem damage, which was calculated based on the results of the above damage analyses obtained with the default mapDamage model. This was also done for the alignments of the two Late Pleistocene horse CGG10022 and CGG10023 nuclear genomes obtained from [4], and used in this study for comparison. Rescaling and trimming parameters for each library are given in **Table S1**.

Microbial profiling

Microbial profiling of the ancient DNA extracts was also performed. Analyses and results are described in the section S2.8 of the Supplemental Information file available online.

Error rates

Error in DNA sequences can arise from DNA damage, nucleotide misincorporations by polymerase during PCR or erroneous base calls at sequencing. We determined the error rates for all genome sequences, i.e., those obtained from ancient samples after trimming and re-scaling, and those generated from modern horses (excluding the Thoroughbred horse Twilight from which the reference sequence was built; EquCab2.0; [5]). We followed the procedure described in [2] (see Supplemental section S4.4 of reference), where, after mapping to EquCab2.0, the alleles were defined as ancestral or derived using an outgroup genome (here, *Equus asinus somaliensis* Shakira, [23]) and a single high quality horse genome representing the "perfect genome" (here EquCab2.0). As any living horse should have the same expected number of derived alleles as the "perfect genome", a relative increase in derived alleles compared to the "perfect genome" is interpreted as an increase in the error rate.

Formally, for a given sample $_S$, the expected number of derived alleles is estimated by

$$E[d_S] = D_S - D_S\varepsilon_S + A_S\varepsilon_S$$

where A_S represents the true number of ancestral alleles, D_S and d_S, the true and observed number of derived alleles, and ε_S, the error rate.

The true number of ancestral and derived alleles were estimated by the number of ancestral and derived alleles observed in the "perfect genome" *p*, i.e., a_P and d_P, respectively.

An estimate of the overall error rate sample $_S$ can then be obtained by:

$$\varepsilon_S = (d_S - d_P) / (a_P - d_P)$$

The type-specific error rates were estimated using a maximum likelihood model based on similar assumptions, which are described in details in [2].

In order to derive the estimates, a single base was sampled at each position for each sample. For this purpose, the EquCab2.0 and the *E. asinus somalicus* Shakira genomes [23] were filtered to keep only bases complying with the following criteria: Phred-mapping quality score >35, Phred-base quality score >25. Error rates were estimated for each genome sequence without additional quality filtering, and again applying thresholds of 30 for the minimum mapping quality and 20 for the minimum base quality for each horse genome sequence (**Table S1**).

Horse nuclear variant genotyping

We carried out variant genotyping as described previously (see the Supplemental Information section S8.3 in [2]). We used the PALEOMIX pipeline [17] as applied in [23] and [17], using SAMTools and BCFTools v0.1.19 [24], Picard Tools v1.97 (http://broadinstitute.github.io/picard/), and BEDTools v2.17.0 [25]. Filtering of base-calls was carried out as part of the PALEOMIX genotyping pipeline, using default parameters, except that the maximum per-site depth was set to be greater or equal to the top 0.5% quantile of the read-depth distribution per sample as calculated using the PALEOMIX tool 'depths'.

We called variant, reference sites, and indels. Indels were only used to filter SNPs adjacent to indels, and were not included in the final sequences, and no multiple-sequence alignment was done following genotyping. SNPs were not genotyped for the X chromosome, and nor for the pseudo-chromosome "Un" neither because of its highly repetitive nature potentially causing SNPs to be called in paralogous sequences.

Mitochondrial DNA analyses

Phylogenetic analyses of complete mitochondrial genome sequences

DNA read processing and mapping was performed using the PALEOMIX pipeline [17] as described above. Sequencing reads were aligned to the reference sequence for the horse mitochondrial genome (Accession number NC_001640.1; [19]). Mapping was performed excluding reads showing mapping qualities <25. A consensus was called using a minimum depth-of-coverage of 3 and filtering for base quality ≥30.

We aligned the complete mitochondrial genome sequences of the historical and present-day Przewalski's, Mongolian and hybrid horses generated in this study to a 107-taxon dataset of previously published mitochondrial genome sequences. This comparative dataset contains the mitochondrial genomes sequences of 27 ancient horses and 80 present-day horses, including ten Przewalski's horses [2, 3, 17, 26–31]. Haplotypes found in more than one Przewalski's horse individual were considered only once in downstream phylogenetic analysis. Modelgenerator v0.85 [32] was used to determine the best substitution model for the dataset: Tamura-Nei + invariant sites + 8-class gamma distribution. Under this model, we ran Maximum Likelihood (ML) analyses in PhyML3.0 [33] (**Figure S2A**).

We generated partial mtDNA sequences from an additional sample set composed of ten historical and 20 present-day Przewalski's horses. Methods and results are described in section S3.1.2 of the Supplemental Information file available online.

Y-chromosome DNA analyses

We used Y-chromosome data to compare Przewalski's horses to Late Pleistocene and domesticated horses, as well as to verify the consistency of the International Studbook [6–8] with regards to paternal lineages. To achieve this, we aligned DNA reads from Przewalski's horse stallions of known pedigrees (Ewld5_Theodor, SB159, SB281, SB293, SB533, SB615, SB285, and the hybrid KB7903), as well as from the Mongolian horse KB7754, to a set of 19 Y-chromosome contigs (accession numbers: G72335.1, G72336.1, G72338.1, G72339.1, AB091794.1 [34], JX646942.1, JX646950.1, JX646958.1, JX646966.1, JX646974.1, JX646982.1, JX646990.1, JX646998.1, JX647006.1, JX647014.1, JX647022.1, JX647030.1, JX647038.1, [35, 36], AY532879.1) representing a total 193,857 bp. Reads were aligned using the PALEOMIX pipeline [4] as described in the above.

The consensus sequence was aligned to previously published Y-chromosome data for the donkey Willy (used as outgroup), the Przewalski's horse Prz_Przewalski, the Late Pleistocene horse CGG10023, and the domesticated horses Std_Standardbred and Icelandic [2, 4, 23]. We ran Maximum Likelihood (ML) analyses in PhyML3.0 [33] under a F81 model [37] previously determined as the best model for the dataset by Modelgenerator v0.85 [32] (**Figure S2B**).

Phylogenetic inference using supermatrix of nuclear coding sequences

The phylogenetic relationships between historical and modern Przewalski's horses, modern domesticated horses and Late Pleistocene horses were characterized using *Equus asinus somaliensis* [23] as an outgroup. We followed the method described in [2, 4, 23] as implemented in the PALEOMIX pipeline [17], using the genotyping calls described above. We used a partitioned supermatrix of the coding sequences of protein-coding genes from Ensembl v76 [38] and randomly sampled 25% of the complete protein-coding genes. The final selection, thus,

included 5,094 genes out of 20,449 protein-coding genes initially included in the Ensemble v76. Support was calculated based on 100 bootstrap trees. We used ExaML v2.04 (http://sco.h-its.org/exelixis/software.html) and RAxML v8.1.3 [39] with default parameters, rooted the resulting tree at the midpoint and plotted it using 'ape' [40]. Singleton variants were filtered out as described in [2] in order to limit the impact of post-mortem damage-derived errors in the genomes of the historical horses.

Phylogenies were reconstructed with and without the hybrid KB7903, with and without historical horses, with and without singleton variant filtering in section S3.7.1 of the Supplemental Information file available online.

Demographic reconstruction using the Pairwise Sequentially Markovian Coalescent

We reconstructed the past demographic history of eleven Przewalski's horses and two Mongolian horses by applying the Pairwise Sequentially Markovian Coalescent to their nuclear genome (PSMC; [41]), previously used on the Thoroughbred horse Twilight [2], an Icelandic horse [42] and the Late Pleistocene horse CGG10022 [4]. Horses were selected on the basis of an at least 20-fold genome coverage (as advised in [41] and [2]).

PSMC analyses were performed as described in the Supplemental Information section 9.2 in [2] with slight modifications employing a per-sample maximum coverage threshold. Briefly, a FASTA consensus sequence was generated for each autosome from the BAM files. SNPs were then called as described above. The coverage distribution was inferred with the 'depths' command included in the PALEOMIX pipeline [17], and excluding sites with a depth equal to zero. In addition, bases with Phred quality scores inferior to 35 were excluded. PSMC inference was run using parameters recommended by the developers (number of iterations = 25; maximum $2N_0$ coalescent time = 15; initial θ_0/ρ = 5; [41]). In order to evaluate the spread of the PSMC reconstructions, chromosomal sequences were split into shorter segments of 500 kb and randomly bootstrapped (with replacement) 100 times. For scaling PSMC profiles, we used a calibration point of 4.5 million years (Myr), a mutation rate of 7.242×10^{-9} per site per generation and a generation time of 8 years [2].

Unscaled and scaled PSMC profiles for a larger number of horses are presented in Figures S5.1-4 of the Supplemental Information file available online.

Genetic relationships between horse populations using TreeMix

We investigated the genetic relationships between present-day Przewalski's horses, domesticated and Late Pleistocene horses using the software TreeMix [43]. This method infers population splits and mixtures using genome-wide allele frequencies and a Gaussian approximation to genetic drift. We used the SNP matrix containing 198,932 SNPs described above. We grouped domesticated horses according to their breeds and kept Przewalski's horses as independent populations. We calculated with plink ([44]; --freq --within --missing) the allele frequencies within each grouping of individuals, which was then converted into the TreeMix input format using the supplied python script ("plink2treemix.py"). The genome of the Thoroughbred individual Twilight was excluded from the analyses in order to avoid biases due to the fact that this genome was used as the reference genome during mapping of all the other genomes, where derived homozygous SNPs likely represent genotyping errors. TreeMix was run considering a global perturbation of populations (-global), the tree was rooted using the CGG10022 sample, and 1,000 SNPs per block. The results of the TreeMix analyses were plotted as population trees using the TreeMix R functions (**Figure S2C**). Other TreeMix trees constructed with and without the hybrid KB7903, grouping Przewalski's horses into a single population, and considering 1 to 4 migration events into the TreeMix model (-m 1-4) can be found in section S3.7.2 of the Supplemental Information file available online.

Single Tandem Repeat analyses

Single Tandem Repeats (STRs) were also typed in 28 present-day Przewalski's horses at KSVM-HUJ. This sample set also includes seven present-day Przewalski's horses sequenced in Copenhagen (SB285, SB293, SB339, SB524, SB528, SB533, SB615). The results were compared to a dataset of 15 donkeys and 183 domesticated horses representing nine different breeds (43 Arabian, 19 Quarter, 24 Warm Blood, 20 Pony, 15 Thoroughbred, eight Cold Blood, one Andalusia, one Tennessee Walker, and 52 "local" horses). We targeted twelve published microsatellite loci: ASB2, ASB17, ASB10 [45], LEX65 [46], NVHEQ21 [47], SGCV33 [48], HTG7 [49], HMS6 [50], HTG4 [51], COR062 [52], LEX68 [46], HMS1 [50]. Fluorescent genotyping with dye-labelled primer sets was performed on either an ABI-3100 or ABI-3730 capillary system. Gene scans were inspected using the Genotyper software (Applied Biosystems). Population genetic analyses were implemented via GenAlEx 6.5 [53] and Structure 2.3.1 software [54–57]. In Bayesian Structure 2.3.1 analyses, the final optimal k value, k=5, was inferred using Evanno's Δk method [58] implemented in STRUCTURE HARVESTER [59] in which biological relevance was also considered. Even though an increase in k values could further separate the groups within the domesticated horse breeds, the structure of the Przewalski's horse cluster remained stable (**Figure S2D**). The results of the analyses of STRs are further discussed in section S3.3 of the Supplemental Information file available online.

Principal Component Analyses of horse SNPs genotyped in this study

We merged the files summarizing sequence variation (VCF files, including non-variant sites) for the Przewalski's horses, domesticated and Late Pleistocene horses with bcftools (version 1.1-57-g4f97f5b) only keeping biallelic variants on the autosomes without missing genotypes in any of the samples (command 'bcftools view -g "^miss" -q 0.00001') and strictly passing quality filters as described in the same section. We then converted the variant sites in the merged VCF file into plink format using VCFtools version 0.1.12 [60] to obtain a plink file containing 198,932 variants. The merged plink file generated for TreeMix analyses was used as input for PCA in the EIGENSTRAT program (version EIG5.0.1, [61]**)** with no outlier iterations (options '-numoutlieriter: 0'; **Figure S2E**).

Principal Component Analyses of a historical Przewalski's horse and horse SNPs genotyped in McCue et al., 2102 and Petersen et al., 2013

Previously, 744 present-day horses from 33 breeds, as well as eight Przewalski's horses were genotyped using the Illumina SNP50 Beadchip designed to target 54,000 SNPs ([1, 62]). As part of the present study, one historical Przewalski horse specimen (EP7, museum specimen 5217 ZI S-Pt-3093, 130 years old) was analysed on the same Equine SNP50 microarray. Genotyping on the microarrays was performed according to standard protocols provided by the manufacturer (Illumina®) and SNPs were filtered based on GenTrain scores of <0.6. In order to investigate genome-wide structure among present-day horses, we compared the individuals sequenced in our study (Late Pleistocene, domesticated, historical and present-day Przewalski's horses) with the extensive horse SNP dataset available. We excluded two sequenced Przewalski's horse individuals SB615 (KB4064) and SB293 (KB7413) as they were already genotyped in [1].

To achieve this, we applied a principal component analysis on a dataset constructed by merging Petersen et al. and McCue et al.'s genotypes with genotypes retrieved from the complete genome sequences that we generated in this study. For each locus genotyped in [62], we extracted the corresponding genotypes from the VCF files obtained for the horses sequenced here using bcftools version 1.1-57-g4f97f5b [24]. The extracted VCF file was converted into plink format using vcftools version 0.1.12 [60]. The plink file was merged with the dataset from [62] using plink [44], where sites with non-biallelic genotypes (37 sites) were removed from further analyses. In addition, we added to the merged dataset, eight Przewlski's horses from [1] (BL12822,

KB3063, KB3838, KB4064, KB4541, KB6842, KB7413 and KB7838) and the ancient individual EP7 genotyped in this study. The merged plink file was later filtered with 0.05 as cut-off for the maximum missingness per locus (--geno 0.05) using plink and restricting to sites called in EP7, which resulted in a set of 20,949 sites. Furthermore, via IBD estimation with plink (--genome), we found two pairs of Franches-Montagnes individuals Mon_FM0431/FM6 and Mon_FM2218/FM21 with high pihat values (0.9948 and 0.9947, respectively). To avoid spurious inference due to the likely duplication of the genotypes, the genotypes obtained from the genomes of samples Mon_FM0431 and Mon_FM2218 sequenced here were removed from this analysis. The final merged plink file was used as input for PCA in the EIGENSTRAT program (version EIG5.0.1, [61]**)** without outlier iterations (option '-numoutlieriter: 0'; **Figure S2F**).

Principal Component Analyses of horses sequenced in this study using genotype likelihoods

For ancient genomes, the combination of low coverage and higher error rate due to post-mortem DNA damage makes genotypes particularly difficult to determine accurately. To account for this, we compared historical Przewalski's horses, and present-day domesticated and Przewalski's horses using a genotype likelihood-based approach as implemented in ANGSD [63]. Briefly the method relies on a maximum-likelihood framework for Generation Sequencing data alignments, which condenses the mapped reads and quality scores into genotype likelihoods under the SAMtools model [24]. The genotype likelihood represents the probability of the sequencing data given a genotype in a particular individual, at a particular site under a genotype likelihood model [63]. To estimate the posterior genotype probabilities, we used priors for allele frequencies obtained by maximum likelihood estimates of the genome-wide site frequency spectrum. Here, genotype likelihoods were estimated using the "angsd" function, assuming a Hardy-Weinberg equilibrium, and considering sites with a minimum mapping quality of 25 (-minMapQ), a minimum base quality of 20 (-minBQ), and for which data was available in at least two individuals (-minInd). Genotype likelihoods were used to compare historical and present-day Przewalski's horses, present-day domesticated horses and Late Pleistocene horses via PCA. This was achieved using the function "ngsCovar" in ngsTools [64], which implements a probabilistic approach to estimate the genotype covariance matrix [65] and default parameters (minor allele frequency cutoff of 0.01: option '-minmaf 0.01'; **Figure 1C**).

Results of PCA are further discussed in section S3.8.1 of the Supplemental Information file available online.

Horse segmental duplications

We identified segmental duplications using mrCaNaVaR v0.31 and mrFAST v2.0.0.5 [66]. BED files were manipulated using BEDTools v2.0.17 [25]. The analysis was carried out using the Thoroughbred horse (Twilight) and the modern Przewalski's horse samples. We exclude samples Prz_D2630 and Prz_D2631 obtained from [3] to avoid biases due to differences in sample preparation and sequencing methods.

We first masked the reference sequence to account for existing duplication in the EquCab2.0 reference genome. To do this, we generated overlapping 36 bp kmers across the entire nuclear genome, using a step size of 5 bp. The genome was then indexed using mrFAST (using a window-size of 12), the 36-mers were mapped back to this genome, and the location of any 36-mer mapping to more than one region was masked in the genome. We additionally masked the regions corresponding to the UCSC repeat-tracts ("Interrupted repeats", "Repeat Masker", and "Simple repeats"). Consequently, a total of 1.2 Gbp of the reference genome was masked, representing ~49% of the EquCab2.0 reference genome.

The masked FASTA reference genome was then indexed using mrFAST as above. We next split the reads of the genomes of interest into non-overlapping 36-mers, starting 10 bp from the 5'-end, thereby excluding the more variable (quality-wise) 5'- and 3'-regions of the read. These were

mapped to the masked FASTQ using mrFAST as described above. Before investigating the copy-number variation in our samples, we expanded the masked regions in the FASTQ sequences by 36 bp, to account for lower coverage in regions proximate to the masked regions, and indexed the resulting FASTQ sequences using mrCaNaVaR, specifying no gaps.

We called read-depth and copy-numbers for each window in each sample using mrCaNaVaR and the alignments generated by mrFAST as described above. We defined as segmental duplications any regions consisting of at least five continuous copy-windows ("CW"), covering at least a total of 10,000 bp. In such CW, the copy-number of each window was required to exceeded the mean by at least 3 times the standard-deviation (calculated using the control regions identified by mrCaNaVaR), while allowing one non-trailing window, which exceeded the mean by at least 2 times the standard-deviation. The mean and standard deviation were calculated separately for the X chromosome of male individuals.

Prior to collecting segmental duplications, we excluded putative mis-assembled regions in which the copy number exceeded 50 times the ploidy of the given chromosome in any sample. We furthermore excluded any region in which segmental duplications were present in the Thoroughbred horse (Twilight), in order to account for segmental duplications present in other domesticated horses and duplications not accounted for in the EquCab2.0 reference genome. This resulted in the exclusion of 176 million bp. We called segmental duplications in the Przewalski's horses using the same method as described above.

We excluded from the following analyses samples SB281, for which the distribution of copy-numbers called for control regions for each sample (excluding the X chromosome) deviated from the remaining samples, in order to ensure that results were comparable across samples. We also excluded the hybrid horse KB7903 from further analyses. We next concatenated the filtered BED files and calculated the number of individuals for which a given region was called using the BEDTools command 'genomecov' with arguments '-bg', and determined overlap with annotated sequences from Ensemble v76 ([38], excluding pseudo-genes and introns) using the BEDTools intersect command with arguments "-loj". For each Ensembl v76 transcript, we recorded the maximum number of samples for which a given segmental duplication overlapped that transcript at any window. We carried out functional enrichment on genes located in Przewalski's horse segmental duplications as described below.

Functional enrichment

Functional enrichment was carried out on the online platform WebGestalt (WEB-based Gene SeT AnaLysis Toolkit; [67]) using humans and mice as model organisms after converting unique horse Ensembl Gene Identifier using a correspondence table generated with the Biomart from Ensembl. We used different databases to investigate enrichment of:

- Pathways: KEGG (03/21/2011), WikiPathways (11/11/2012),
- Phenotypes: Mammalian Phenotype Ontology (for mouse gene ID, 04/10/2013) and Human Phenotype Ontology (for human gene ID 04/10/2013),
- Phenome-Wide Association Study: PheWAS catalog (05/20/2014; for human gene ID only),
- Disease associated genes: GLAD4U (1/26/2013, for human gene ID only).

We used default parameters and Benjamini-Hochberg correction for multiple tests [68], and retained the top-10 significant enrichment clusters.

Genotyping of ancestral informative markers

We collected potential ancestry informative markers using the criteria based on those described in [69]. We excluded the ancient, low-quality Przewalski's horses due to the higher error-rate, and limited the domesticated horses to those included in the phylogenetic analyses described below, in order to ensure an even representation of breeds. We furthermore excluded the

Przewalski's horse hybrid from the analyses. As a result, the final dataset analysed comprised 14 Przewalski's horses (KB7674, KB7713, KB7848, SB281, SB293, SB339, SB4329, SB533, SB615, KB3879, KB7856, Prz_D2630, Prz_D2631, Prz_Przewalski) and 18 domesticated (Qrt_A5659, Qrt_A5964, Mon_FM1030, Mon_FM1785, Mon_FM1932, Mon_FM1951, Mor_EMS595, Std_M977, Std_M5256, Std_Standardbred, Arabian, Fjord, Icelandic, KB7754, Mng_D2628, Mng_D2629) and the two Late Pleistocene horses (CGG10022, CGG10023). We did not filter SNPs by physical proximity (unlike [69] who suggest a minimum distance of 0.3 cM in humans), to accommodate downstream analyses for which different marker densities may be required.

We only kept sites where genotyping could be performed in:

- all, or all but one, Przewalski's horses,

and

- all, or all but one, or all but two, Late Pleistocene and domesticated horses (or all but two for domesticated and Late Pleistocene horses),

and among these sites, those for which one allele was found in:

- all, or all but one, Przewalski's horses,

and in:

- none, one, or two domesticated horses.

We carried out functional enrichment, as described above, for the genes where these SNPs were found (**Table S2**).

Selection scans using F-$_{ST}$ estimates

We estimated the F-statistics (F_{ST}; [70]) between the domesticated and Przewalski's horse population across the genome as a proxy for genetic differentiation. We compared present-day Przewalski's horses with two sets of domesticated horses: one set containing 29 samples (DOM) and one set containing 12 Franches-Montagnes horses (FM). Descriptions of the panels are given in Table S6.1 of the Supplemental Information file available online.

F_{ST} was estimated on a genome-wide scale using the ngsFSt command in the ngsTools package [64, 65], which implements a method-of-moments estimator of F_{ST} that accounts for genotype uncertainty and uses a 2D-SFS as prior information [65]. F_{ST} was computed in 50,000 bp sliding windows with a step size of 10,000 bp. A genotype likelihood model similar to the one implemented in SAMtools version 0.1.18 [24] was used and the posterior probability of a site being variable was estimated for every position of the genome. For each group comparison, we calculated the 2D Site Frequency Spectrum using the command realSFS 2dsfs from ANGSD. The F_{ST} value the two groups considered was then calculated for each genomic position considered in the 2D SFS calculation and using the ngsFst command from the ngsPopGen package [65]. Fst values within sliding genomic windows were then calculated using the per site information retrieved from ngsFst and disregarding windows where coverage was less than 90%.

Those genomic regions showing a local and significant increase of the F_{ST} value were regarded as candidate regions for selection (outliers selection scans). We used two different approaches to detect those regions as described in detail in [23] (see their Supplemental section 7.2). In the first approach ("Gene-max"), genes were ranked according to their F_{ST} value and the outliers in the top 10% gene of the resulting ranked list were extracted. This provided a total number of 2,509 gene candidates underpinning population differentiation. In the second approach ("Region-peaks"), we used a smoothing procedure to define those regions and to obtain a local maximum of F_{ST} values (peaks) and retained as outliers the top 10% regions showing local maxima. This, for "Region-peaks", 2,847 and 2,887 genes depending on whether the DOM and FM panels were used for domesticated horses, respectively.

From each method ("Gene-max" and "Region-peaks"), we obtained the overlapping gene IDs from Ensembl version 76 [38] and we assigned the corresponding mean F_{ST} value, using

IRanges as in [23]. Genes identified by F_{ST}-based selection scans are listed in Table S6.2 of Supplemental Information file available online. We then performed functional enrichment as described above (**Table S2**).

Selection scans using SweeD

We carried out selection scans on each chromosome using the program SweeD [71]. Genomic regions under selective sweeps were detected as regions where ancestral alleles show significantly lower frequencies (corresponding to lower a-parameters in SweeD) than expected under a neutral model. The observed frequency distribution was determined from the Przewalski's horses genomes, whereas the frequency distribution under the neutral model was obtained through simulations using MSMS [72] under a demographic model selected beforehand through Approximate Bayesian Computation (ABC).

ABC demographic model selection for MSMS simulations

In order to determine the parameters of the demographic models to be used in MSMS simulations, we simulated four different demographic models before selecting the one best fitting the data using ABC. Such demographic models were constructed based on the results of PSMC (see "Demographic reconstruction based on nuclear genome information"). Firstly, we determined the age of the effective population size minima and maxima in the main PSMC profile and located these at the midpoint of the corresponding segment, after merging adjacent segments with identical population sizes. We next calculated the population size at each minima and maxima as the population size of the segments overlapping these averaged across all bootstraps. We defined t1 as representing the beginning of the captive population of Przewalski's horses, 8 generations ago (corresponding to roughly 64 years BP using 8 years as the generation time). Unless otherwise specified, we simulated 1,000 sequences of 10 kb for each set of parameters and for each model. Graphical descriptions of the tested models are given in Figure S6.1 of the Supplemental Information file available online.

From the PSMC demographic scaffold, we derived four models for simulations:

- Model PSMC #1, which represents the basic model derived from the PSMC results, with a constant population size from t_2 to t_0. We simulated a total of 1 million sequences using this model.
- Model PSMC #2, which corresponds to PSMC #1, except that the population size at t_2 was set to 1,000, 2,000, 3,000, …, 10,0000, and values 10,000, 20,000, 30,000, …, 10 million, with a constant population size from t_2 to t_0. We simulated a total of 1 million sequences using this model.
- Simple Model, in which we modelled exponential population growth from t_2 to t_0, ending with population sizes of 1,000, 2,000, 3,000,…, 10,000, and values 10,000, 20,000, 30,000, …, 10 million.
- Complex Model, in which we modelled exponential population growth from the t2 to t_1, followed by exponential decline to t_0. This model was built to possibly account for the recent historical bottleneck documented for Przewalski's horses. We considered population sizes 1,000, 10,000, and 50,000 to 10 millions in increments of 50,000, while requiring that the population size at t_0 be strictly less than the population size at t_1. We simulated 100 sequences using each set of parameters.

We calculated the scaled mutation rate θ as $4N_0 \times \mu \times N_{sites}$, and the scaled recombination rate ρ using the θ/ρ ratio estimated by PSMC. Each model was simulated in MSMS, and from the simulation output, a range of population statistics was calculated using 'mstatspop' (available at http://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html). These included the total number of variants, Watterson's theta (and its Achaz correction), Tajima's theta (and its Achaz correction), Fu and Li's theta, the haplotype diversity (HapW), Tajima's D, Fu and Li's D, Fu and Li's F, Achaz's Y, Fs, R2, the MD standard deviation, skewness and kurtosis and

the nucleotide diversity (Pi W). The matrix of summary statistics was then reduced to a limited number of dimensions using PCA and the first two or three principal components, which explained a minimum of 97.5% of the total variance (data not shown).

For ABC-based model selection, the same population statistics were calculated from the EquCab2.0 reference nuclear genome: we first computed 10 kb windows across EquCab2.0, starting and ending 100 kb from chromosome termini. We next collected pairs of adjacent windows in which at least 10,000 sites were called in all samples, selected the first 10,000 sites called, converted these to MS format and calculated population statistics using 'mstatspop'. We furthermore required a minimal distance of 100 kb between the start of the present pair of windows and the end of the previously selected pair of windows in order to avoid linkage between adjacent windows, and considered the mode for each distribution of summary statistics as the most likely observed value. These values were converted to predicted values according to the PCA performed on the simulated data of each model, and the two or three first corresponding components were used as transformed observed values. ABC was performed using the 'abc' R package [73] using local linear regression and accepting 1,000 simulations. Model comparison was performed using the 'postpr' function of the 'abc' R package and accepting 1,000 simulations, following PCA transformation of the simulated summary statistics across all models, and prediction of the observed values based on the three first principal components (which explained 98.7% of the total variance). The PSMC #1 model was identified as the demographic model best fitting the observed data by calculating posterior model probabilities and Bayes Factor (0.1819-0.2409). We choose this model as the most likely model for further analyses in SweeD.

Selection scan using SweeD

VCF files containing the genotype calls for the Przewalski's horses were converted into MS files by random-sampling heterozygous sites and considering all non-reference alleles as derived. The allele frequency distribution under the neutral model was obtained by simulating 30,000 blocks of size 5 million bp under the Model PSMC #1 (see above) using MSMS. The distribution of the alpha-value from the Przewalski's horse and simulated genotypes were obtained through the SweeD run with a grid size of 10 kb. We identified candidate sites for selective sweep on the basis of significant Qvalues, i.e., p-values after correction for multiple testing using a False Discovery Rate of 0.05 (calculated in R using the "p.adjust" function, "BH" method, [68]). We identified those annotated sequences from Ensembl v76 [38] that overlap with the selected regions as determined by SweeD. Genes identified by SweeD-based selection scans are listed in Table S6.2 of Supplemental Information file available online. We then performed functional enrichment as described above (**Table S2)**.

Joint demographic inference using dadi (Diffusion Approximation for Demographic Inference)

We refined the recent demographic history of the two populations of domesticated and Przewalski's horses using the program dadi (version 1.6.3), where the demographic parameters are estimated based on the diffusion approximation to the SFS [74]. We first estimated the 2-Dimension SFS (2D-SFS) between Przewalski's horses and domesticated horses, where the Przewalski's horse dataset was composed of the 14 present-day individuals and the domesticated dataset comprised of either 12 Franches-Montagnes horses (FM-DOM) or 16 other domesticated horses (OT-DOM). Descriptions of the panels are given in Table S5.1 of the Supplemental Information file available online.

To accommodate for different sequencing depths among samples, we used ANGSD [63] and obtained a Maximum Likelihood estimate of the 2D-SFS, taking sequencing data uncertainty into account. We filtered sites with a mapping quality < 30, a site quality < 20, a global minimum depth < 120, a global maximum depth > 1,000, and with data for at least 10 individuals. Only autosomal chromosomes were considered. The spectra were polarised following a strict identity

rule on the genomes of a range of non-horse equids sequenced so far (i.e. only bases found in all those genome are considered ancestral - other cases are recoded as Ns). 1,365,669,235 sites were considered for the analysis between PRZ and FM-DOM, and 1,365,665,795 sites for the analysis between PRZ and OT-DOM.

We fixed the demographic history before the split between Przewalski's horses and domesticated horses using the model inferred by PSMC analyses above on the sample with the highest sequencing depth (SB281). The ancestral effective population size was therefore set to 157,602 and it changed over time. To model recent size changes at the split time, we used as template the model proposed in [4], specifically imposing an exponential decay of size until 3,000 from the divergence time to 10,000 years ago. At that time populations then instantaneously changed to a separate sizes, which remained constant until present. Analyses were run considering the same generation time and using the mutation rate as for the PSMC analyses described above.

We tested different models of population split between Przewalski's horses and domesticated horses, each one having a distinct set of free parameters to be estimated:
- Model A: no gene flow;
- Model B: symmetric gene flow between PRZ and FM-DOM/OT-DOM;
- Model C: asymmetric gene flow between PRZ and FM-DOM/OT-DOM;
- Model D: gene flow only from PRZ to FM-DOM/OT-DOM;
- Model E: gene flow only from FM-DOM/OT-DOM to PRZ.

Tested demographic models are also described in section S5.2 of the Supplemental Information file available online.

We then allowed for different levels and directions of gene flow, from the divergence time to present time, between populations. For each model, we ran the program 20 times from different initial values to ensure convergence, and retained the parameters set with the highest likelihood. **Table S3** show estimates of demographic parameters for all tested models using FM-DOM and OT-DOM as domesticated horses, respectively.

Finally, we tested whether gene flow occurred continuously from the divergence time to the present, or if it stopped at some pre-selected point in the past. In this analysis, we used parameters estimated in model C (**Table S3**), which was most-likely scenario for both sets of domesticated horses. We let divergence time and gene flow parameters free to be re-estimated, while setting the time for the putative end of gene flow fixed at different values. These times were chosen to represent some events that could have had an impact on breeding opportunities between PRZ and domesticated horses, and are:

a) 26,000 years ago at the beginning of the Last Glacial Maximum (LGM);
b) 19,000 years ago at the end of the LGM;
c) 11,700 years ago at the beginning of the Holocene;
d) 9,000 years ago at the beginning of the Holocene climatic optimum;
e) 5,500 years ago at the beginning of horse domestication;
f) 2,000 years ago at the beginning of the Iron Age;
g) 200 years ago at approximately the discovery of the Przewalski's horse population by the Western world.

Exploring demographic history using projections of test genomes onto reference populations

We used the genome projection method, where test genomes are projected onto reference populations as described in [75] to explore models of demographic history for the domesticated and Przewalski's horses. We used genomes obtained from 15 present-day Przewalski's horses and 28 domesticated horses representing a range of domesticated horse breeds and types. Horse genomes were used either as test genomes, or grouped into reference panels: refPRZ for the 14 non-hybrid Przewalski's horse genomes, refFM for the 12 Franches-Montagnes horse genomes, and refDOM for the full range of 28 domesticated horses, including Franches-Montagnes horses.

Descriptions of the panels are given in Table S5.5 of the Supplemental Information file available online.

The projection, *w*, as described in [75], explores how different a test genome is from a set of genomes from a reference population (reference panel). For sites belonging to each derived allele frequency category of the reference panel, the expected number of sites with a derived allele in a genome from the reference population can be calculated. The observed number of test-genome sites with a derived allele is compared to the expected number to determine the projection. In the projection, a value greater than one means that it is more likely than expected that a derived allele will be observed in the test genome for that frequency category, and a value less than one means it is less likely than expected that a derived allele will be observed in the test genome for that frequency category. A member of the reference population would give a projection of $w = 1$ for every frequency category. Depending on the demographic relationship between the test and reference populations, the projection may deviate from $w = 1$ in a variety of ways.

To calculate the projection, we retained sites that were variable in at least one of the genomes sampled. The projection was determined for every individual against one of three reference panels, refPRZ, refFM and refDOM, for sites that are polymorphic in the reference panel. Based on results described in section S5.3 of the Supplemental Information file available online and to avoid the effects of population structure within the reference panel, we focused on the projections for refFM, rather than refDOM, in further analyses.

We used PSMC estimates of past horse population size and the Maximum Likelihood estimates calculated in dadi from each of the five population models described above (see "Joint demographic inference using dadi (Diffusion Approximation for Demographic Inference)"). For each model, we determined the expected projection using *fastsimcoal2* [76]. We compared these projections to the observed projection using the sum of least squares (LSS) score to assess how well each model recovers the projection. The observed projections were determined using SB159 as the test Przewalski's horse genome (this genome shows minimal, if any, admixture with domesticated horses), and Mon_FM0450 as the test Franches-Montagnes genome. refPRZ and refFM were used as the reference panels.

Using the model that fits best according to dadi analyses, we considered whether an improved fit of the model could be obtained considering changes in population decay and migration rate. We used *fastsimcoal2* [76], Brent's algorithm and the LSS score to vary one model parameter at a time and improve the fit. Simulated parameters are reported and the LSS scores for the adjusted models C are reported in **Table S3**. Four models were considered. Model C1 only considered a single asymmetric migration, the rate of which was estimated here. Model C2 changed the size of the exponential population decay and the effective population size of Przewalski's horse, allowing a single asymmetric migration rate as in Model C (Model C2_1) and allowing the asymmetric migration rate to vary (Model C2_2). Model C3 considered no population size changes but allowed two different asymmetric migration rates in the past and present, in order to test for potential consequences of past climatic changes. We allowed the migration rate to change once, keeping it constant before and after the time of migration rate change. Finally, Model C4 allowed both population size changes similar to that in Model C2 and two different asymmetric migration rates, plus a pulse of admixture from Franches-Montagnes to Przewalski's horses at some recent time in the past.

Projection plots, can be found in section S5.3 of the Supplemental Information file available online, where genome projections are also used to provide further evidence for population structure within horses and domestic admixture in Przewalski's horses.

Average genome-wide heterozygosity

For each horse described in this study, the average heterozygosity was estimated as the number of segregating sites by calculating the Watterson estimator theta [77]. This was done using the program ANGSD ([63]; http://popgen.dk/wiki/index.php/ANGSD) that allows accounting for the uncertainty in the genotyping due to low genome coverage by using genotype likelihoods and priors based on site frequency spectrum (SFS). Priors for the autosomes were constructed using the SFS [78] for chromosome 22. The average heterozygosity was calculated as the genome-wide average of the theta estimated in 50 kb overlapping windows with a step-size of 10 kb of the genome, as described in [4] (Supplemental Information S4.2). For high-quality genomes we excluded windows in which less than 45 out of 50 kb (90%) were covered. For lower quality/historical genomes we excluded windows in which less than 25 out of 50 kb (50%) were covered, as well as transitions in order to control for damage-derived misincorporations, mainly affecting C→T and G→A misincorporation rates in ancient and historical samples (**Table S1; Figure 3A**). Heterozygosity levels in modern individuals calculated including transitions are shown in **Figures S3.9-10** of the Supplemental Information file available online.

The vast majority of present-day horses showed reduced heterozygosity values compared to Late Pleistocene horses (**Figure 3A**). The average heterozygosity values were lower in modern Przewalski's than in domesticated horses (Mann-Whitney test, p-value = 1.88×10^{-3}) and historical PHs (Mann-Whitney test, p-value = 8.40×10^{-4}). The latter could reflect the genetic bottleneck associated with captivity, or the increased error rate of our historical genomes. However, considering only transversions (for which the error rate is reduced; **Table S1**), we note that the historical Paratype specimen displays a higher heterozygosity than one domesticated Icelandic horse ($\log\text{-}\theta_w$ = 3.61 vs 3.39) showing similar per base error rates (0.099% vs 0.094%). As its heterozygosity was also higher than in any modern Przewalski's horse, this suggests that the extreme demographic bottleneck experienced in the first half of the 20th century has reduced the heterozygosity within contemporary Przewalski's horses, in line with pedigree–informed reconstructions.

Genome-wide inbreeding coverage

We estimated inbreeding in Przewalski's horses as the proportion of genomic fragments being mostly homozygous, or "homozygous-by-descent" (HBD), following terminologies and methodologies similar to those in [79] and applied to complete horse genomes in [4]. We excluded the historical horses because their low sequencing depth-of-coverage was not compatible with the detection of inbreeding tracts.

We used the Watterson's theta estimator across the genome (see section "Average genome-wide heterozygosity") to detect HBD regions. In these analyses, we replaced missing data by the closest neighbouring valid values on a per-chromosome basis. We determined the coordinates of the HBD genomic tracks by identifying segments with local changes in the average estimates for Watterson's theta along the genome using the R package 'changepoints' ([80]; http://cran.r-project.org/web/packages/changepoint/index.html) and the binary segmentation algorithm (method="BinSeg") allowing up to 10 breakpoints per chromosome (Q=10). For each segment, we calculated the heterozygosity estimates given by the average of (log(theta)), which were used to weight genomic track lengths (deduced from their external coordinates). Their density was then plotted with the 'density' R function (from 100,0000 points) and a bimodal distribution of the Watterson's theta estimator values indicated inbreeding in the corresponding individual. We determined the coordinate of the lowest point (pit) between the two modes using numerically estimates of first and second derivatives of the density function (using absolute tolerance of 10^{-7}). The pit coordinate provided a threshold used to classify genomic segments as showing "high" or "low" heterozygosity (vertical red line). The total size of the "low" heterozygosity regions was divided by the total size of all regions to give the proportion of HBD

tracks, or "inbreeding coverage estimate". The average of log(theta) values and positions of the genomic tracks were determined both including and excluding transitions to account for misincorporations arising from post-mortem DNA damage in ancient genomes (**Figure 3B**). Genomic track length distributions including and excluding transitions are shown in **Figure S3.11** of the Supplemental Information file available online.

F-inbreeding coefficients

We also directly estimated inbreeding coefficients in a genotype likelihood framework to accommodate low-depth genomic data, as the method presented in the previous section could not be applied on our set of historical genomes, as those were sequenced at relatively low depth-of-coverage. We used a new method based on that described in [81], where the probability given the data that each position is Identical By Descent (IBD) is calculated as well as the site allele frequency (F-inbreeding coefficient). Here, a Hidden Markov Model (HMM) that, apart from estimating genome-wide inbreeding level, also attempts to infer the location of IBD tracts. Assuming a model with two states (IBD and not IBD) and three possible observations (homozygous reference, heterozygous and homozygous alternate, all expressed in genotype likelihoods), the probability that a site is IBD calculated and the most likely path for the observed data is found. The per-individual inbreeding coefficient, F-HMM, is the average of the per-site IBD probabilities. We only analysed a randomly sampled subset of 0.1% of the sites in order to avoid the confounding effects of linkage. We ran analyses separately for domesticated and Przewalski's horses so that biases due to population structure that can artificially inflate inbreeding estimates (known as the "Wahlund effect") could be reduced. All analyses were run until convergence, which was reached when the average likelihood difference was $<10^{-7}$. Each step was replicated 100 times in order to avoid convergence to local maxima (**Figure 3B**). Comparisons between inbreeding coverage, F-, F-HMM, and pedigree-based inbreeding coefficient are discussed in section S3.6 of the Supplemental Information file available online.

NGSAdmix admixture analyses

Genomic structure within modern and ancient horses was investigated using the function "NGSAdmix" implemented in ANGSD that handles genotype likelihoods in a Maximum Likelihood framework [82]. The analyses were run considering two to 20 ancestral populations (-K), a minimum minor allele frequency of 5% (-minMaf) and a maximum number of Expectation Maximization iterations of 5,000 (-maxiter).

We retained K=8 as the optimal number of ancestral populations, as it allows retrieving the known structure that defines the present-day domesticated horse breeds (**Figure 4A** and **S3A**), including: one cluster for Standardbred, one for Franches-Montagnes, one for Arabian, Thoroughbred and Quarter horses in accordance with the latter two descending from Arabian horses, two clusters for Mongolian (including the historical horse Emgl1), Norwegian Fjord and Icelandic horses in line with previous phylogenetic results [2], while the Morgan horses appear as a mix of five other domesticated groups. One additional cluster was identified to describe Late Pleistocene horses CGG10022 and CGG10023, whereas two clusters were found to be unique to the Przewalski's horse population. Full NGSAdmix clustering results for K=2-20 are presented in section S4.1 of the Supplemental Information file available online.

Gene flow between Przewalski's horses and Mongolian horses using the f_3-statistics.

In order to test for admixture between Przewalski's horses and Franches-Montagnes or Mongolian horses in the ancestry of each single Przewalski's horse, we calculated f_3-statistics [83] using the 'threepop' program (-k 5000) provided in the TreeMix package. For each trio, we merged and filtered the VCF files with bcftools (-g "^mis –q 0.00001"), then converted the

merged VCF file to plink using vcftools (version 0.1012). Finally, we converted the plink files to the TreeMix format using the supplied python script ("plink2treemix.py").

Here we use the same notation as [83], where the test population C is represented by each Przewalski's horse individual, and the ancestral populations A is a Przewalski's horse individual, and B, is either the Franches-Montagnes or the Mongolian population. Following [83], negative f_3(C;A,B) values suggest that C has contribution from both ancestral populations A and B Test significance was evaluated on the basis of adjusted p-value, which are Z-scores corrected for multiple testing using the function 'p.adjust' with the Holm correction [84] in R (**Table S4**).

Gene flow between Przewalski's horses and other horse populations using the D-statistics.

We investigated potential gene flow between Przewalski's horses and present-day domesticated horses, as well as between Przewalski's horses and Late Pleistocene horses using D-statistics [85, 86] and *E. asinus somalicus* Shakira [23] as an outgroup. We applied the test to our horse genome dataset as described in the Supplemental materials S12.1 of [2]. We excluded from our tests the genome of the Thoroughbred horse Twilight in order to avoid biases due to the fact that this genome was used as reference genome (EquCab2.0), thus making Twilight significantly closer to the reference than any other horse.

Briefly, using the framework and notation from [2], one outgroup is selected, as well as three test-taxa defined as H1, H2 and H3. We test the null hypothesis that the topology ((H1, H2), H3), outgroup) is correct, and in the case it is rejected, the outcome of the test is compatible with the presence gene flow between H3 and H1, or between H3 and H2. This is also compatible with the presence of population sub-structure and/or admixture with an un-sampled and presently unknown ghost population [86]. Note also that the calculation of the D-statistics can be significantly affected in cases where H1 and H2 show strong differences in error rates (for example if H2 has a relative high error rate compared to H1, it will make a surplus of BABA events, by reverting the most common BABA sites, potentially leading to negative D-statistics, supporting gene flow between H1 and H2 [2, 86].

Our implementation of the ABBA-BABA test is described in [2], where we estimated the standard error of the D-statistics using "*delete-m Jackknife for unequal m*" based on 10 Mbp blocks, chosen to accommodate for the large linkage disequilibrium in horses [5]. The Z-score were calculated as a measure of significance; absolute values greater than 3 are usually used as an indication of statistically significant deviation from the null hypothesis. However, in order to control the family-wise error rate arising from the large number of tests performed, we corrected the Z-scores for multiple testing using the function 'p.adjust' with the Holm correction [84] in R. A test was considered significant, if it showed an adjusted p-value smaller than 0.05. Finally, in addition to ABBA-BABA tests based on sites where both transitions and transversions were observed, the analyses were run taking into consideration transversions only in order to reduce the effect of base error in the data when using ancient genomes.

We tested the following tree topologies:

- *((Domesticated1, Domesticated2), Przewalski), Outgroup,* where "Domesticated1" and "Domesticated2" are two different domesticated horses, in order to detect gene flow between present-day/historical Przewalski's horses and domesticated horses. The results of these tests suggest that Przewalski's horses are equidistant to all domesticated horse types.
- *((Przewalski1, Domesticated1), Domesticated2), Outgroup)*, where we verified that all domesticated horses were closer to any other domesticated horse than to any present-day and historical Przewalski's horse.
- *((Przewalski1, Przewalski2), Domesticated), Outgroup)*, where "Przewalski1" and "Przewalski2" are two different present-day Przewalski's horses (to reduce the impact of larger error rates on the data characteristic of ancient genomes), excluding the hybrids KB7903 and Ewld5_Theodor, in order to investigate potential stratification within the

Przewalski's horse population as a result of different amounts of gene flow with domesticated horses. Here, we showed that not all pairs of Przewalski's horses are equidistant to any given domesticated horse, with a large proportion of the tests showing a significant deviation from the tree topology and the D-statistic estimates displaying a large variance. Moreover, we could not identify particular domesticated horses showing more genetic affinity to Przewalski's horses than the others. These results suggest variable amounts of admixture with domesticated horses in the ancestry of Przewalski's horses.

- *((Paratype, Przewalski), Domesticated), Outgroup),* where "Przewalski" was any present-day Przewalski's horse, except the hybrid horse KB7903 in order to identify which, if any, domesticated or Przewalski's horse type has been more involved in the gene flow detected using the test above. The rationale behind this test was that any pre-captivity Przewalski's horse captured in the wild is hypothesized to show a minimal genetic contribution from domesticated horses (or at least represent the pre-captivity level of domesticated admixture in Przewalski's horses). Using one of those pre-captivity horses as H1 should then allow for detecting admixture in present-day Przewalski's horses. Here, we used the genome of the Paratype, as it is the pre-captivity Przewalski's horse sequenced with the highest average coverage (3.7-fold) and lowest error rate (0.10%). It is likely that the extent of significant gene flow from domesticated to Przewalski's horses is overestimated in the later tests owing to the use of the Paratype ancient genome, with a larger error rate, was used here as H1.

We also quantified the amount of domesticated ancestry in each Przewalski's horse. To achieve this, we used the results of the test of the two ((Przewalski1, Przewalski2), Domesticated), Outgroup) and tree topologies ((Przewalski1, Domesticated2), Domesticated), Outgroup), following the methodology presented in [83, 86, 87]. The rationale is to compare the difference in ABBA and BABA counts in both cases. As the latter represents a situation of total admixture (as two domesticated horses are lined up as H2 and H3), it offers a way to normalize the admixture level detected in the former tests, providing a minimal estimate for the admixture level observed. More specifically, we first calculated the difference Diff_P = nABBA – nBABA for the ((Przewalski1, Przewalski2), Domesticated), Outgroup) tree topology, where "Przewalski1" is Paratype, SB339, SB274, SB533 or SB293, and "Domesticated" a given domesticated horse. For the same Domesticated horse, we then calculated the values for the difference Diff_D = nABBA – nBABA for all the possible ((Przewalski1, Domesticated2), Domesticated), Outgroup) tree toplogies, where "Przewalski1" is one of the Przewalski1 horses listed above, and "Domesticated2" another horse belonging to the same horse type Mongolian, Franches-Montagnes, Standardbred, or Quarter horses as domesticated horses. For a given Przewalski1, we calculate, for each Przewalski2, the ratio Diff_P / Diff_D *100, which varies according to the Domesticated used as H3.

On the basis of the results of these analyses, we identified the horses SB339, SB274, SB533 and SB293 as those the least genetically impacted by domesticated horses, and we used their genomes as H1 (Przewalski1) to test the ((Przewalski1, Przewalski2), Domesticated), Outgroup) tree topologies (**Table S4**). In these tests, the impact of genome error rate is reduced by the use of high-quality genomes as H1. We confirmed previous results with horses SB528, SB524, SB966 being the most impacted by domesticated horses, and SB339, SB274, SB533 and SB293 the least impacted.

The results of the D-statistic tests are further discussed in section S4.3 of the Supplemental Information file available online, which also presents the results and discussion of the test for tree topologies (((Przewalski, Domesticated), Ancient), Outgroup), where "Ancient" was the CGG10022 or the CGG10023 ancient horses.

Chromosome painting by locus-specific inference

We used the program LAMP v2.5 [88], which estimates, for each individual, its maximum likelihood locus-specific ancestry given a number K=2 of ancestral populations (the alleles of a diploid individual originated either from the domesticated and Przewalski's ancestral populations). In the analyses, the Przewalski's horse population comprised 15 individuals (including the hybrid KB7903) and the full dataset of 28 domesticated horse genomes (**Table S1**). We considered that both populations have been mixing for eight generations, corresponding to the foundation of the Przewalski's horse pedigree. We assumed chromosome-specific recombination rates (set according to [89] and ranging from 0.58 Cm/Mb for chromosome 26 to 2.33 cm/Mb for chromosome 27), and an admixture proportion (a parameter) of 65% and 35%, i.e., approximately the proportions of domesticated and Przewalski's horse chromosomes in the joint SNP panel. As inconsistent introgression proportions were calculated when masking genotypes not showing maximum quality, i.e., equal to 99, the Prz_Przewalski horse, sequenced at a lower coverage (9.65-fold, **Table S1**), we excluded this horse from our analyses.

As in the Przewalski's horse population, the contribution from domesticated horses has not been equal in all Przewalski's lineages, in violation of the random mating assumption since the first admixture event in LAMP, we re-assigned chromosomal windows showing a deviation from the global the expected admixture proportions in order to avoid erroneous inference of locus ancestry. We also masked uncertain genomic regions, which may represent ancestral polymorphisms, or poorly assembled regions. The results of the LAMP analyses are further discussed in section S4.4.1 of the Supplemental Information file available online.

Mendelian trait genotyping

We determined whether alleles associated with Mendelian traits present in the Przewalski's horse population were found in genomic blocks introgressed from domesticated horse. Using the genotypes calls, we examined SNPs at 49 loci underlying Mendelian traits (e.g., diseases, coat coloration, other phenotypes; **Table S5**). These include traits collected in the Online Mendelian Inheritance in Animals (OMIA) database [90], in [2], and in [91] as described in [4]. An attempt was made to call SNPs at all loci for both modern and ancient samples as described above, and when this was not possible, the number of high-quality nucleotides (Phred base quality score >= 35) observed at the given site was recorded instead (**Table S5**). The results of the Mendelian trait analyses are further discussed in sections S3.5.2 and S4.4.6 of the Supplemental Information file available online.

Supplemental references

1. McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., Distl, O., Guérin, G., Hasegawa, T., Hill, E. W., et al. (2012). 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.

2. Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., et al. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature *499*, 74–78.

3. Do, K.-T., Kong, H.-S., Lee, J.-H., Lee, H.-K., Cho, B.-W., Kim, H.-S., Ahn, K., and Park, K.-D. (2014). Genomic characterization of the Przewalski′s horse inhabiting Mongolian steppe by whole genome re-sequencing. Livest. Sci. *167*, 86–91.

4. Schubert, M., Jónsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., Albrechtsen, A., Dupanloup, I., Foucal, A., Petersen, B., et al. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. Proc. Natl. Acad. Sci. *111*, E5661–E5669.

5. Lander, E. S., and Lindblad-Toh, K. with Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T. L., Adelson, D. L., Bailey, E., Bellone, R. R., et al., Broad Institute Genome Sequencing Platform and Broad Institute Whole Genome Assembly Team (2009). Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. Science *326*, 865–867.

6. Mohr, E. (1959). Das Urwildpferd (Equus przewalski Poljakoff 1881) Mit 87 Abbildungen Die Neue Brehm-Bücherei 249. (Wittenberg Lutherstadt: A. Ziemsen Verlag).

7. Volf, J., Kus, E., and Prokopova, L. (1991). General studbook of the Przewalski horse. (Prague (Czech Republic): Zoological Garden Prague).

8. Zimmermann, W. (2014). International Przewalski's Horse Studbook (Equus ferus przewalskii). (Cologne Zoo, SPARS version).

9. Seal, U., Lacy, R., Zimmermann, W., Ryder, O., and Princee, F. Captive Breeding Specialist Group.SSC/IUCN. (Apple Valley, MN).

10. Orlando, L., Metcalf, J. L., Alberdi, M. T., Telles-Antunes, M., Bonjean, D., Otte, M., Martin, F., Eisenmann, V., Mashkour, M., Morello, F., et al. (2009). Revising the recent evolutionary history of equids using ancient DNA. Proc. Natl. Acad. Sci. *106*, 21754–21759.

11. Ginolhac, A., Vilstrup, J., Stenderup, J., Rasmussen, M., Stiller, M., Shapiro, B., Zazula, G., Froese, D., Steinmann, K. E., Thompson, J. F., et al. (2012). Improving the performance of true single molecule sequencing for ancient DNA. BMC Genomics *13*, 177.

12. Der Sarkissian, C., Ermini, L., Jónsson, H., Alekseev, A. N., Crubezy, E., Shapiro, B., and Orlando, L. (2014). Shotgun microbial profiling of fossil remains. Mol. Ecol. *23*, 1780–1798.

13. Gilbert, M. T. P., Tomsho, L. P., Rendulic, S., Packard, M., Drautz, D. I., Sher, A., Tikhonov, A., Dalén, L., Kuznetsova, T., Kosintsev, P., et al. (2007). Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. Science *317*, 1927–1930.

14. Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. U. S. A. *110*, 15758–15763.

15. Seguin-Orlando, A., Schubert, M., Clary, J., Stagegaard, J., Alberdi, M. T., Prado, J. L., Prieto, A., Willerslev, E., and Orlando, L. (2013). Ligation bias in illumina next-generation DNA libraries: implications for sequencing ancient genomes. PloS One *8*, e78575.

16. Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajkovic, D., Kucan, Z., et al. (2009). Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science *325*, 318–321.

17. Schubert, M., Ermini, L., Sarkissian, C. D., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M. D., Fernández, R., Kircher, M., McCue, M., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. *9*, 1056–1082.

18. Lindgreen, S. (2012). AdapterRemoval: Easy Cleaning of Next Generation Sequencing Reads. BMC Res. Notes *5*, 337.

19. Xu, X., and Arnason, U. (1994). The complete mitochondrial DNA sequence of the horse, Equus caballus: extensive heteroplasmy of the control region. Gene *148*, 357–362.

20. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

21. Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., AL-Rasheid, K. A., Willerslev, E., Krogh, A., and Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. BMC Genomics *13*, 178.

22. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinforma. Oxf. Engl. *29*, 1682–1684.

23. Jónsson, H., Schubert, M., Seguin-Orlando, A., Ginolhac, A., Petersen, L., Fumagalli, M., Albrechtsen, A., Petersen, B., Korneliussen, T. S., Vilstrup, J. T., et al. (2014). Speciation with gene flow in equids despite extensive chromosomal plasticity. Proc. Natl. Acad. Sci. *111*, 18655–18660.

24. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R., 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

25. Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis: BEDTools: the Swiss-Army Tool for Genome Feature Analysis. In Current Protocols in Bioinformatics, A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo, and J. R. Yates, eds. (Hoboken, NJ, USA: John Wiley & Sons, Inc.), pp. 11.12.1–11.12.34.

26. Achilli, A., Olivieri, A., Soares, P., Lancioni, H., Kashani, B. H., Perego, U. A., Nergadze, S. G., Carossa, V., Santagostino, M., Capomaccio, S., et al. (2012). Mitochondrial genomes from

modern horses reveal the major haplogroups that underwent domestication. Proc. Natl. Acad. Sci. *109*, 2449–2454.

27. Lippold, S., Knapp, M., Kuznetsova, T., Leonard, J. A., Benecke, N., Ludwig, A., Rasmussen, M., Cooper, A., Weinstock, J., Willerslev, E., et al. (2011). Discovery of lost diversity of paternal horse lineages using ancient DNA. Nat. Commun. *2*, 450.

28. Goto, H., Ryder, O. A., Fisher, A. R., Schultz, B., Kosakovsky Pond, S. L., Nekrutenko, A., and Makova, K. D. (2011). A Massively Parallel Sequencing Approach Uncovers Ancient Origins and High Genetic Variability of Endangered Przewalski's Horses. Genome Biol. Evol. *3*, 1096–1106.

29. Jiang, Q., Wei, Y., Huang, Y., Jiang, H., Guo, Y., Lan, G., and Liao, D. J. (2011). The complete mitochondrial genome and phylogenetic analysis of the Debao pony (Equus caballus). Mol. Biol. Rep. *38*, 593–599.

30. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. PLoS ONE *7*, e34131.

31. Xu, S., Luosang, J., Hua, S., He, J., Ciren, A., Wang, W., Tong, X., Liang, Y., Wang, J., and Zheng, X. (2007). High altitude adaptation and phylogenetic analysis of Tibetan horse based on the mitochondrial genome. J. Genet. Genomics Yi Chuan Xue Bao *34*, 720–729.

32. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J., and Mclnerney, J. O. (2006). 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.

33. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

34. Iwase, M., Satta, Y., Hirai, Y., Hirai, H., Imai, H., and Takahata, N. (2003). The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. Proc. Natl. Acad. Sci. *100*, 5258–5263.

35. Wallner, B., Vogl, C., Shukla, P., Burgstaller, J. P., Druml, T., and Brem, G. (2013). Identification of Genetic Variation on the Horse Y Chromosome and the Tracing of Male Founder Lineages in Modern Breeds. PLoS ONE *8*, e60015.

36. Wallner, B., Vogl, C., Shukla, P., Burgstaller, J. P., Druml, T., and Brem, G. (2013). Identification of genetic variation on the horse y chromosome and the tracing of male founder lineages in modern breeds. PloS One *8*, e60015.

37. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. *17*, 368–376.

38. Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2013). Ensembl 2013. Nucleic Acids Res. *41*, D48–D55.

39. Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688–2690.

40. Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.

41. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.

42. Andersson, L. S., Larhammar, M., Memic, F., Wootz, H., Schwochow, D., Rubin, C.-J., Patra, K., Arnason, T., Wellbring, L., Hjälm, G., et al. (2012). Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. Nature *488*, 642–646.

43. Pickrell, J. K., and Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet. *8*, e1002967.

44. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575.

45. Breen, M., Lindgren, G., Binns, M. M., Norman, J., Irvin, Z., Bell, K., Sandberg, K., and Ellegren, H. (1997). Genetical and physical assignments of equine microsatellites—first integration of anchored markers in horse genome mapping. Mamm. Genome *8*, 267–273.

46. Coogle, L., and Bailey, E. (1999). Equine dinucleotide repeat loci LEX064 through LEX070. Anim. Genet. *30*, 71–72.

47. Bjørnstad, G., Midthjell, L., and Røed, K. H. (2000). Characterization of ten equine dinucleotide microsatellite loci: NVHEQ21, NVHEQ54, NVHEQ67, NVHEQ70, NVHEQ75, NVHEQ77, NVHEQ79, NVHEQ81, NVHEQ82 and NVHEQ83. Anim. Genet. *31*, 78–79.

48. Godard, S., Vaiman, D., Oustry, A., Nocart, M., Bertaud, M., Guzylack, S., Mériaux, J. C., Cribiu, E. P., and Guérin, G. (1997). Characterization, genetic and physical mapping analysis of 36 horse plasmid and cosmid-derived microsatellites. Mamm. Genome *8*, 745–750.

49. Marklund, S., Ellegren, H., Eriksson, S., Sandberg, K., and Andersson, L. (2009). Parentage testing and linkage analysis in the horse using a set of highly polymorphic microsatellites. Anim. Genet. *25*, 19–23.

50. Guékin, G., Bertaud, M., and Amigues, Y. (2009). Characterization of seven new horse microsatellites: HMS1, HMS2, HMS3, HMS5, HMS6, HMS7 and HMS8. Anim. Genet. *25*, 62–62.

51. Ellegren, H., Johansson, M., Sandberg, K., and Andersson, L. (2009). Cloning of highly polymorphic microsatellites in the horse. Anim. Genet. *23*, 133–142.

52. Tallmadge, R. L., Hopman, T. J., Schug, M. D., Aquadro, C. F., Bowling, A. T., Murray, J. D., Caetano, A. R., and Antczak, D. F. (1999). Equine dinucleotide repeat loci cor061-cor080. Anim. Genet. *30*, 462–466.

53. Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research--an update. Bioinformatics *28*, 2537–2539.

54. Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles: TECHNICAL ARTICLE. Mol. Ecol. Notes *7*, 574–578.

55. Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Resour. *9*, 1322–1332.

56. Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.

57. Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics *164*, 1567–1587.

58. Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. *14*, 2611–2620.

59. Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. *4*, 359–361.

60. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al., 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

61. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

62. Petersen, J. L., Mickelson, J. R., Rendahl, A. K., Valberg, S. J., Andersson, L. S., Axelsson, J., Bailey, E., Bannasch, D., Binns, M. M., Borges, A. S., et al. (2013). Genome-Wide Analysis Reveals Selection for Important Traits in Domestic Horse Breeds. PLoS Genet. *9*, e1003211.

63. Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics *15*.

64. Fumagalli, M., Vieira, F. G., Linderoth, T., and Nielsen, R. (2014). ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics *30*, 1486–1487.

65. Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., and Nielsen, R. (2013). Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data. Genetics *195*, 979–992.

66. Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O., et al. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. Nat. Genet. *41*, 1061–1067.

67. Zhang, B., Kirov, S., and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. *33*, W741–W748.

68. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

69. Baye, T. M., Tiwari, H. K., Allison, D. B., and Go, R. C. (2009). Database mining for selection of SNP markers useful in admixture mapping. BioData Min. *2*, 1.

70. Wright, S. (1965). The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. Evolution *19*, 395–420.

71. Pavlidis, P., Zivkovic, D., Stamatakis, A., and Alachiotis, N. (2013). SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. Mol. Biol. Evol. *30*, 2224–2234.

72. Ewing, G., and Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics *26*, 2064–2065.

73. Csilléry, K., François, O., and Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC): *R package: abc*. Methods Ecol. Evol. *3*, 475–479.

74. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet *5*, e1000695.

75. Yang, M. A., Harris, K., and Slatkin, M. (2014). The Projection of a Test Genome onto a Reference Population and Applications to Humans and Archaic Hominins. Genetics *198*, 1655–1670.

76. Excoffier, L., Dupanloup, I., Huerta-SÃ¡nchez, E., Sousa, V. C., and Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. PLoS Genet. *9*, e1003905.

77. Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. *7*, 256–276.

78. Korneliussen, T. S., Moltke, I., Albrechtsen, A., and Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics *14*, 289.

79. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., et al. (2013). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.

80. Killick, R., and Eckley, I. (2014). changepoint:an R package for changepoint analysis. J. Stat. Softw. *58*, 1–19.

81. Vieira, F. G., Fumagalli, M., Albrechtsen, A., and Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. Genome Res. *23*, 1852–1861.

82. Skotte, L., Korneliussen, T. S., and Albrechtsen, A. (2013). Estimating Individual Admixture Proportions from Next Generation Sequencing Data. Genetics *195*, 693–702.

83. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient Admixture in Human History. Genetics *192*, 1065–1093.

84. Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scand. J. Stat. *6*, 65–70.

85. Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., et al. (2010). A Draft Sequence of the Neandertal Genome. Science *328*, 710–722.

86. Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. *28*, 2239–2252.

87. Cahill, J. A., Green, R. E., Fulton, T. L., Stiller, M., Jay, F., Ovsyanikov, N., Salamzade, R., St. John, J., Stirling, I., Slatkin, M., et al. (2013). Genomic Evidence for Island Population Conversion Resolves Conflicting Theories of Polar Bear Evolution. PLoS Genet. *9*, e1003345.

88. Pasaniuc, B., Sankararaman, S., Kimmel, G., and Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. Bioinforma. Oxf. Engl. *25*, i213–221.

89. Swinburne, J. E., Boursnell, M., Hill, G., Pettitt, L., Allen, T., Chowdhary, B., Hasegawa, T., Kurosawa, M., Leeb, T., Mashima, S., et al. (2006). Single linkage group per chromosome genetic linkage map for the horse, based on two three-generation, full-sibling, crossbred horse reference families. Genomics *87*, 1–29.

90. Nicholas, F. W. (2003). Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. Nucleic Acids Res. *31*, 275–277.

91. Doan, R., Cohen, N., Harrington, J., Veazy, K., Juras, R., Cothran, G., McCue, M. E., Skow, L., and Dindot, S. V. (2012). Identification of copy number variants in horses. Genome Res. *22*, 899–907.

Table S1. Information about the horse genomes used in this study

Horse ID (Centre for GeoGenetics CGG database number)	Studbook#/Alternative ID	Sex	Horse type	Age (years)
Holoype (CGG_1_016159)	n/a	M	Przewalski	137
Paratype (CGG_1_016160)	n/a	F	Przewalski	116
Ewld1_Bijsk1 CGG_1_015486	SB11	M	Przewalski	107
Ewld2_Bijsk2 CGG_1_015487	SB12	F	Przewalski	105
Ewld3 (CGG_1_015489)	SB55	F	Przewalski	109
Ewld4 (CGG_1_015490)	SB57	F	Przewalski	99
Ewld5_Theodor (CGG_1_015493)	SB56	M	Hybrid	88
Ewld7_Wallach (CGG_1_015495)	n/a	M	Hybrid	86

Ewld1 (CGG_1_015494)	SB59	M	Hybrid	103
Emgl1 (CGG_1_015491)	n/a	F	Mongolian	102
Emgl3 (CGG_1_015492)	n/a	F	Mongolian	103
KB7674	SB285	M	Przewalski	Modern
KB7713	SB524	F	Przewalski	Modern
KB7848	SB528	F	Przewalski	Modern
SB281	SB281	M	Przewalski	Modern
SB293	SB293	M	Przewalski	Modern
SB339	SB339	F	Przewalski	Modern
SB4329	SB159	M	Przewalski	Modern
SB533	SB533	M	Przewalski	Modern
SB615	SB615	M	Przewalski	Modern

KB3879	SB274	F	Przewalski	Modern
KB7856	SB966	F	Przewalski	Modern
KB7903	n/a	n/a	Przewalski x Domesticated F1 hybrid	Modern
KB7754	n/a	n/a	Mongolian	Modern
Mon_FM1798	n/a	M	Franches-Montagnes	Modern
Mon_FM1932	n/a	M	Franches-Montagnes	Modern
Mon_FM1948	n/a	M	Franches-Montagnes	Modern
Mon_FM2218	n/a	M	Franches-Montagnes	Modern
Mon_FM1190	n/a	M	Franches-Montagnes	Modern
Mon_FM1041	n/a	M	Franches-Montagnes	Modern
Mon_FM1951	n/a	M	Franches-Montagnes	Modern
Mon_FM0467	n/a	M	Franches-Montagnes	Modern
Mon_FM0431	n/a	M	Franches-Montagnes	Modern
Mon_FM1785	n/a	M	Franches-Montagnes	Modern
Mon_FM1030	n/a	M	Franches-Montagnes	Modern
Mon_FM0450	n/a	M	Franches-Montagnes	Modern
Mor_EMS595	96	F	Morgan	Modern
Qrt_A5964 QH6	75	F	Quarter	Modern
Qrt_A5659 QH5	74	F	Quarter	Modern
Qrt_A1543 QH1	70	F	Quarter	Modern
Qrt_A2085 QH4	73	M	Quarter	Modern
Std_M977 STDB3	78	F	Standardbred	Modern
Std_M5256 STDB12	81	M	Standardbred	Modern
Std_M1009 STDB9	90	M	Standardbred	Modern
Std_Standardbred	n/a	M	Standardbred	Modern
Mng_D2629	n/a	F	Mongolian	Modern
Mng_D2628	n/a	M	Mongolian	Modern
Fjord	n/a	F	Norwegian Fjord horse	Modern
Icelandic	n/a	M	Icelandic horse	Modern
Arabian	n/a	F	Arabian horse	Modern
Prz_D2630	n/a	M	Przewalski	Modern
Prz_D2631	n/a	F	Przewalski	Modern
Prz_Przewalski	n/a	M	Przewalski	Modern

CGG10022	n/a	F	Late Pleistocene	42,692 ± 891 cal.
CGG10023	n/a	M	Late Pleistocene	16,099 ± 192 cal.

*Orlando et al., 2009; Ginolhac et al., 2012
**Gilbert et al., 2007
***Dabney et al. 2013

Reference	Sequenced in	Library name Index	Sample type	Powder used (mg)
This study	Copenhagen	PRZ1_CGATGA	Tooth	78
		PRZ1_ACTGCC		
		PRZ2_GCAACG		
		PH_H1_1_CTATCA	Hair	500
		PH_H1_2_TGTGAC		
		PH_H1_3_TGAGCC		
		PH_H1_4_CGATGA		
		PH_H1_5_GAGATA		
		PH_H1_6_CTGACA		
This study	Copenhagen	PH_P2_1_TGATGC	Hair	500
		PH_P2_2_TAGATG		
		PH_P3_GTATCT	Hair	<50
This study	Copenhagen	Ewld1_Bjisk1_TAGATG	Tooth	76
		Ewld1_Bjisk1_TGAGCC	Tooth	216
		Ewld1_Bjisk1_CTATCA		
This study	Copenhagen	Ewld2_Bjisk2_TGTGAC	Tooth	100
		Ewld2_Bjisk2_AGCATG	Tooth	225
		Ewld2_Bjisk2_CGTATA		
This study	Copenhagen	Ewld3_GTATCT	Bone	330
This study	Copenhagen	Ewld4_TGCATA	Tooth	84
This study	Copenhagen	Ewld5_Theodore_CGCTAT	Tooth	228
		Ewld5_Theodore_CGTATA	Tooth	225
		Ewld5_Theodore_CTATCA	Tooth	225
This study	Copenhagen	Ewld7_Wallach_TCTCGC	Tooth	110

This study	Copenhagen	Ewld1_TGAACA	Tooth	300
This study	Copenhagen	Emgl1_CAGCTA	Tooth	154
		Emgl1_GTGTAT	Tooth	294
		Emgl1_TGAACA		
This study	Copenhagen	Emgl2_ACATAC	Tooth	136
		Emgl3_CGCTAT	Tooth	241
		Emgl3_CAGCTA		
This study	Copenhagen	KB7674_1BIS AGCATG	Immortalized c	n/a
		KB7674_2BIS TGTGAC		
This study	Copenhagen	KB7713_1BIS GTATCT	Immortalized c	n/a
		KB7713_2BIS GACGAC		
This study	Copenhagen	KB7848_1BIS GACCGG	Immortalized c	n/a
		KB7848_2BIS CGTATA		
This study	Copenhagen	SB281_1BIS GTGTAT	Immortalized c	n/a
		SB281_2BIS ACGCAT		
This study	Copenhagen	SB293_1BIS CGATGA	Immortalized c	n/a
		SB293_2BIS TAGATG		
This study	Copenhagen	SB339_1BIS TAGATG	Immortalized c	n/a
		SB339_2BIS CGATGA		
This study	Copenhagen	SB4329_1BIS CGTATA	Immortalized c	n/a
		SB4329_2BIS GACCGG		
This study	Copenhagen	SB533_1BIS CTATCA	Immortalized c	n/a
		SB533_2BIS CTGACA		
This study	Copenhagen	SB615_1BIS ACATAC	Immortalized c	n/a
		SB615_2BIS TCTCGC		

This study	Copenhagen	KB3879_1BIS TCTCGC KB3879_2BIS ACATAC	Immortalized c	n/a
This study	Copenhagen	SB966_1BIS GACGAC SB966_2BIS GTATCT	Immortalized c	n/a
This study	Copenhagen	KB7903_1BIS ACGCAT KB7903_2BIS GTGTAT	Immortalized c	n/a
This study	Copenhagen	KB7754_1BIS TGTGAC KB7754_2BIS AGCATG	Immortalized c	n/a
This study	Bern	Mon_FM1798	Blood	n/a
This study	Bern	Mon_FM1932	Blood	n/a
This study	Bern	Mon_FM1948	Blood	n/a
This study	Bern	Mon_FM2218	Blood	n/a
This study	Bern	Mon_FM1190	Blood	n/a
This study	Bern	Mon_FM1041	Blood	n/a
This study	Bern	Mon_FM1951	Blood	n/a
This study	Bern	Mon_FM0467	Blood	n/a
This study	Bern	Mon_FM_431	Blood	n/a
This study	Bern	Mon_FM1785	Blood	n/a
This study	Bern	Mon_FM1030	Blood	n/a
This study	Bern	Mon_FM0450	Blood	n/a
This study	Minnesota	Mor_EMS595	Blood	n/a
This study	Minnesota	Qrt_A5964	Blood	n/a
This study	Minnesota	Qrt_A5659	Blood	n/a
This study	Minnesota	Qrt_A1543	Blood	n/a
This study	Minnesota	Qrt_A2085	Blood	n/a
This study	Minnesota	Std_M977	Blood	n/a
This study	Minnesota	Std_M5256	Blood	n/a
This study	Minnesota	Std_M1009	Blood	n/a
Orlando et al., 2013; Schubert et al., 2014	-	-	-	-
Do et al., 2014	-	-	-	-
Do et al., 2014	-	-	-	-
Orlando et al., 2013; Schubert et al., 2014	-	-	-	-
Orlando et al., 2013; Schubert et al., 2014	-	-	-	-
Orlando et al., 2013; Schubert et al., 2014	-	-	-	-
Do et al., 2014	-	-	-	-
Do et al., 2014	-	-	-	-
Orlando et al., 2013; Schubert et al., 2014	-	-	-	-

Orlando et al., 2013; Schubert et al., 2014	-	ACTGCC_DI	-	-
		ACTTGA		
		CGTAGT		
		CTTGTA_AP		
		CTTGTA_APem		
		CTTGTA_TG		
		GCAACG_TI		
		TGACCA_AP		
		TGACCA_TG		
		TGCAGG_SI		

Schubert et al., 2014	-	13-idx1	-	-
		13-idx10		
		13-idx11		
		13-idx12		
		13-idx2		
		13-idx24		
		13-idx25		
		13-idx26		
		13-idx3		
		13-idx4		
		13-idx9		

Extraction method	USER treatment	Rescaling option	Trimming (bp)		
			5'-end	3'-end	Nuclear genome
Silica*	USER-	--forward	7	7	2.9E-02
	USER+	No rescaling	8	8	3.8E-01
	USER+	No rescaling	8	8	5.6E-01
Hair-silica**	USER-	Defaults	4	4	9.2E-02
	USER-	Defaults	4	4	2.1E-02
	USER+	Defaults	4	4	7.8E-02
	USER+	Defaults	4	4	8.9E-02
	USER+	Defaults	4	4	6.3E-02
	USER+	Defaults	4	4	7.8E-02
				Total	1.4
Hair-silica**	USER-	Defaults	5	5	1.6E-02
	USER-	Defaults	5	5	1.3E-01
Dabney***	USER+	Defaults	0	0	3.6
				Total	3.7
Silica*	USER-	Defaults	1	1	4.5E-02
Dabney***	USER+	Defaults	1	1	7.6E-01
	USER+	Defaults	1	1	1.0E-01
				Total	9.1E-01
Silica*	USER-	Defaults	8	8	1.5E-02
Dabney***	USER+	Defaults	9	9	5.0E-01
	USER+	Defaults	9	8	4.0E-01
				Total	9.1E-01
Silica*	USER-	Defaults	0	0	1.6E-03
Silica*	USER-	Defaults	0	0	5.4E-04
Dabney***	USER-	Defaults	7	7	1.0E-03
Dabney***	USER+	No rescaling	1	1	2.2
Dabney***	USER+	No rescaling	1	1	2.1
				Total	4.3
Silica*	USER-	Defaults	5	5	5.1E-04

Silica*	USER-	Defaults	5	5	4.3E-04
Silica*	USER-	Defaults	6	6	2.2E-03
Dabney***	USER+	Defaults	3	3	9.2E-02
	USER+	Defaults	4	4	2.3E-02
				Total	1.2E-01
Silica*	USER-	Defaults	6	6	1.5E-02
Dabney***	USER+	No rescaling	7	7	3.4E-03
	USER+	No rescaling	7	7	4.4E-03
				Total	2.3E-02
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	2.9
					16.8
				Total	19.7
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	8.0
					11.9
				Total	19.9
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	3.5
					14.6
				Total	18.0
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	5.3
					18.1
				Total	23.4
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	4.6
					14.7
				Total	19.3
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	4.0
					19.0
				Total	23.0
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	6.5
					14.1
				Total	20.5
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	5.8
					14.6
				Total	20.4
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	4.2
					15.4
				Total	19.6

Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	2.9
					20.0
				Total	22.9
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	5.5
					14.5
				Total	20.1
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	6.7
					16.2
				Total	22.8
Qiagen's DNeasy Blood and Tissue Kit	n/a	n/a	n/a	n/a	3.5
					15.9
				Total	19.5
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	22.6
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	10.8
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	8.7
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	12.1
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	16.4
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	8.0
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	19.2
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	15.3
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	11.6
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	14.9
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	17.0
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	16.2
GE Healthcare Nucleon Bacc2 kit	n/a	n/a	n/a	n/a	14.0
Purogene Blood Kit C/Qiagen's DNeasy Blood a	n/a	n/a	n/a	n/a	14.0
Purogene Blood Kit C/Qiagen's DNeasy Blood a	n/a	n/a	n/a	n/a	13.7
Purogene Blood Kit C/Qiagen's DNeasy Blood a	n/a	n/a	n/a	n/a	12.6
Purogene Blood Kit C/Qiagen's DNeasy Blood a	n/a	n/a	n/a	n/a	12.3
Purogene Blood Kit C/Qiagen's DNeasy Blood a	n/a	n/a	n/a	n/a	12.4
Purogene Blood Kit C/Qiagen's DNeasy Blood a	n/a	n/a	n/a	n/a	11.4
Purogene Blood Kit C/Qiagen's DNeasy Blood a	n/a	n/a	n/a	n/a	7.7
-	-	n/a	n/a	n/a	12.3
-	-	n/a	n/a	n/a	24.1
-	-	n/a	n/a	n/a	23.5
-	-	n/a	n/a	n/a	7.7
-	-	n/a	n/a	n/a	8.5
-	-	n/a	n/a	n/a	10.8
-	-	n/a	n/a	n/a	16.9
-	-	n/a	n/a	n/a	25.3
-	-	n/a	n/a	n/a	9.7

-	-	--diff-hangs	0	0	3.0
		--diff-hangs	0	0	9.4E-01
		--diff-hangs	0	0	7.7E-01
		No rescaling	0	0	1.5
		No rescaling	0	0	7.0E-01
		--diff-hangs	0	0	9.6
		Defaults	0	0	3.0E-01
		No rescaling	0	0	8.1E-01
		No rescaling	0	0	8.4
		Defaults	0	0	9.0E-01
				Total	2.4E+01
-		--diff-hangs	0	0	2.3
		--diff-hangs	0	0	6.3E-01
		--diff-hangs	0	0	9.0E-03
		--diff-hangs	0	0	1.2
		--diff-hangs	0	0	1.1E-01
		Defaults	0	0	6.5E-01
		Defaults	0	0	1.5E-01
		Defaults	0	0	2.0
		--diff-hangs	0	0	1.2E-01
		--diff-hangs	0	0	1.5E-01
		--diff-hangs	0	0	2.6E-02
				Total	7.4

Coverage (X)		Error rate (% errors per base)		Heterozygosity
Ratio X-chromosome/Nuclear	Mitochrondrial genome	pre-filtering	post-filtering (mapping scores >30, base-quality scores >20)	ites per kilobas
	6.3			
	53.0			
	85.0			
	21.0			
	4.8			
	21.0			
	23.0			
	16.0			
	20.0			
0.4	250.0	0.317	0.258	1.2
	7.6			
	54.0			
	690.0			
0.9	750.0	0.150	0.099	0.742
	21.0			
	400.0			
	250.0			
0.4	670.0	0.257	0.178	1.353
	21.0			
	540.0			
	1000.0			
1.1	1600.0	0.550	0.463	2.354
0.5	3.0	1.188	0.568	
0.5	0.1	0.348	0.041	
	1.4			
	250.0			
	260.0			
0.4	510.0	0.661	0.622	2.659
0.6	0.2	1.512	1.089	

0.4	1.8	1.006	0.712	
1.0	2.3 210.0 57.0 270.0	0.406	0.306	n/a
1.0	13.0 8.3 12.0 33.0	0.717	0.546	n/a
0.5	602.1 3331.4 3933.5	0.147	0.043	1.533
1.0	226.1 329.5 555.6	0.149	0.047	1.546
1.0	129.0 516.4 645.4	0.134	0.038	1.265
0.5	79.6 284.7 364.4	0.158	0.046	1.707
0.5	165.9 496.6 662.5	0.139	0.045	1.399
1.0	978.0 4408.9 5386.9	0.135	0.039	1.365
0.5	2122.2 4471.8 6594.0	0.158	0.051	1.525
0.5	1017.0 2508.8 3525.8	0.160	0.050	1.237
0.5	1073.9 3787.1 4860.9	0.151	0.051	1.122

1.0	84.3 558.5 642.8	0.136	0.036	1.226
1.0	1374.2 3647.8 5022.0	0.145	0.040	1.452
0.5	3291.5 7573.4 10864.9	0.143	0.038	1.9
0.5	464.0 2002.8 2466.8	0.135	0.035	0.532
0.5	n/a	0.316	0.026	0.579
0.5	n/a	0.424	0.032	0.600
0.5	n/a	0.441	0.033	0.549
0.5	n/a	0.412	0.030	0.566
0.5	n/a	0.195	0.032	0.563
0.5	n/a	0.262	0.040	0.568
0.5	n/a	0.206	0.025	0.566
0.5	n/a	0.177	0.023	0.540
0.6	n/a	0.178	0.021	0.552
0.6	n/a	0.200	0.027	0.544
0.5	n/a	0.262	0.031	0.562
0.5	n/a	0.172	0.023	0.565
1.1	n/a	0.222	0.029	0.564
0.9	n/a	0.220	0.034	0.554
1.1	n/a	0.220	0.023	0.514
1.1	n/a	0.194	0.027	0.555
0.6	n/a	0.196	0.022	0.519
1.1	n/a	0.228	0.030	0.497
0.6	n/a	0.238	0.038	0.530
0.5	n/a	0.211	0.030	0.497
0.5	-	0.253	0.092	0.473
0.5	-	0.062	0.019	0.978
0.9	-	0.059	0.018	0.598
0.9	-	0.170	0.067	0.575
0.5	-	0.228	0.094	0.592
0.9	-	0.166	0.068	0.483
0.5	-	0.065	0.022	1.261
1.1	-	0.059	0.021	1.331
0.5	-	0.198	0.077	1.451

0.9	-	0.265	0.103	0.798
0.5	-	0.363	0.207	1.058

breeding coverag	HMM inbreeding coefficie
n/a	0.057
n/a	0.130
n/a	0.000
n/a	0.000
n/a	n/a
n/a	n/a
n/a	0.000
n/a	n/a

n/a	n/a
n/a	n/a
n/a	n/a
0.120	0.060
0.131	0.084
0.283	0.244
0.052	0.005
0.181	0.119
0.194	0.139
0.125	0.065
0.280	0.193
0.388	0.321

0.266	0.189
0.201	0.134
0.010	0.000
0.129	0.111
0.133	0.035
0.041	0.010
0.125	0.074
0.183	0.128
0.100	0.062
0.114	0.090
0.097	0.079
0.139	0.102
0.093	0.060
0.133	0.097
0.092	0.059
0.085	0.051
0.074	0.054
0.167	0.120
0.093	0.067
0.089	0.055
0.157	0.110
0.195	0.171
0.149	0.125
0.207	0.165
0.259	0.220
0.034	0.026
0.006	0.000
0.099	0.080
0.135	0.188
0.285	0.262
0.248	0.176
0.241	0.170
0.160	0.074

0.002	n/a
0.000	n/a

Table S2. Function enrichment for Przewalski's horse Ancestry Informative Marke

Przewalski's horse Ancestry Inforr			
Enrichment database	**Function**	**Database ID**	**p-value (human)**
KEGG pathways	arrhythmogenic right ventricular cardiomyopathy	5412	0.0013
Wikipathway		WP2118	0.0018
KEGG pathways	dilated cardiomyopathy	5414	0.0013
KEGG pathways	hypertrophic cardiomyopathy	5410	0.0013
KEGG pathways	protein digestion and absorption	4974	0.0013
PheWAS	disorders of muscle, ligament, and fascia	PheWAS:728	0.0084
	fasciitis	PheWAS:728.7	0.0084
	diseases of sebaceous glands	PheWAS:706	0.0365
Diseases	craniofacial Abnormalities	DB_ID:PA446836	0.0238

	mental retardation	DB_ID:PA444929	0.0279
	metaplasia	DB_ID:PA444941	0.0147
	musculoskeletal abnormalities	DB_ID:PA445000	0.0255
	neoplastic processes	DB_ID:PA445078	0.0279
	pathologic neovascularization	DB_ID:PA445080	0.0147
	nervous system malformations	DB_ID:PA445092	0.0182
	pathologic processes	DB_ID:PA445265	0.0416
	respiratory tract diseases	DB_ID:PA445525	0.0279
	skin diseases, vascular	DB_ID:PA446487	0.0100
			F_{ST}-based scans
Enrichment	**Function****	**Database ID**	**Min. p-value***

Wikipathways	non-odorant GPCRs	WP1396	3.89E-07
	mRNA processing	WP310	8.37E-08
	one carbon metabolism and related pathways	WP1770	2.53E-06
	plurinetwork	WP1763	7.07E-11
	epithelium TarBase	WP2002	7.00E-04
	IL-3 signaling pathway	WP286	1.40E-03
	kit receptor signaling pathway	WP407	4.31E-07
	Id signaling pathway	WP512	3.00E-04
	focal adhesion	WP306	4.17E-09

	chemokine signaling pathway	WP2292	3.78E-07
	Alzheimers disease	WP2059	1.24E-06
	integrated pancreatic cancer pathway	WP2256	8.85E-09
	muscle cell TarBase	WP2005	6.53E-07
	apoptosis	WP1254	3.22E-05
	striated muscle contraction	WP383	2.60E-03
	lymphocyte TarBase	WP2004	3.00E-07
	MAPK cascade	WP422	3.77E-02
	glycogen metabolism	WP500	4.15E-02
KEGG pathway	glycerophospholipid metabolism	564	6.81E-10

cytokine-cytokine receptor interaction	4060	1.59E-07
chemokine signaling pathway	4062	4.89E-06
dilated cardiomyopathy	5414	3.74E-09
ECM-receptor interaction	4512	1.53E-08
hypertrophic cardiomyopathy (HCM)	5410	1.70E-10
pathways in cancer	5200	2.39E-14
pancreatic secretion	4972	3.65E-07
gastric acid secretion	4971	6.00E-04
pancreatic cancer	5212	2.00E-04
long-term depression	4730	2.00E-04

	metabolic pathways	1100	2.82E-20
	focal adhesion	4510	1.07E-10
	vascular smooth muscle contraction	4270	1.19E-05
	phosphatidylinositol signaling system	4070	2.40E-08
	long-term potentiation	4720	7.00E-04
	renal cell carcinoma	5211	7.00E-04
Phenotypes	cellular phenotype	MP:0005384	9.90E-10
	mortality/aging	MP:0010768	9.90E-10
	abnormal survival	MP:0010769	4.69E-09
	preweaning lethality	MP:0010770	4.68E-09
	abnormal cell physiology	MP:0005621	9.90E-10
	abnormal phagocyte morphology	MP:0008251	2.54E-07
	abnormal myeloid leukocyte morphology	MP:0008250	5.40E-06

	lethality during fetal growth through weaning	MP:0010832	1.09E-06
Diseases	metabolic diseases	DB_ID:PA444938	1.50E-09
	genetic translocation	DB_ID:PA445914	2.15E-08
	multiple endocrine neoplasia type 1	DB_ID:PA446760	3.17E-08
	disease susceptibility	DB_ID:PA443919	4.42E-10
	genetic predisposition to disease	DB_ID:PA446882	2.25E-10
	mood disorders	DB_ID:PA447209	1.89E-08
	immune system diseases	DB_ID:PA444602	4.86E-09
	cancer or viral infections	DB_ID:PA128407012	3.25E-13
	neoplasms	DB_ID:PA445062	4.87E-08
	leukemia, myeloid	DB_ID:PA444761	8.00E-04
	leukemia	DB_ID:PA444750	3.78E-09
	tracheoesophageal fistula	DB_ID:PA445912	1.60E-03
	HIV	DB_ID:PA447230	1.43E-09
	death	DB_ID:PA443842	4.81E-06
	heart diseases	DB_ID:PA444368	3.85E-07
			SweeD-based selection
Enrichment	**Function**	**Database ID**	**p-value***
Wikipathways	AMPK signaling	WP1403	1.48E-02
	angiogenesis overview	WP1993	1.48E-02

	focal adhesion	WP306	3.65E-02
	mRNA processing	WP411	2.94E-02
	peptide GPCRs	WP24	1.48E-02
	regulation of actin cytoskeleton	WP51	3.25E-02
KEGG pathways	Alzheimer's disease	5010	3.33E-02
	phagosome	4145	3.15E-02
	GnRH signaling pathway	4912	2.16E-02
	cardiac muscle contraction	4260	2.16E-02
	hypertrophic cardiomyopathy (HCM)	5410	2.16E-02

	metabolic pathways	1100	2.16E-02
	RNA transport	3013	3.15E-02
	mRNA surveillance pathway	3015	2.16E-02
	Vascular smooth muscle contraction	4270	2.42E-02
	VEGF signaling pathway	4370	2.16E-02
Phenotypes	external ophthalmoplegia	HP:0000544	*8.84E-02*
	ophthalmoparesis	HP:0000597	*1.70E-01*
	ophthalmoplegia	HP:0000602	*1.72E-01*

abnormality of the seventh cranial nerve	HP:0010827	*2.84E-01*
abnormality of facial musculature	HP:0000301	*2.91E-01*
abnormality of facial soft tissue	HP:0011799	*3.24E-01*
abnormality of the cranial nerves	HP:0001291	*3.24E-01*
cranial nerve paralysis	HP:0006824	*3.24E-01*
facial palsy	HP:0010628	*2.84E-01*
reduced consciousness/confusion	HP:0004372	*3.24E-01*
increased adrenocorticotropin level	MP:0005129	n.i.
abnormal food preference	MP:0001432	n.i.
abnormal voluntary movement	MP:0003491	n.i.

	abnormal involuntary movement	MP:0003492	n.i.
	abnormal motor capabilities/coordination/movement	MP:0002066	n.i.
	increased bone trabecula number	MP:0010868	n.i.

	abnormal trabecular bone volume	MP:0010877	n.i.
	decreased epididymal fat pad weight	MP:0009289	n.i.
	increased insulin sensitivity	MP:0002891	n.i.
PheWAS	torsion dystonia	PheWAS:333.4	4.14E-02
	acute pancreatitis	PheWAS:577.1	4.14E-02
	cancer of larynx	PheWAS:149.4	4.14E-02
	impetigo	PheWAS:686.2	4.14E-02

	bronchitis	PheWAS:497	4.35E-02
	diseases of white blood cells	PheWAS:288	4.35E-02
	mixed hyperlipidemia	PheWAS:272.13	4.35E-02
Diseases	psychosomatic disorder NOS	DB_ID:PA165109034	1.30E-03
	psychophysiologic disorders	DB_ID:PA445454	1.30E-03
	Kaposi sarcoma	DB_ID:PA445599	1.90E-02
	alagille syndrome	DB_ID:PA446376	1.90E-02
	encephalomyelitis	DB_ID:PA444023	1.90E-02
	dizygotic twins	DB_ID:PA165108245	3.46E-02
	mitochondrial diseases	DB_ID:PA447172	3.46E-02

	starvation	DB_ID:PA445734	4.13E-02
	ring chromosomes	DB_ID:PA445572	4.42E-02

*p-value: adjusted p-value, after Benjamini-Hochberg correction for multiple tests.
**function identified as significant in a least one FM and one DOM panel (using the hum
“n.i.” pathway non identified in the model organism.

rs, gene candidates for selection based on FST and SweeD genomic scans.

native Markers			
Human Gene ID	**p-value (mouse)**	**Mouse Gene ID**	**Gene Names**
ENSG00000157388	0.0007	ENSMUSG00000015968	*CACNA1D*
ENSG00000143127		ENSMUSG00000090210	*ITGA10*
ENSG00000157388	0.0007	ENSMUSG00000015968	*CACNA1D*
ENSG00000143127		ENSMUSG00000090210	*ITGA10*
ENSG00000157388	0.0007	ENSMUSG00000015968	*CACNA1D*
ENSG00000143127		ENSMUSG00000090210	*ITGA10*
ENSG00000157388	0.0007	ENSMUSG00000015968	*CACNA1D*
ENSG00000143127		ENSMUSG00000090210	*ITGA10*
ENSG00000182871	0.0007	ENSMUSG00000001435	*COL18A1*
ENSG00000204291		ENSMUSG00000028339	*COL15A1*
ENSG00000099260	n/a	n/a	*PALMD*
ENSG00000147316			*MCPH1*
ENSG00000099260	n/a	n/a	*PALMD*
ENSG00000147316			*MCPH1*
ENSG00000099260	n/a	n/a	*PALMD*
ENSG00000147316			*MCPH1*
ENSG00000147316	n/a	n/a	*MCPH1*
ENSG00000152217			*SETBP1*

ENSG00000147316	n/a	n/a	*MCPH1*
ENSG00000152217			*SETBP1*
ENSG00000182871	n/a	n/a	*COL18A1*
ENSG00000143369			*ECM1*
ENSG00000147316	n/a	n/a	*MCPH1*
ENSG00000152217			*SETBP1*
ENSG00000182871	n/a	n/a	*COL18A1*
ENSG00000143369			*ECM1*
ENSG00000182871	n/a	n/a	*COL18A1*
ENSG00000143369			*ECM1*
ENSG00000182871	n/a	n/a	*COL18A1*
ENSG00000147316			*MCPH1*
ENSG00000182871	n/a	n/a	*COL18A1*
ENSG00000143369			*ECM1*
ENSG00000182871	n/a	n/a	*COL18A1*
ENSG00000143369			*ECM1*
ENSG00000204291	n/a	n/a	*COL15A1*
ENSG00000182871			*COL18A1*

Max. p-value*	**Gene Name list / Number of genes**

5.00E-04	*ADORA3, AVPR1B, C5AR1, CALCRL, CCKAR, CCR1, CCR2, CCR3, CCR8, CCR9, CCRL2, CNR2, CXCR6, EMR1, F2RL3, GHSR, GPR56, GPR77, GPRC5B, GRM3, HCRTR2, HRH3, HRH4, HTR1D, HTR2B, HTR4, HTR6, LPAR1, MC4R, MTNR1B, NIACR1, PROKR2,*
5.00E-04	*ADAD1, AKAP1, ANKRD33B, CLK2, CLP1, CSDE1, DDX20, DDX4, DDX41, DHX9, DND1, EIF2AK2, EIF2D, EIF4E, ELAVL3, EXOSC7, GRSF1, HNRNPR, ILF3, LSM11, MSI2, PCBP1, PCBP3, PPP1R14B, PRMT2, PRPF4, PUM2, RALY, RBM18, RBM22, RBM25, RBM26, RPS20, RPS28, RSRC2, SBNO1, SF3A1, SF3A2, SNRNP35, SNRPA, SNRPB2,*
9.00E-04	*AHCYL1, CBS, CEPT1, CHDH, CHKA, DNM1, ETNK1, GAD1, GPX1, GPX3, MTHFR, PLD1, TYMS,*
1.10E-03	*ACVR1, CDK2, CHD4, CUBN, DGKA, DHX9, DNMT3B, EP400, GATA6, GRSF1, GSK3B, HCK, HDAC4, LRP5, MAPK1, MBD3, MDM2, MLL2, MTOR, MYBL2, NACC1, NANOG, OTX2, PAF1, PIAS2, PIK3CD, POU2F1, PTPN11, RBPJ, RCOR2, RELA, RYBP, SATB1, SATB2, SF1, SMAD2, SMAD4, SMAD7, SMARCA4, SOX2, TCF7, TLE1,*
3.30E-03	*ADIPOR2, ATP2A2, BCKDHB, BCL6, BRI3BP, CAND1, CDCP1, CSDE1, DHX40, FBXW11, FNDC3B, HARS, HDAC4, HOXC8, HSD17B12, KIT, LAMC1, LAMC2, LAMTOR5, LCLAT1, LY6K, MTX1, NOTCH2, NUFIP2, PDLIM7, PHC2, PICALM, PLK1, PPP1R7, RAB5C, SERP1, SLC12A2, SLC25A24,*
2.20E-03	*BAD, BCL2L1, CCR3, CSF2RB, ENPP3, HCK,*
7.30E-03	*BAD, CBLB, CLTC, CSF2RB, EPOR, HCK, KIT, LYN, MAPK1, PRKCB, PTPN11, PTPN6, STAP1, TNFRSF10B, VAV1, YES1,*
5.10E-03	*CDK2, IGF1, IGF1R, NGF, RBL1, SMAD4,*
7.60E-03	*BAD, CAPN1, COL2A1, DOCK1, GSK3B, HCK, IGF1, IGF1R, ITGA10, ITGA2, ITGA8, LAMA2, LAMA3, LAMB2, LAMC1, LAMC2, MAP2K5, MAPK1, MAPK4, MYL6, MYLK2, PAK4, PAK7, PIK3CD, PIK3R4, THBS2, THBS3, TNR, VAV1, VEGFB, VWF,*

1.21E-02	*ADCY2, CCL1, CCL11, CCL7, CCR1, CCR2, CCR3, CCR8, CCR9, CXCR6, ELMO1, GNB3, GSK3B, HCK, LYN, MAPK1, NRAS, PIK3CD, PLCB1, PLCB3, PLCB4, PRKCB, RASGRP2,*
9.50E-03	*APP, ATF6, ATP2A2, BAD, CACNA1C, CACNA1D, CAPN1, CASP9, GAPDH, GSK3B, MAPK1, PLCB1, PLCB3, PLCB4, TNFRSF1A*
1.90E-02	*APP, BLM, BRF1, CAB39, CAMK2A, CAPN1, CASP9, CDK2, DAG1, DCC, DFFB, ESR1, ESR2, HNF4A, IAPP, MAPK1, MAPK4, MDM2, MYT1, NCK1, NOXA1, PCNA, PLK1, PLN, PRKAA2, RARB, RB1CC1, RBL1, RRM1,*
4.16E-02	*ADIPOR2, ATP2A2, CAMTA1, CAND1, CDCP1, CDKAL1, CHORDC1, COIL, CSDE1, CSF1, DHX40, E2F3, EIF4E, ERBB3, ESR1, FADS2, FNDC3B, HDAC4, HSD17B12, HSDL1, ITGA2, KIT, LAMC1, LAMC2, LAMTOR5, LCLAT1, MOV10, NAPG, NCEH1, NOTCH2, NUFIP2, PDE3A, PDLIM7, PICALM, PKN2, PLK1, POLA2, PPP1R7, RAB27B,*
4.15E-02	*BAD, BCL2L1, BOK, CASP9, CRADD, DFFB, DIABLO, IGF1, IGF1R, IRF4, IRF6, MDM2,*
4.16E-02	*CASQ2, MYL9, MYOM1*
4.15E-02	*ADIPOR2, ATP2A2, BCKDHB, BCL6, BRI3BP, CAND1, CDCP1, CDKAL1, CHORDC1, COIL, CSDE1, DHX40, DNMT3B, E2F3, EIF4E, FADS2, FBXW11, FNDC3B, HARS, HDAC4, HSD17B12, HSDL1, ITGA2, KIT, LAMC1, LAMC2, LAMTOR5, LCLAT1, LY6K, MOV10, MSI2, MTX1, NAPG, NCEH1, NOTCH2, NUFIP2, PDE3A, PDLIM7, PHC2, PICALM, PKN2, PLK1, POLA2, PPP1R7, RAB27B, RAB5C, RFT1, RHOG, SAC3D1, SERP1,*
4.15E-02	*NRAS, PLCB3*
4.16E-02	*GBE1, PPP2R5B*
4.80E-06	*AGPAT3, AGPAT9, CDS2, CEPT1, CHKA, DGKA, DGKB, DGKD, DGKE, ETNK1, LCLAT1, LPCAT3, LYPLA1, LYPLA2,*

5.93E-06	*ACVR1, BMP2, CCL1, CCL11, CCL7, CCR1, CCR2, CCR3, CCR9, CD27, CSF1, CSF2RB, CXCR6, EDAR, EPOR, GDF5, IFNG, IL17RA, IL2, IL20RB, IL7, KIT, LTBR, OSM,*
2.15E-05	*ADCY2, CCL1, CCR1, CCR2, CCR3, CCR9, CXCR6, HCK, NRAS, PIK3CD, PLCB1, PLCB3, PLCB4, PRKCB, RASGRP2, VAV1,*
7.13E-05	*ADCY2, ATP2A2, CACNA1C, CACNA1D, CACNA2D4, CACNG3, DAG1, IGF1, ITGA1, ITGA10, ITGA2, ITGA8, LAMA2, MYL2, PLN,*
9.32E-05	*COL2A1, COL6A6, DAG1, HMMR, HSPG2, ITGA1, ITGA10, ITGA2, ITGA8, LAMA2, LAMA3, LAMB2, LAMC1, LAMC2, SDC1,*
2.00E-04	*ATP2A2, CACNA1C, CACNA1D, CACNA2D4, CACNG3, DAG1, IGF1, ITGA1, ITGA10, ITGA2, ITGA8, LAMA2, MYL2, PRKAA2, PRKAG1, RYR2, TCEB2, TNNT2, TPM1, TTN,*
6.00E-04	*BAD, BCL2L1, BMP2, CASP9, CBLB, CDK2, DCC, DVL3, E2F3, EGLN2, FGF6, GLI2, GSK3B, HSP90AA1, IGF1, IGF1R, ITGA2, KIT, LAMA2, LAMA3, LAMB2, LAMC1, LAMC2, MAPK1, MDM2, MTOR, NRAS, PIAS2, PIAS3, PIK3CD, PLD1, PRKCB, PTCH2, RALB, RALBP1, RARB, RASSF5, RELA, SMAD2, SMAD4, STK36, STK4, TCF7, TFG, TGFB2,*
7.00E-04	*ADCY2, ATP1A1, ATP1B1, ATP1B3, ATP2A2, CCKAR, CELA2A, CELA2B, PLA2G2A, PLA2G2E, PLA2G5, PLCB1, PLCB3, PLCB4, PRKCB, RAB27B, RAB3D, RAB8A, RYR2,*
7.00E-04	*ATP1A1, MYLK2, PLCB3, PLCB4*
1.20E-03	*BAD, SMAD2, PIK3CD, CASP9, VEGFB, PLD1*
2.20E-03	*PLCB4, NRAS, IGF1, PLCB3, PLA2G2E, IGF1R, PRKCB, PLA2G5, PLA2G2A, PLCB1*

2.50E-03	*ACACB, ACSBG2, ACSM1, ACSM3, ACSM5, ACSS3, ADA, ADH1, ADH1A, ADK, AGMAT, AGPAT3, AGPAT9, AHCYL1, ALDH3B2, ALG14, ALG3, AMDHD1, AMPD1, AMT, ASNS, ATP5A1, ATP5F1, ATP6V0A2, ATP6V0B, ATP6V1A, ATP6V1B1, ATP6V1E1, ATP6V1H, B3GAT2, B3GNT4, B3GNT6, B4GALT2, BCKDHB, CBR3, CBS, CCBL2, CDS2, CEPT1, CHKA, CYP11B2, CYP7A1, DGKA, DGKB, DGKD, DGKE, DGUOK, DNMT3B, DPM3, DPYD, DPYS, DTYMK, ENO2, ENPP1, ENPP3, ENPP7, ETNK1, GAD1, GALE, GALNT10, GALNT13, GALNT8, GALNT9, GAPDH, GBA, GBE1, GDA, GLCE, GLDC, GLYCTK, GPT, GUK1, H6PD, HMGCL, HPD, HSD17B12, IDH3A, IDH3B, IMPA2, KYNU, LCLAT1, LSS, MAN1A2, MAT2B, MBOAT2, MCEE, MOGS, MTHFR, NDUFA2, NDUFA7, NDUFB10, NDUFV2, NDUFV3, NME7, NMNAT1, NT5C2, PAH, PDXK, PI4K2A, PI4KB, PIK3C2G, PIP5K1A,*
2.50E-03	*BAD, COL2A1, COL6A6, DOCK1, GSK3B, IGF1, IGF1R, ITGA1, ITGA10, ITGA2, ITGA8, LAMA2, LAMA3, LAMB2, LAMC1, LAMC2, MAPK1, MYL12A, MYL2, MYL9, MYLK2, PAK4, PAK7, PIK3CD, PPP1CC, PRKCB,*
2.50E-03	*ADCY2, CACNA1C, CALCRL, MYL6, MYL6B, MYL9, MYLK2, PLA2G2A, PLA2G2E, PLA2G5, PLCB1, PLCB3, PLCB4, PPP1CC,*
2.70E-03	*CDS2, DGKA, DGKB, DGKD, DGKE, IMPA2, PI4K2A, PI4KB, PIK3C2G, PIK3CD, PIP4K2B, PIP5K1A, PLCB1, PLCB3, PLCB4, PLCZ1,*
2.50E-03	*PLCB4, NRAS, CACNA1C, PLCB3*
2.50E-03	*NRAS, PAK7, PIK3CD, VEGFB*
2.50E-05	278
2.50E-05	339
2.50E-05	313
2.87E-05	278
5.03E-05	265
1.00E-04	75
2.00E-04	75

2.60E-03	178
5.01E-09	127
4.61E-08	97
5.67E-07	32
4.81E-06	202
5.72E-06	200
5.72E-06	58
9.08E-06	173
2.00E-04	308
1.60E-03	134
1.60E-03	17
2.50E-03	191
2.50E-03	6
6.80E-03	173
6.80E-03	34
9.20E-03	46

scan

Human Gene ID	p-value*	Mouse Gene ID	Gene
ENSG00000078142	n.i.	n.i.	*PIK3C3*
ENSG00000111725	n.i.	n.i.	*PRKAB1*
ENSG00000170890	n.i.	n.i.	*PLA2G1B*
ENSG00000089159	n.i.	n.i.	*PXN*

ENSG00000006607	2.21E-02	ENSMUSG00000034066	*FARP2*
ENSG00000089159		ENSMUSG00000029528	*PXN*
ENSG00000101654	8.91E-05	ENSMUSG00000009535	*RNMT*
ENSG00000111786		ENSMUSG00000029538	*SRSF9*
n.i.		ENSMUSG00000024527	*AFG3l2*
n.i.		ENSMUSG00000054256	*MSI1*
n.i.		ENSMUSG00000029536	*GATC*
n.i.		ENSMUSG00000041645	*DDX24*
ENSG00000185231	n.i.	n.i.	*MC2R*
ENSG00000176358	n.i.	n.i.	*TAC4*
ENSG00000078142	2.21E-02	ENSMUSG00000033628	*PIK3C3*
ENSG00000089159		ENSMUSG00000029528	*PXN*
ENSG00000157388	2.48E-02	ENSMUSG00000015968	*CACNA1D*
ENSG00000111775		ENSMUSG00000041697	*COX6A1*
ENSG00000078142	2.43E-02	ENSMUSG00000033628	*PIK3C3*
ENSG00000176014		ENSMUSG00000001473	*TUBB6*
ENSG00000157388	1.23E-02	ENSMUSG00000015968	*CACNA1D*
ENSG00000170890		ENSMUSG00000029522	*PLA2G1B*
ENSG00000157388	1.23E-02	ENSMUSG00000015968	*CACNA1D*
ENSG00000111775		ENSMUSG00000041697	*COX6A1*
ENSG00000157388	1.23E-02	ENSMUSG00000015968	*CACNA1D*

ENSG00000111725		ENSMUSG00000029513	*PRKAB1*
ENSG00000122971	1.23E-02	ENSMUSG00000029545	*ACADS*
ENSG00000110871		ENSMUSG00000041733	*COQ5*
ENSG00000111775		ENSMUSG00000041697	*COX6A1*
ENSG00000078142		ENSMUSG00000033628	*PIK3C3*
ENSG00000170890		ENSMUSG00000029522	*PLA2G1B*
n.i.		ENSMUSG00000027427	*Polr3F*
ENSG00000167272	2.43E-02	ENSMUSG00000060152	*POP5*
ENSG00000085415		ENSMUSG00000079614	*SEH1L*
ENSG00000135097	1.23E-02	ENSMUSG00000054256	*MSI1*
ENSG00000101654		ENSMUSG00000009535	*RNMT*
ENSG00000157388	1.59E-02	ENSMUSG00000015968	*CACNA1D*
ENSG00000170890		ENSMUSG00000029522	*PLA2G1B*
ENSG00000170890	1.23E-02	ENSMUSG00000029522	*PLA2G1B*
ENSG00000089159		ENSMUSG00000029528	*PXN*
ENSG00000122971	n.i.	n.i.	*ACADS*
ENSG00000125871		n.i.	*C20orf72*
ENSG00000122971	n.i.	n.i.	*ACADS*
ENSG00000141385		n.i.	*AFG3L2*
ENSG00000125871		n.i.	*C20orf72*
ENSG00000122971	n.i.	n.i.	*ACADS*

ENSG00000125871		n.i.	*C20orf72*
ENSG00000122971	n.i.	n.i.	*ACADS*
ENSG00000125871		n.i.	*C20orf72*
ENSG00000122971	n.i.	n.i.	*ACADS*
ENSG00000125871		n.i.	*C20orf72*
ENSG00000122971	n.i.	n.i.	*ACADS*
ENSG00000125871		n.i.	*C20orf72*
ENSG00000122971	n.i.	n.i.	*ACADS*
ENSG00000125871		n.i.	*C20orf72*
ENSG00000122971	n.i.	n.i.	*ACADS*
ENSG00000125871		n.i.	*C20orf72*
ENSG00000122971	n.i.	n.i.	*ACADS*
ENSG00000125871		n.i.	*C20orf72*
ENSG00000122971	n.i.	n.i.	*ACADS*
ENSG00000185231		n.i.	*MC2R*
n.i.	*5.44E-02*	ENSMUSG00000045569	*MC2R*
n.i.		ENSMUSG00000060807	*SERPINA6*
n.i.	*5.44E-02*	ENSMUSG00000029545	*ACADS*
n.i.		ENSMUSG00000021198	*UNC79*
n.i.	*5.44E-02*	ENSMUSG00000024527	*AFG3L2*
n.i.		ENSMUSG00000029516	*CIT*

n.i.		ENSMUSG00000045569	*MC2R*
n.i.		ENSMUSG00000054256	*MSI1*
n.i.		ENSMUSG00000029513	*PRKAB1*
n.i.		ENSMUSG00000024539	*PTPN2*
n.i.		ENSMUSG00000060807	*SERPINA6*
n.i.		ENSMUSG00000021198	*UNC79*
n.i.	*6.13E-02*	ENSMUSG00000024527	*AFG3L2*
n.i.		ENSMUSG00000015968	*CACNA1D*
n.i.		ENSMUSG00000029516	*CIT*
n.i.		ENSMUSG00000029513	*PRKAB1*
n.i.		ENSMUSG00000024539	*PTPN2*
n.i.	*6.13E-02*	ENSMUSG00000024527	*AFG3L2*
n.i.		ENSMUSG00000015968	*CACNA1D*
n.i.		ENSMUSG00000029516	*CIT*
n.i.		ENSMUSG00000045569	*MC2R*
n.i.		ENSMUSG00000054256	*MSI1*
n.i.		ENSMUSG00000029513	*PRKAB1*
n.i.		ENSMUSG00000024539	*PTPN2*
n.i.		ENSMUSG00000060807	*SERPINA6*
n.i.		ENSMUSG00000021198	*UNC79*
n.i.	*5.44E-02*	ENSMUSG00000034066	*FARP2*

n.i.		ENSMUSG00000024539	*PTPN2*
n.i.	*6.13E-02*	ENSMUSG00000029516	*CIT*
n.i.		ENSMUSG00000034066	*FARP2*
n.i.		ENSMUSG00000045569	*MC2R*
n.i.		ENSMUSG00000054256	*MSI1*
n.i.		ENSMUSG00000029513	*PRKAB1*
n.i.		ENSMUSG00000024539	*PTPN2*
n.i.		ENSMUSG00000060807	*SERPINA6*
n.i.		ENSMUSG00000021198	*UNC79*
n.i.	*6.13E-02*	ENSMUSG00000029522	*PLA2G1B*
n.i.		ENSMUSG00000029513	*PRKAB1*
n.i.	*6.13E-02*	ENSMUSG00000015968	*CACNA1D*
n.i.		ENSMUSG00000029522	*PLA2G1B*
n.i.		ENSMUSG00000029513	*PRKAB1*
ENSG00000149474	n/a	n/a	*CSRP2BP*
ENSG00000157837		n/a	*SPPL3*
ENSG00000135127	n/a	n/a	*CCDC64*
ENSG00000175354		n/a	*PTPN2*
ENSG00000149474	n/a	n/a	*CSRP2BP*
ENSG00000006607		n/a	*FARP2*
ENSG00000006607	n/a	n/a	*FARP2*

ENSG00000175354		n/a	*PTPN2*
ENSG00000175354	n/a	n/a	*PTPN2*
ENSG00000157837		n/a	*SPPL3*
ENSG00000149474	n/a	n/a	*CSRP2BP*
ENSG00000006607		n/a	*FARP2*
ENSG00000175970		n/a	*UNC119B*
ENSG00000152214	n/a	n/a	*RIT2*
ENSG00000175970		n/a	*UNC119B*
ENSG00000185231	n/a	n/a	*MC2R*
ENSG00000170099		n/a	*SERPINA6*
ENSG00000185231	n/a	n/a	*MC2R*
ENSG00000170099		n/a	*SERPINA6*
ENSG00000078142	n/a	n/a	*PIK3C3*
ENSG00000089159		n/a	*PXN*
ENSG00000125844	n/a	n/a	*RRBP1*
ENSG00000125846		n/a	*ZNF133*
ENSG00000135097	n/a	n/a	*MSI1*
ENSG00000170099		n/a	*SERPINA6*
ENSG00000135127	n/a	n/a	*CCDC64*
ENSG00000152214		n/a	*RIT2*
ENSG00000122971	n/a	n/a	*ACADS*

ENSG00000141385		n/a	*AFG3L2*
ENSG00000111775		n/a	*COX6A1*
ENSG00000078142	n/a	n/a	*PIK3C3*
ENSG00000111725		n/a	*PRKAB1*
ENSG00000101639	n/a	n/a	*CEP192*
ENSG00000128789		n/a	*PSMG2*
ENSG00000175354		n/a	*PTPN2*

an and/or mouse model).

Table S3. Support and parameters for the demographic models explored using d

Panel of domesticated h	Log-likelihood/Parameters	
		A: No gene flow
Franches-Montagnes (FM-DOM)	**Log-likelihood**	-144,249
	Divergence time (years ago)	17,455
	Effective population size for Przewalski's horses	3,216
	Effective population size for domesticated horses	7,696
	Symmetric gene flow rate (x 10^-5)*	n/a
	Przewalski-to-domesticated gene flow rate (x 10^-5)*	n/a
	Domesticated-to- Przewalski gene flow rate (x 10^-5)*	n/a
	End of gene flow (time in years ago)	**26,000**
	Log-likelihood	-159,754
	Divergence time (years ago)	26,003
	Effective population size for Przewalski's horses	3,034
	Effective population size for domesticated horses	7,206
	Przewalski-to-domesticated gene flow rate (x 10^-5)	27.9
	Domesticated-to- Przewalski gene flow rate (x 10^-5)	1.5
Panel of domesticated h	**Log-likelihood/Parameters**	
		A: No gene flow
Others (OT-DOM)	**Log-likelihood**	-196,520
	Divergence time (years ago)	15,817
	Effective population size for Przewalski's horses	2,839
	Effective population size for domesticated horses	29,997
	Symmetric gene flow rate (x 10^-5)*	n/a
	Przewalski-to-domesticated gene flow rate (x 10^-5)*	n/a
	Domesticated-to- Przewalski gene flow rate (x 10^-5)*	n/a
	End of gene flow (time in years ago)	**26,000**
	Log-likelihood	-222,264
	Divergence time (years ago)	26,584
	Effective population size for Przewalski's horses	2,839

	Effective population size for domesticated horses	45,397
	Przewalski-to-domesticated gene flow rate (x 10^-5)	31.7*
	Domesticated-to- Przewalski gene flow rate (x 10^-5)	0.2

Reference panel	**LSS/Parameters**	
		A: No gene flow
refPRZ	**LSS**	3.4988
refFM	**LSS**	0.9024
	LSS/Parameters	**Model C1**
	NPRZ	3,183
	NANC	3,000
	TMIG (years)	n/a
	FFP (%)	0
	MFP1 (x 10^-5)	32.00
	MPF1 (x 10^-5)	24.00
	MFP2 (x 10^-5)	n/a
	MPF2 (x 10^-5)	n/a
refPRZ	**LSS**	0.2887
refFM	**LSS**	0.4378

Migration rates are reported as the fraction of the sink population made up by new migrants.
*Gene flow rates are reported as the fraction of the sink population made up by new migrants.
"refPRZ": Przewalski's horse reference genome panel
"refFM": Franches-Montagnes reference genome panel
“LSS”: Sum of least squares score
“NPRZ”: effective population size of Przewalski’s horse today
“NANC”: effective population size of both domesticated and Przewalski's horse populations 10,000 yBP
“TMIG”: time for the change of migration rate. Migration rates are reported as the fraction of the sink popu
“FFP”: present-time migration pulse of Franches-Montagnes into Przewalski’s horses
“MPF1”: migration rate of Przewalski’s horses into Franches-Montagnes horses before the change of mig
“MFP1”: migration rate of Franches-Montagnes horses into Przewalski’s horses before the change of mig
“MFP2”: migration rate of Franches-Montagnes horses into Przewalski’s horses after the change of migra

Model	NPRZ

C1	3,183
C2_1	4,500
C2_2	4,500
C3	3,183
C4	4,500

adi and genome projections.

dadi ANALYSES

Models		
B: Symmetric gene flow	**C: Asymmetric gene flow**	**: Gene flow from PRZ to DOI**
-133,158	-109,792	-112,140
52,925	54,083	48,607
3,632	3,183	2,934
12,387	16,517	13,292
1.9	n/a	n/a
n/a	4.2	4.6
n/a	0.5	n/a

MODEL C		
19,000	**11,700**	**9,000**
-144,610	-149,822	-145,701
19,495	55,467	55,476*
3,218	3,255	3,604
7,699	11,295	17,182
31.7	10.8	8.6
0.6	2.9	2.0

Models		
B: Symmetric gene flow	**C: Asymmetric gene flow**	**: Gene flow from PRZ to DOI**
-204,546	-164,293	-165,002
34,681	35,382	34,047
2,692	2,242	2,203
50,089	75,109	63,818
1.9	n/a	n/a
n/a	6.97	7.29
n/a	0.33	n/a

MODEL C		
19,000	**11,700**	**9,000**
-198,635	-215,960	-216,631
20,276	43,857	37,508
2,839	2,739	2,755

28,892	57,467	104,283
31.7*	13.9	11.9
1.5	6.6	4.3

Genome Projections

Models

B: Symmetric gene flow	C: Asymmetric gene flow	D: Gene flow from PRZ to FM
1.9589	0.8347	1.4468
2.1137	3.0062	2.5717

MODELS C

Model C2_1	Model C2_2	Model C3
4,500	4,500	3,183
6,000	6,000	3,000
n/a	n/a	23,200
0	0	0
0.54	6.90	45.00
4.16	1.70	90.00
n/a	n/a	34.00
n/a	n/a	0.43
0.4545	0.0586	0.3454
0.218	0.0439	0.2487

ulation made up by new migrants

gration rate
gration rate
ation rate

Parameters

NANC	TMIG (years)	FFP (%)

3,000	n/a	0
6,000	n/a	0
6,000	n/a	0
3,000	23,200	0
6,000	23,200	4

: Gene flow from DOM to PRZ
-144,249
17,455
3,216
7,696
n/a
n/a
<0.001

5,500	2,000	200
-134,579	-118,841	-110,458
55,476*	55,476*	54,377
3,462	3,286	3,180
17,280	17,480	16,508
6.8	5.3	4.3
1.2	0.7	0.6

: Gene flow from DOM to PRZ
-196,407
16,604
2,891
31,720
n/a
n/a
0.13

5,500	2,000	200
-206,287	-183,492	-165,967
37,000	36,428	35,525
2,599	2,369	2,244

95,500	92,188	77,187
10.3	8.8	7.3
1.6	0.5	0.3

E: Gene flow from FM to PRZ
3.4925
0.8734
Model C4
4,500
6,000
23,200
4
80.00
40.00
0.50
1.30
0.0657
0.0066

				LSS	
MFP1 ($x10^{-5}$)	MPF1 ($x10^{-5}$)	MFP2 ($x10^{-5}$)	MPF2 ($x10^{-5}$)	refPRZ	refFM

32	24	n/a	n/a	0.2887	0.4378
0.54	4.16	n/a	n/a	0.4545	0.218
6.9	1.7	n/a	n/a	0.0586	0.0439
45	90	34	0.43	0.3454	0.2487
80	40	0.5	1.3	0.0657	0.0066

Table S4. Results of admixture tests based on f3- and D-statistics

f3-statistics tests in the form (B,C;A)					
Source population B	Source individual C	Sink individual A	f3-statistics	Standard deviation	Z-score
Franches-Montagnes	KB7903	SB285	-0.0314	0.0011	-27.78
	KB7903	SB293	-0.0229	0.0009	-24.79
	KB7903	Prz_D2631	-0.0274	0.0011	-24.24
	KB7903	Prz_D2630	-0.0264	0.0011	-23.52
	KB7903	SB966	-0.0328	0.0014	-22.79
	KB7903	SB159	-0.0222	0.001	-22.53
	KB7903	SB285	-0.0217	0.001	-22.47
	KB7903	SB533	-0.0229	0.001	-22.42
	KB7903	SB281	-0.0225	0.001	-22.2
	KB7903	SB339	-0.0236	0.0011	-21.7
	KB7903	Prz_Przewalski	-0.0241	0.0012	-20.55
	KB7903	SB524	-0.0256	0.0013	-20.01
	KB7903	SB528	-0.0264	0.0013	-19.76
	KB7903	SB615	-0.0281	0.0015	-18.99
	SB281	SB293	-0.0252	0.0014	-18.65
	SB281	SB159	-0.0236	0.0014	-16.69
	SB281	SB285	-0.0207	0.0013	-15.85
	SB281	SB339	-0.0221	0.0014	-15.78
	SB281	Prz_D2630	-0.0227	0.0014	-15.76
	SB281	SB533	-0.0202	0.0013	-15.08
	SB281	SB615	-0.0256	0.0019	-13.76
	SB281	Prz_D2631	-0.0189	0.0014	-13.34
	SB281	Prz_Przewalski	-0.0206	0.0016	-12.53
	SB281	SB966	-0.0151	0.0015	-10.28
	SB281	SB285	-0.0121	0.0014	-8.82
	SB281	SB524	-0.0116	0.0015	-7.56
	SB524	SB528	-0.0253	0.0018	-14.23
	SB524	SB966	-0.0236	0.0019	-12.28

SB524	SB285	-0.0176	0.0015	-11.98
SB524	Prz_D2631	-0.0158	0.0017	-9.49
SB524	Prz_Przewalski	-0.0134	0.0018	-7.34
SB524	Prz_D2630	-0.0119	0.0016	-7.25
SB524	SB339	-0.0110	0.0017	-6.41
SB524	SB615	-0.0131	0.0021	-6.37
SB524	SB533	-0.0068	0.0015	-4.52
SB524	SB285	-0.0061	0.0016	-3.79
SB524	SB159	-0.0048	0.0015	-3.26
SB285	Prz_D2631	-0.0182	0.0016	-11.37
SB285	Prz_D2630	-0.0158	0.0016	-10.08
SB285	SB528	-0.0152	0.0023	-6.75
SB285	SB524	-0.0103	0.0019	-5.34
SB285	SB966	-0.0099	0.002	-4.97
SB285	SB533	-0.0091	0.002	-4.67
SB285	SB339	-0.0066	0.0016	-4.17
SB285	Prz_Przewalski	-0.0070	0.0019	-3.76
SB285	SB293	-0.0048	0.0014	-3.38
SB159	SB285	-0.0304	0.0018	-17.3
SB159	SB533	-0.0283	0.0017	-16.81
SB159	SB293	-0.0232	0.0015	-15.26
SB159	SB339	-0.0157	0.0017	-8.99
SB159	Prz_D2630	-0.0130	0.0018	-7.25
SB159	Prz_Przewalski	-0.0128	0.0023	-5.53
SB159	Prz_D2631	-0.0079	0.0019	-4.2
SB293	SB533	-0.0157	0.002	-7.79
SB293	Prz_D2630	-0.0140	0.0021	-6.81
SB293	SB159	-0.0113	0.0019	-5.91
SB293	SB339	-0.0084	0.002	-4.1
SB293	SB285	-0.0092	0.0022	-4.08
SB966	SB524	-0.0134	0.0023	-5.86
SB966	Prz_D2631	-0.0106	0.0021	-5.17
SB966	SB615	-0.0105	0.0025	-4.21
SB966	SB285	-0.0070	0.002	-3.57

	Prz_Przewalski	Prz_D2630	-0.0089	0.002	-4.5
	Prz_Przewalski	SB285	-0.0076	0.0021	-3.54
	Prz_Przewalski	SB533	-0.0071	0.0021	-3.34
	Prz_Przewalski	SB159	-0.0052	0.0017	-3.07
	SB339	SB293	-0.0077	0.0019	-4.04
	SB339	Prz_D2630	-0.0067	0.002	-3.42
	Prz_D2631	Prz_D2630	-0.0069	0.0018	-3.76
	SB533	SB285	-0.0080	0.0022	-3.63
Mongolian	KB7903	SB285	-0.0335	0.0009	-36.45
	KB7903	SB966	-0.0355	0.0011	-33.44
	KB7903	Prz_D2631	-0.0287	0.0009	-31.28
	KB7903	Prz_D2630	-0.0278	0.0009	-29.43
	KB7903	SB281	-0.0232	0.0008	-29.19
	KB7903	SB293	-0.0235	0.0008	-29.16
	KB7903	SB524	-0.0279	0.001	-28.74
	KB7903	SB339	-0.0246	0.0009	-28.18
	KB7903	SB159	-0.0227	0.0008	-27.94
	KB7903	SB533	-0.0237	0.0009	-27.49
	KB7903	SB274	-0.0226	0.0008	-27.11
	KB7903	SB528	-0.0284	0.001	-27.07
	KB7903	SB615	-0.0302	0.0011	-26.49
	KB7903	Prz_Przewalski	-0.0260	0.0011	-24.67
	SB281	SB293	-0.0272	0.0011	-25.11
	SB281	SB159	-0.0255	0.0011	-23.09
	SB281	SB274	-0.0224	0.0011	-20.95
	SB281	SB339	-0.0242	0.0012	-20.41
	SB281	Prz_D2630	-0.0250	0.0012	-20.37
	SB281	SB615	-0.0290	0.0015	-19.88
	SB281	SB533	-0.0219	0.0011	-19.45
	SB281	Prz_D2631	-0.0207	0.0012	-17.37
	SB281	Prz_Przewalski	-0.0229	0.0014	-16.68
	SB281	SB966	-0.0165	0.0012	-13.91
	SB281	SB285	-0.0133	0.0011	-11.69
	SB281	SB524	-0.0134	0.0012	-10.76
	SB281	SB528	-0.0048	0.0013	-3.57
	SB524	SB528	-0.0273	0.0014	-20.1
	SB524	SB966	-0.0250	0.0014	-17.35
	SB524	SB285	-0.0182	0.0011	-16.08
	SB524	Prz_D2631	-0.0154	0.0013	-11.49
	SB524	SB615	-0.0140	0.0016	-8.7
	SB524	Prz_D2630	-0.0114	0.0014	-8.3
	SB524	Prz_Przewalski	-0.0135	0.0016	-8.27
	SB524	SB339	-0.0100	0.0014	-7.18
	SB524	SB533	-0.0052	0.0013	-3.89
	SB524	SB274	-0.0042	0.0013	-3.19
	SB285	Prz_D2631	-0.0198	0.0013	-15.45
	SB285	Prz_D2630	-0.0173	0.0013	-13.77
	SB285	SB528	-0.0173	0.0016	-10.64

	SB285	SB524	-0.0121	0.0015	-8.13
	SB285	SB966	-0.0109	0.0015	-7.14
	SB285	SB339	-0.0072	0.0013	-5.73
	SB285	Prz_Przewalski	-0.0083	0.0015	-5.39
	SB285	SB293	-0.0046	0.0012	-3.84
	SB159	SB274	-0.0340	0.0013	-26.76
	SB159	SB533	-0.0314	0.0013	-24.91
	SB159	SB293	-0.0254	0.0012	-21.47
	SB159	SB339	-0.0169	0.0014	-12.38
	SB159	Prz_D2630	-0.0140	0.0014	-9.92
	SB159	Prz_Przewalski	-0.0143	0.0018	-8.15
	SB159	Prz_D2631	-0.0082	0.0015	-5.51
	SB293	SB533	-0.0183	0.0015	-12.38
	SB293	Prz_D2630	-0.0159	0.0015	-10.59
	SB293	SB159	-0.0134	0.0015	-9.02
	SB293	SB339	-0.0101	0.0015	-6.68
	SB293	SB274	-0.0109	0.0017	-6.36
	SB966	SB524	-0.0144	0.0017	-8.53
	SB966	Prz_D2631	-0.0099	0.0016	-6.23
	SB966	SB615	-0.0112	0.0019	-6.05
	SB966	SB285	-0.0066	0.0015	-4.33
	Prz_Przewalski	Prz_D2630	-0.0085	0.0017	-5.14
	Prz_Przewalski	SB274	-0.0075	0.0017	-4.33
	Prz_Przewalski	SB533	-0.0070	0.0017	-4.03
	SB339	SB293	-0.0082	0.0014	-5.7
	SB339	Prz_D2630	-0.0074	0.0015	-4.88
	SB274	SB533	-0.0098	0.0015	-6.42
	SB274	SB159	-0.0059	0.0014	-4.13
	SB533	SB274	-0.0089	0.0017	-5.34
	Prz_D2631	Prz_D2630	-0.0077	0.0015	-5.18

Adjusted p-value
4.00E-167
4.56E-133
3.06E-127
7.96E-120
1.57E-112
4.93E-110
1.50E-109
3.28E-109
2.90E-107
8.02E-103
3.04E-91
1.66E-86
2.27E-84
5.56E-78
3.38E-75
1.98E-60
1.19E-54
1.69E-54
2.10E-53
6.22E-49
9.58E-41
2.50E-38
7.40E-34
2.93E-22
2.09E-16
4.82E-12
1.58E-43
1.18E-32

Domestic admixture proportions estimated on the basis of D					
H2	Breeds	H1 = Paratype		H1 = SB293	
		Mean	Standard deviation	Mean	Standard deviation
KB7903	Franches-Montagnes	35.29	5.09	34.83	4.67
	Mongolian	47.47	20.72	44.71	19.59
	American Quarter Horse	42.83	2.31	42.11	2.24
	Standardbred	31.11	4.09	30.4	3.88
SB528	Franches-Montagnes	13.65	1.6	13.11	1.42
	Mongolian	21.43	9.51	17.83	7.91
	American Quarter Horse	15.28	1.26	14.76	0.93
	Standardbred	11.42	1.58	10.21	1.58
SB524	Franches-Montagnes	13.3	1.56	12.14	1.52
	Mongolian	19.61	8.89	15.2	6.87
	American Quarter Horse	14.67	1.24	13.35	0.9
	Standardbred	11.33	1.86	9.69	1.67
SB966	Franches-Montagnes	11.91	1.68	11	1.37
	Mongolian	19.95	8.71	15.49	6.72
	American Quarter Horse	14.02	1.01	12.78	1.29
	Standardbred	9.75	1.66	8.24	1.59
SB281	Franches-Montagnes	10.51	1.75	9.78	1.34
	Mongolian	17.91	7.94	14.03	6.08
	American Quarter Horse	11.29	1.07	10.18	0.8
	Standardbred	8.73	1.37	7.47	1.06
SB615	Franches-Montagnes	10.97	1.35	9.3	1.28
	Mongolian	17.75	7.64	11.89	5.14
	American Quarter Horse	11	1.07	9.18	0.6
	Standardbred	8.09	1.14	6.14	1.03
Prz_Przewalski	Franches-Montagnes	6.12	1.06	5.18	1.01
	Mongolian	10.65	4.7	5.93	2.77
	American Quarter Horse	6.74	0.88	5.35	1.13
	Standardbred	5.22	0.7	3.59	0.52

3.26E-31	**Prz_D2631**	Franches-Montagnes	4.85	0.81	4.64	0.74
5.70E-19		Mongolian	10.36	5.18	7.02	4.84
1.88E-11		American Quarter Horse	3.6	0.93	2.97	0.5
1.25E-11		Standardbred	3.48	0.96	2.75	0.65
3.24E-08	**SB285**	Franches-Montagnes	4.09	1.02	3.22	0.77
3.48E-08		Mongolian	7.98	3.45	3.22	2.32
1.23E-03		American Quarter Horse	4.21	1	3.08	0.79
3.07E-03		Standardbred	4.06	0.56	2.39	0.79
3.69E-02	**SB159**	Franches-Montagnes	3.57	1.26	2.46	0.63
2.03E-28		Mongolian	8.3	3.63	3.36	1.65
1.95E-21		American Quarter Horse	3.61	0.86	2.04	1.03
3.54E-09		Standardbred	3	0.69	1.46	0.36
6.81E-06	**Prz_D2630**	Franches-Montagnes	2.5	0.84	2.01	0.9
1.61E-05		Mongolian	6.69	4.44	4.3	2.62
7.05E-04		American Quarter Horse	2.5	1.01	1.8	0.73
3.07E-03		Standardbred	1.7	0.73	0.73	0.66
2.95E-02	**SB339**	Franches-Montagnes	3.9	1.06	3.03	0.93
4.12E-02		Mongolian	7.21	3.34	2.75	1.6
9.23E-65		American Quarter Horse	3.74	0.74	2.49	0.86
3.04E-61		Standardbred	2.79	0.71	1.36	0.53
5.13E-50	**SB274**	Franches-Montagnes	2.26	1.06	1.48	0.68
5.21E-17		Mongolian	5.21	3.23	0.43	0.32
2.30E-11		American Quarter Horse	1.5	0.6	0.73	0.73
3.28E-06		Standardbred	1.39	0.53	0.32	0.21
3.09E-03	**SB533**	Franches-Montagnes	1.89	1.23	0.65	0.4
1.01E-12		Mongolian	5.4	2.81	0.31	0.32
2.76E-09		American Quarter Horse	1.69	0.66	0.56	0.38
5.30E-07		Standardbred	1.23	1.08	0.86	0.36
3.20E-03	**SB293**	Franches-Montagnes	1.17	0.79	n/a	n/a
2.79E-03		Mongolian	4.95	2.68	n/a	n/a
5.87E-07		American Quarter Horse	1.31	0.96	n/a	n/a
1.11E-05		Standardbred	1.37	0.57	n/a	n/a
3.35E-03						
3.77E-02						

1.23E-03
3.77E-02
3.77E-02
3.62E-02
2.32E-03
4.22E-02
2.74E-02
3.69E-02
3.81E-288
1.41E-242
3.62E-212
6.40E-188
8.27E-185
1.34E-184
2.06E-179
1.58E-172
8.70E-170
9.21E-165
3.53E-159
8.34E-159
3.55E-152
3.70E-132
9.81E-137
7.48E-116
6.60E-96
5.28E-90
1.10E-89
1.68E-85
6.52E-82
2.47E-65
1.81E-60
1.74E-41
2.68E-29
6.69E-25
1.01E-02
2.09E-87
2.69E-65
2.49E-56
2.21E-28
8.02E-16
2.00E-14
2.04E-14
4.94E-11
5.73E-03
1.93E-02
2.47E-52
1.18E-40
1.64E-24

4.31E-14
4.60E-11
1.25E-06
4.13E-06
5.54E-03
3.22E-155
1.19E-134
2.17E-100
7.71E-33
9.63E-22
4.65E-14
2.88E-06
8.59E-33
1.87E-24
5.51E-17
5.69E-10
4.03E-08
3.23E-15
8.30E-08
2.22E-07
1.80E-03
4.98E-05
1.99E-03
4.32E-03
1.23E-06
1.71E-04
3.15E-08
3.42E-03
3.54E-06
4.20E-06

)-statistics tests in the form (Outgroup,(H3,(H2,H1)))					
H1 = SB533		H1 = SB274		H1 = SB339	
Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
34.5	4.72	33.86	4.74	32.82	4.69
44.73	19.65	44.42	19.52	43.39	19.53
42.11	2.18	41.94	2.27	40.62	2.16
30.86	3.79	30.44	3.82	29.44	3.7
12.39	1.4	11.72	1.39	10.4	1.13
17.85	8.01	17.3	7.68	15.6	7.35
14.52	0.83	14.41	1.05	12.57	0.87
10.56	1.48	10.28	1.53	9.01	1.28
11.74	1.62	10.89	1.42	9.46	1.19
15.36	7.05	14.86	6.72	12.99	6.42
13.44	0.93	13.16	1.04	11.22	1.21
10.29	1.37	9.9	1.66	8.56	1.35
10.58	1.36	9.73	1.37	8.28	1.04
15.65	6.76	15.18	6.58	13.31	5.97
12.8	0.85	12.55	0.86	10.55	0.92
8.84	1.38	8.4	1.53	6.98	1.18
9.42	1.53	8.48	1.31	6.99	1.35
14.26	6.19	13.6	5.88	11.75	5.4
10.26	0.68	10	1.05	7.92	0.76
8.1	1.01	7.63	1	6.22	0.81
8.92	1.45	7.96	1.23	6.51	1.26
12.06	5.19	11.47	4.99	9.54	4.2
9.25	0.75	8.93	1.14	6.87	0.77
6.77	1.03	6.25	1.02	4.87	0.73
4.91	0.87	4.04	0.82	2.52	0.84
6.16	2.76	5.68	2.74	3.56	1.61
5.44	0.61	5.21	0.81	3.15	0.7
4.21	0.79	3.79	0.58	2.38	0.45

4.17	0.88	3.24	0.81	1.69	0.83
7.15	4.6	6.63	4.89	4.58	3.55
2.92	0.76	2.67	1.24	0.75	0.41
3.31	0.7	2.87	0.59	1.41	0.64
2.65	0.76	1.81	0.72	0.69	0.25
3.49	1.98	2.76	2.29	0.88	0.57
3.01	1.01	2.82	1.3	0.86	0.32
2.91	1.16	2.51	0.71	1.06	0.79
2.04	0.66	1.04	0.76	0.86	0.83
3.55	1.6	2.9	1.57	0.73	0.6
2.09	0.42	1.8	0.5	0.61	0.56
2.07	0.45	1.63	0.46	0.25	0.23
1.76	0.82	0.75	0.44	1	0.85
4.02	2.93	4.36	2.44	2.92	1.59
1.93	1.02	1.65	1.33	0.56	0.68
1.37	1.18	0.94	0.56	0.91	0.47
2.61	0.95	1.59	0.93	n/a	n/a
2.93	1.51	2.23	1.63	n/a	n/a
2.56	0.58	2.24	0.9	n/a	n/a
2.01	0.43	1.51	0.43	n/a	n/a
1.03	0.69	n/a	n/a	1.62	0.97
0.69	0.56	n/a	n/a	2.31	1.74
0.61	0.25	n/a	n/a	2.3	0.94
0.59	0.5	n/a	n/a	1.54	0.44
n/a	n/a	1.04	0.7	2.68	1
n/a	n/a	0.69	0.57	3.03	1.61
n/a	n/a	0.61	0.25	2.62	0.61
n/a	n/a	0.6	0.5	2.04	0.44
0.66	0.41	1.51	0.69	3.13	0.99
0.31	0.32	0.43	0.32	2.85	1.71
0.56	0.39	0.74	0.74	2.56	0.9
0.86	0.36	0.32	0.21	1.38	0.54

Table S5. Admixture proportion of genomic blocks containing sites underlying Mendelian traits.

Chromosome	Coordinate	Phenotype	Gene	Mutation	SB159	SB274	SB281	SB285	SB293	SB339	SB524	SB528	SB533	SB615	SB966	Prz_D2630	Prz_D2631	KB7903
1	########	Racing performance	*ACTN2*	A>G	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	########	Leopard complex spotting and cation channel congenital stationary night blindness	*TRPM1*	C>T	.	.	C/T	.	.	C/T	.	.	.	.	.	.	.	.
1	########	Hereditary equine dermal isomerase B asthenia	*PPIB*	G>A	G4	.	.	.	.	.	.	.	.	.	.	G2	G1	.
1	########	Lavender foal syndrome	*MYO5A*	1 DEL	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	########	Cerebellar abiotrophy	*TOE1*	C>T	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	########	Racing performance	*COX4/1*	C>T	T/T	.	C/T	C/T	C/T	T/T	C/T	C/T	.	C/T	C/T	T/T	T/T	C/T
3	########	Chestnut coat color	*MC1R (1)*	C>T	.	C/T	C/T	.	C/T	.	.	.	C/T	C/T	.	.	.	C/T
3	########	Chestnut coat color	*MC1R (2)*	G>A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	########	Sabino spotting	*KIT (1)*	A>T	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	########	Tobiano spotting pattern	*KIT (2)*	G>A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	########	Larger body size	*LCORL/NCAPG*	T>C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	########	Racing performance	*PON1*	C>T	.	.	.	.	.	.	.	.	.	.	.	C6	.	.
4	########	Racing performance	*PDK4 (1)*	C>A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	########	Racing performance	*PDK4 (2)*	G>A	.	.	.	.	.	.	.	A/G	.	.	.	.	.	A/G
4	########	Racing performance	*ACN9*	C>T	.	.	.	.	.	.	.	C/T	.	.	.	.	.	.
4	########	Congenital myotonia	*CLCN1*	A>C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	########	Junctional epidermolysis bullosa	*LAMC2*	1 C ins	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	########	Splashed white coat	*PAX3*	C>T	.	.	.	.	.	.	.	.	C7	.	C4	C3	.	.
6	########	Silver coat color	*PMEL17*	G>A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	########	Larger body size	*HMGA2*	C>T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	T/T	C/T
9	########	Severe combined immunodeficiency	*DNAPK*	5 DEL	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	########	Wither height	*ZFAT (1)*	C>T	.	.	.	.	.	.	T/T	T/T	.	.	C/T	.	.	.
9	########	Wither height	*ZFAT (2)*	C>A	.	.	.	.	.	.	A/A	A/A	.	.	A/C	.	.	.
9	########	Wither height	*ZFAT (3)*	G>A	.	.	.	.	.	.	A/A	A/A	.	.	A/G	.	.	.
9	########	Wither height	*ZFAT (4)*	G>A	.	.	.	.	.	.	A/A	A/A	.	.	A/G	.	.	.
9	########	Larger body size	*ZFAT (5)*	C>T	.	.	.	.	.	.	.	.	.	.	.	.	.	.
10	########	Malignant hyperthermia	*RYR1*	C>G	.	.	.	C7	.	.	.	.	C2	C7	.	C3	.	.
10	########	Racing performance	*CKM*	G>A	.	.	.	A/A	.	.	.	A3,G3	.	A/G	G7	.	A/G	A/A
10	########	Polysaccharide storage myopathy	*GYS1*	C>T	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11	########	Equine hyperkalemic periodic paralysis	*SCN4A*	C>T	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11	########	Glanzmann Thrombasthenia	*ITGA2B*	10 DEL	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11	########	Larger body size	*LASP1*	G>A	A1	A1	G4	A1	A2,G2	A5	A1	?	A3	A4	A1,G4	?	?	A2
14	########	Dwarfism	*PROP1 (1)*	G>C	.	C/G	.	.	.	.	.	.	.	.	.	G3	G4	.
14	########	Dwarfism	*PROP1 (2)*	T>C	.	C/T	.	.	.	.	.	.	.	.	.	T5	.	.
14	########	Dwarfism	*ND*	G>A	A/G	A/G	.	A/G	.	A/A	A/G	A/G	A/G	A/A	A/G	A/G	A/A	.
14	########	Champagne dilution	*SLC36A1*	G>C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
14	########	Autosomal recessively inherited chondrodysplasia	*SLC26A2*	A>G	.	.	.	.	.	.	.	.	.	.	.	.	.	.
16	########	Macchiato, hearing loss	*MITF (1)*	T>C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
16	########	Splashed white coat	*MITF (2)*	5 DEL	.	.	.	.	.	.	.	.	.	.	.	.	.	.
16	########	Splashed white coat	*MITF (3)*	11 DEL	.	.	.	.	.	.	.	.	.	.	.	.	.	.
17	########	Lethal white foal syndrome	*EDNRB*	GA>CT	.	.	.	.	.	.	.	.	.	.	.	.	.	.
18	########	Optimum racing position	*MSTN*	T>C	.	.	.	.	.	.	.	.	.	.	.	.	T36	.
21	########	Cream coat color	*SLC45A2*	G>A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22	########	Racing performance	*COX4/2*	C>T	.	.	.	.	.	.	.	C/T	.	.	.	.	.	.
22	########	Black and bay color	*ASIP*	11 DEL	.	.	.	.	.	.	.	.	.	.	.	.	.	.
23	########	Pattern of locomotion (altered gait)	*DMRT3*	C>A	.	.	C3	.	.	.	.	.	.	.	.	.	C3	.
26	########	Foal immunodeficiency syndrome	*SLC5A3*	C>T	.	.	.	.	.	.	.	.	.	.	.	.	.	.
X	########	Androgen insensitivity syndrome (AIS)	*AR*	A>G	.	.	.	A4	.	.	.	.	.	.	.	.	.	.
X	########	Incontinentia pigmenti	*IKBKG*	C>T	.	.	C7	.	.	.	.	.	.	.	.	C2	.	.

"INS": insertion; "DEL": bp deletion. A dot (.) stands for same alleles as the reference sequence EquCab2.0. Variants located in white, red and orange cells represent alleles of Przewalski's, domesticated or hybrid horse ancestry, respectively.