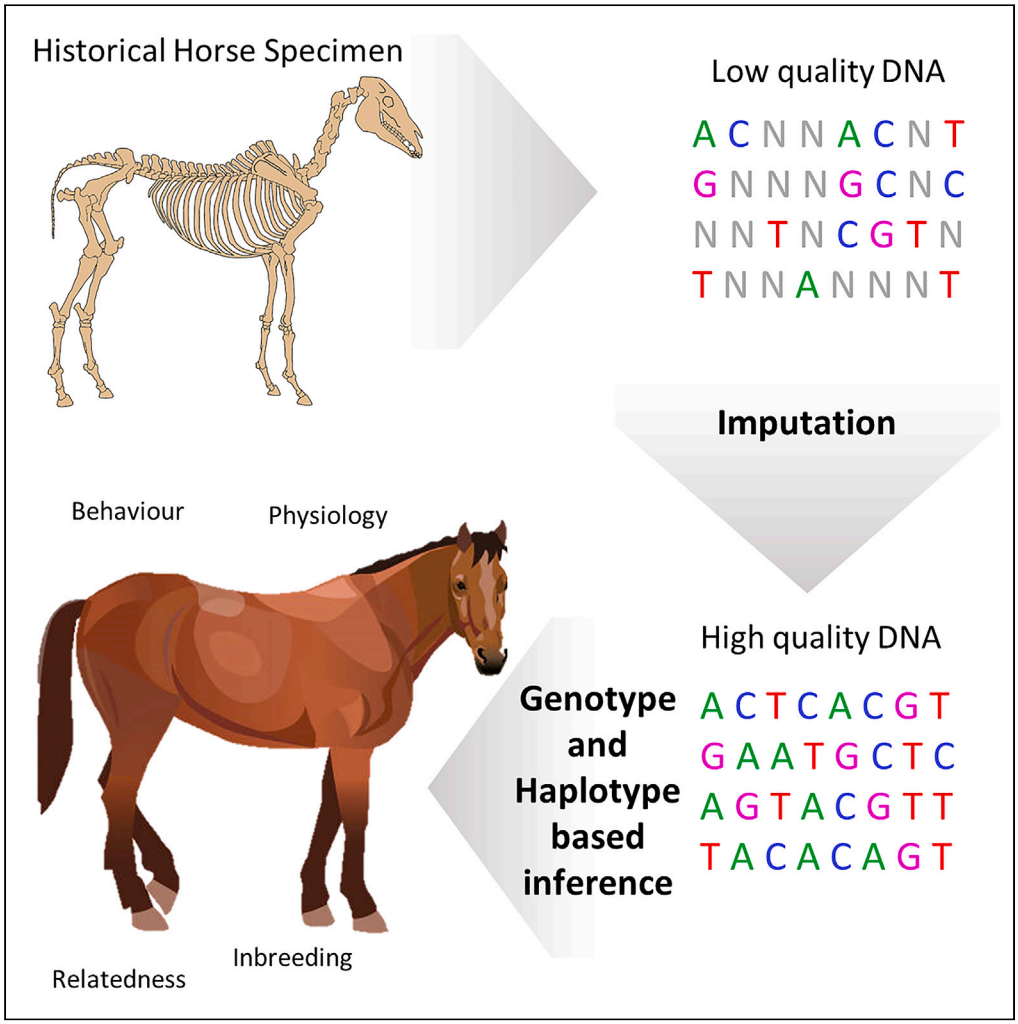
iScience

CellPress
OPEN ACCESS

Article

Imputed genomes of historical horses provide insights into modern breeding



Evelyn T. Todd, Aurore Fromentier, Richard Sutcliffe, ..., Ted Kalbfleisch, Jessica L. Petersen, Ludovic Orlando

ludovic.orlando@univ-tlse3.fr

Highlights

We use modern diversity to statistically impute four historically important horse genomes

Imputed SNPs provide insights into behavioral and physiological traits of historical horses

We characterize genomic changes in Clydesdales, Thoroughbreds, and Przewalski's horses

Modern horses show higher levels of inbreeding than historical specimens

Todd et al., iScience 26, 107104
July 21, 2023
https://doi.org/10.1016/j.isci.2023.107104

iScience

CellPress
OPEN ACCESS

Article

Imputed genomes of historical horses provide insights into modern breeding

Evelyn T. Todd,[1] Aurore Fromentier,[1] Richard Sutcliffe,[2] Yvette Running Horse Collin,[1] Aude Perdereau,[3] Jean-Marc Aury,[3] Camille Èche,[4] Olivier Bouchez,[4] Cécile Donnadieu,[4] Patrick Wincker,[3] Ted Kalbfleisch,[5] Jessica L. Petersen,[6] and Ludovic Orlando[1,7,*]

SUMMARY

Historical genomes can provide important insights into recent genomic changes in horses, especially the development of modern breeds. In this study, we characterized 8.7 million genomic variants from a panel of 430 horses from 73 breeds, including newly sequenced genomes from 20 Clydesdales and 10 Shire horses. We used this modern genomic variation to impute the genomes of four historically important horses, consisting of publicly available genomes from 2 Przewalski's horses, 1 Thoroughbred, and a newly sequenced Clydesdale. Using these historical genomes, we identified modern horses with higher genetic similarity to those in the past and unveiled increased inbreeding in recent times. We genotyped variants associated with appearance and behavior to uncover previously unknown characteristics of these important historical horses. Overall, we provide insights into the history of Thoroughbred and Clydesdale breeds and highlight genomic changes in the endangered Przewalski's horse following a century of captive breeding.

INTRODUCTION

Horses have provided power, speed, and long-distance mobility to human societies for over four thousand years.[1,2] The advent of the first American and European studbooks from the early 18th century marked a new era in horse breeding management, in which the selection for desirable traits resulted in the emergence of strongly differentiated subpopulations. These subpopulations have provided the basis for the many hundreds of horse breeds that we know today,[3] each representing a unique biocultural heritage developed to support humans with specific tasks or for use in sport. With the rise of mechanical transportation from the late 19th century,[4] the role of horses has changed in modern western societies to an increased and almost exclusive focus on leisure and sport. The dramatic decline in demand for horsepower during the 20th century has resulted in the collapse of many once-important breeds, with the current horse population estimated at 60 million worldwide today.[5]

The increasing genomic resources available for horses have started to reveal the genetic consequences of modern breeding practices and underlying demographic collapses. For example, pre-18th century genomes of domestic horses were found to have a higher level of heterozygosity and lower genetic load than modern genomes.[6] Modern traditional working breeds, including coldblood draught horses such as Friesians, show particularly elevated genetic load, possibly due to deprecation of horsepower in agriculture.[7] The intensive selection for performance has also resulted in disproportionate contributions of highly successful individuals in sport horses, such as racing Thoroughbreds. As expected in a closed population under selection, the amount of genetic diversity has decreased since the founding of these breeds.[8] In some instances, however, inbreeding has been linked to reduced racing performance.[9,10]

In addition to domestic horses, the endangered population of Przewalski's horses (PH) became extinct in the wild in 1969 and was saved by captive breeding programs throughout the 20th century. PH split from domestic horses ~35,000–45,000 years ago and were once considered the only wild horses remaining in the modern world.[11] However, archaeological horse remains ~5,500 years old from Botai, Central Asia carry genomes directly ancestral to modern PH, suggesting that PH represent the feral descendant of the earliest domestic horses known in the archaeological record.[12] Captive breeding of PH has led to higher

[1]Centre d'Anthropobiologie et de Génomique de Toulouse (CAGT), CNRS UMR 5288, Université Paul Sabatier, 37 Allées Jules Guesde, Bâtiment A, 31000 Toulouse, France

[2]Glasgow Museums Resource Centre, 200 Woodhead Road, Nitshill, G53 7NN Glasgow, UK

[3]Genoscope, Institut de biologie François Jacob, CEA, Université d'Evry, Université Paris-Saclay, 91042 Evry, France

[4]GeT-PlaGe - Génome et Transcriptome - Plateforme Génomique, GET - Plateforme Génome & Transcriptome, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, 31326 Castanet-Tolosan Cedex, France

[5]MH Gluck Equine Research Center, University of Kentucky, Lexington, KY 40546-0091, USA

[6]Department of Animal Science, University of Nebraska-Lincoln, 3940 Fair St, Lincoln, NE 68583-0908, USA

[7]Lead contact

*Correspondence: ludovic.orlando@univ-tlse3.fr

https://doi.org/10.1016/j.isci.2023.107104

inbreeding levels, genomic load, and domestic admixture today, relative to the 19th century,[7,11] leading to concerns as to population health and viability leading into the future. The understanding of genetic diversity is therefore critical for future management and conservation strategies in both the PH and domestic breeds.

Historical genomes hold great potential for reconstructing the complex history of modern breeding, involving various and potentially changing selection regimes,[13] admixture,[6] and repeated episodes of demographic collapses (see[14] for a review). Gene candidate genotyping in historical Thoroughbreds suggested that genetic variation strongly associated with sprinting performance entered the pedigree through native British mares rather than imported male founders of oriental origins.[15] The genomes of historical horses have also helped correct the pedigree of PH, which proved particularly difficult to assemble from the different sources kept across multiple institutions throughout the history of captivity.[11] However, the vast majority of sequenced ancient horse genomes are derived from archaeological and paleontological specimens and date several millennia back and beyond ([1,6,12,16–18]; see[19] for a review). Focusing attention on museum specimens from the last few centuries will improve our understanding of the history of modern breeding and its consequences in extant populations, which remains largely overlooked.

Sequencing the genome of museum specimens comes with a number of challenges, mostly relating to DNA preservation.[20] Low DNA quality may limit access to only a fraction of the genome,[21,22] or may require extensive and costly sequencing efforts.[23] Despite its potential to gain information and resolution at the genome-wide scale,[24] statistical imputation has only recently been used in ancient domestic animals (e.g. donkeys[25] and pigs[26]), but has never been applied to ancient horses. In this study, we used imputation to characterize the genomic variation present in four museum horse specimens from the late 19th and early 20th century. The resulting four genomes shed light into the history of the endangered PH as well as that of racing Thoroughbreds and draught Clydesdale horses. The framework presented not only reveals individual phenotypes not preserved in historical records but also clarifies Fthe timing of changing breeding practices and their impact on inbreeding, admixture, and the genetic diversity of modern breeds and subpopulations.

RESULTS AND DISCUSSION

Genome panel and imputation

Previous work reported the genome sequence of one historical Thoroughbred called Dark Ronald from a bone sample preserved at the Martin Luther-Universitat Halle, Wittenberg, Germany (“Ronald” thereinafter).[6] Ronald was an elite-performing English Thoroughbred born in 1905, who provided an important genetic contribution to Warmblood breeds on the European continent, especially in Germany. His genome was characterized using shotgun sequencing at an average depth of coverage of 0.80-fold. Previous work also reported the genome sequences of two historical PH, including the holotype captured in 1872, and preserved since at the Zoological Museum of the Academy of Science in St Petersburg, Russia. The second PH died in 1899, a time when the captive stock that is foundational to the modern subpopulation was not yet formed.[11] DNA preservation was relatively limited and only compatible with genome characterization at an average depth of coverage of 0.32-fold and 2.95-fold, respectively. These two individuals are hereafter referred to as “Holotype” and “Paratype,” respectively.

In order to characterize the genome of a draught horse from the early 20th century, we collected bone material from Baron O’Buchlyvie (“Baron” thereinafter), which is preserved at the Kelvingrove Art Gallery and Museum in Glasgow, Scotland. Baron was a prominent Clydesdale stallion, born in 1900 and famous for being sold for 9,500 guineas at auction in 1911, which is equivalent to approximately 1.5 million euros today, and still represents the record price for a draught horse. Pedigree records indicate that Baron provided an important genetic contribution to the Clydesdale breed. His DNA was extracted and manipulated in the state-of-the-art ancient DNA facilities of the Centre for Anthropobiology and Genomics of Toulouse, France. The final genome coverage was 3.09-fold, following the sequencing of 978 million read pairs from 8 independent Illumina DNA libraries on the HiSeq4000 instrument.

To analyze the genetic makeup of these historical genomes relative to modern horses, we collected an extensive panel of 401 genomes, including 20 PH, 380 domestic horses, and 1 donkey that we used as an outgroup. Since no whole-genome sequences were previously published of Clydesdale horses or their close Shire relatives, we supplemented this panel with 20 Clydesdale and 10 Shire genomes at an average

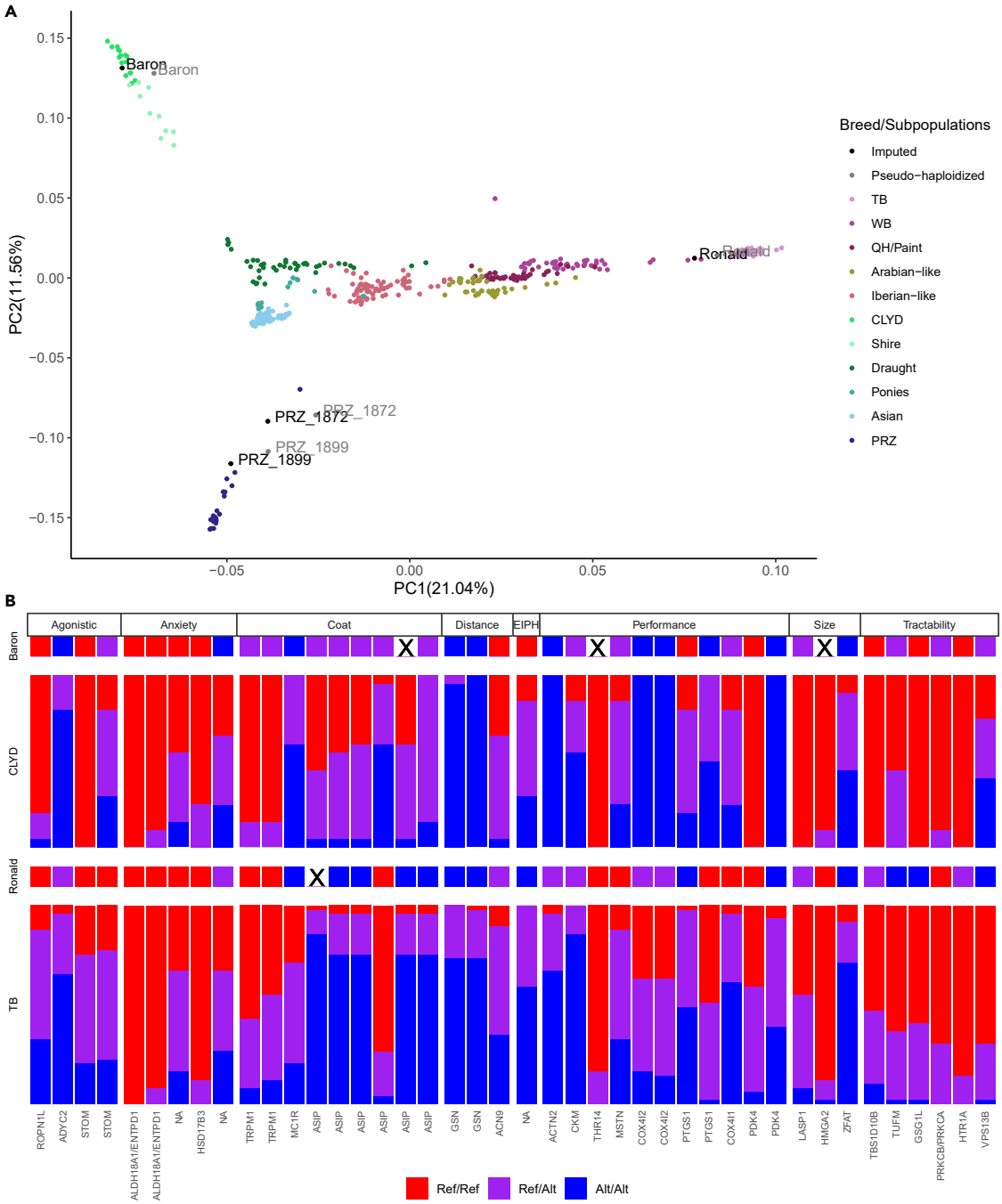


Figure 1. Population structure and genotype frequency of ancient and modern horses

(A) Principal component analysis of modern horse breeds/populations (colored), and imputed ancient genomes (black), conditionning on transversions only and using the "smartpca" function from the EIGENSOFT package (version 6.1.4).[31,32] Pseudo-haploidized ancient genomes are projected and colored in gray. Modern horses are grouped into clades according to phylogenetic clustering (Figure 2A): Asian (n = 72, breeds: DATO, DBAO, JEJU, JICH, JIZI, LNGZ, MONG, MOGO, MZHU, NIMU, NIQI, YAKU), Ponies (n = 13, breeds: DART, DUEL, ERLC, ICEL, SHET, WELS, YAQI, YILI), Draught (n = 41, breeds: DART, HAFL, FRIE, FROM, NORI, PRCH, PRCHx), Iberian-like (n = 75, breeds: ANDA, BACA, CART, CHIC, GALI, FLOR, FOPO, LIPI, LUSI, MGMR, MGPL, MORG, MUST, OLGA, SACR, SORR), Arabian-like (n = 46, breeds: ARAB, AKTK, GALI, REIT, SB), QH/Paint (n = 42, breeds: CURL, MIXD, PAIN, QH), Warmblood (n = 43, breeds: APPA, AMWB, APQH, BAVA, BGWB, DUTC, HANO, HOLS, OLDE, PAxWB, POWB, QH, SWIS, TRAK, WB, WEST, WURT). The breeds/subpopulations of PRZ (n = 19), SHIRE (n = 10), CLYD (n = 20), and TB (n = 49) are individually colored as they are the modern counterparts of the historical individuals.

Figure 1. *Continued*

(B) Genotype frequencies of variants associated with anxiety, coat color, optimal distance, exercise-induced pulmonary hemorrhage (EIPH), racing performance, size, and tractability in Baron, modern Clydesdales (n = 20), Ronald, and modern Thoroughbreds (n = 49). Homozygotes for the allele associated with the positive trait (less agonistic, lower anxiety, derived coat, optimal short distance performance, no EIPH, elite performance, large size, and higher tractability) are colored in red, heterozygotes in purple, and homozygotes for the alternative allele in blue.

depth of coverage of 5.42- to 21.15-fold (Table S1). Combined, the final panel considered for the following analyses consisted of 1 donkey and 430 horses with 73 breeds/subpopulations represented which reflect the wide range of phenotypic diversity found throughout the world. Using Graphtyper,[27] we identified 8,687,237 high-quality, segregating SNPs in these modern equids at a minimum allele frequency of 5%. Those variants were statistically phased using BEAGLE (version 5.1),[28] using the recombination map of Beeson and colleagues.[29]

Principal component analysis (PCA) showed separation of PH from domestic horses on the PC2 axis (Figure 1A). Within domestic breeds, there is segregation of draught Clydesdale and Shire horses from other modern breeds on the PC2 axis and a continuum on the PC1 axis from Asian breeds to Thoroughbreds. This structure is in line with that observed from genotypic data, reported by Petersen and colleagues.[30] Phylogenetic reconstruction also showed strong genetic sub-structuring within modern horse breeds, which we classified into 8 main clades consisting of PH (n = 19), Asian breeds (n = 73), European ponies (n = 12), Draught horses (n = 71), Iberian-like (n = 75), Arabian-like (n = 46), Quarter Horse/Paint (n = 42), and Warmblood/Thoroughbred (n = 92) (Figure 2A).

Using the 8.7 million SNPs in the modern panel (minor allelic frequency $\geq$5%), we statistically imputed the four ancient genomes, which provided 5.91–8.28 million SNP genotypes after filtering for high-quality variants, with a genotype probability score over 0.99 (Figure S1). Down-sampling and re-imputing of 9 high-coverage modern horse genomes (3 each of PH, Clydesdale, and Thoroughbred) showed high genotype accuracy (>99%, Table S2, Figure S2), in line with previous reports in donkeys using sequence data equivalent to 0.75-fold coverage and above.[25,36]

We also carried out a common analysis in ancient DNA research by which ancient genomes are pseudo-haploidized by randomly sampling one read per genome location and PCA projected against modern diversity. We then compared the resulting projections to their PCA placement based on imputed data, which showed highly consistent clustering of ancient samples (Figure 1A). We found that the genome imputation recovered highly similar variants in all four ancient samples, before and after rescaling and trimming the bam files, attesting to the robustness of the methods to postmortem degradation or sequencing errors (Table S3, Figure S3). We used the genomes imputed after rescaling and retrimming for all downstream analysis as a conservative measure.

Warmblood Fragile Foal Syndrome is a monogenetic defect of connective tissue,[37] which is characterized by hyperextensible and fragile skin and also occurs in Thoroughbreds.[38] Pedigree analysis had suggested that the mutation responsible for this syndrome could trace back to Ronald. However, genetic testing based on PCR sequencing ruled out Ronald as a carrier.[39] Although this variant did not pass the minor allele frequency filter in our genotype panel, our approach allowed for the genotyping of millions of genetic variants, including 42 variants that are causative or associated with important phenotypic traits in horses ranging from aesthetics to physiology, and behavior (Figure 1B, Table S4). The imputed genotypes at these loci provided insights into the phenotypes of our two historical domestic horses: Ronald and Baron. Neither of them carried alleles deriving chestnut coat, or variants associated with lighter bay coat color (in five breeds studied), in agreement with photographs and historical reports (Figure 1B).[40–42] Baron was predicted to be heterozygous for one derived mutation at *LASP1*, one of the four loci driving height variation in horses.[43] Since this mutation is almost fixed in modern Shire and Clydesdale horses, we conclude that the selection for increased height in these breeds was still ongoing at start of the 20th century, when Baron was born. Ronald exhibited a range of derived mutations at eleven loci associated with improved racing performance in Thoroughbreds.[44–46] He was not homozygous for alleles at three loci associated with optimal sprint performance,[45] indicating that he was predisposed toward endurance races, in line with his racing record. A variant associated with exercise-induced pulmonary hemorrhage was not found in Ronald, suggesting he may not have been susceptible to this condition that negatively affects racing performance.[47] Finally, Ronald was homozygous for derived mutations in *TUFM*, *GSG1L*, and *VPS13B*, which are associated with tractability and are rare in modern Thoroughbreds.[48] He was also heterozygous for an allele in *HTRA1*,

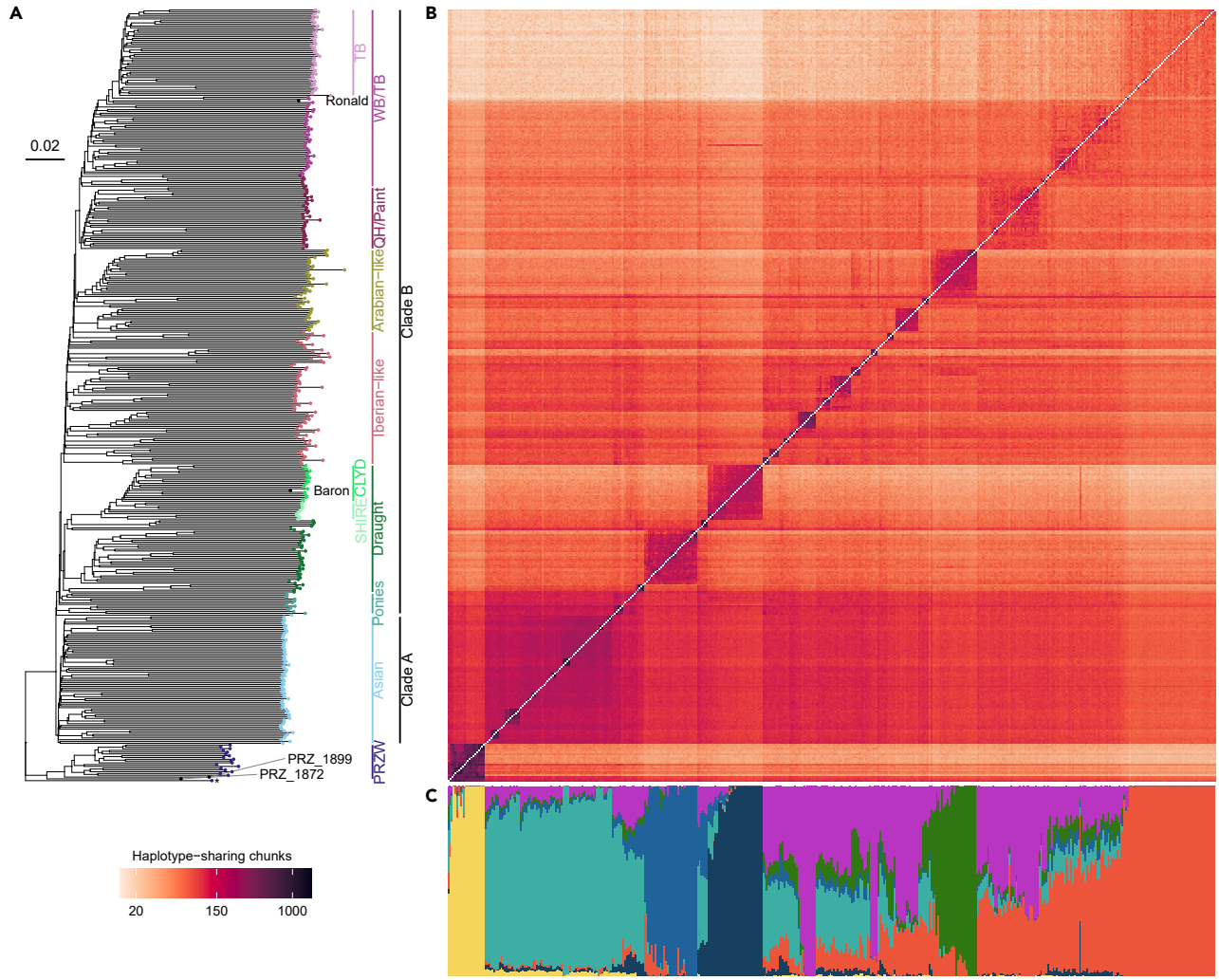


Figure 2. Genetic affinities between ancient horses and modern breeds/subpopulations

(A) Neighbor-joining tree with 100 bootstrap pseudo-replicates constructed in FastMe (version 2.1.4)[33] (n = 434 individuals, n = 8,687,237 SNPs). The four ancient samples are labeled, as are the major clade groupings of horse breeds/populations. The number and breeds of horses in each clade are detailed in Figure 1 legend and in Table S5. The F1 hybrid between a PH and a domestic horse is denoted with an asterisk (*).

(B) Haplotype sharing heatmap estimated using fineSTRUCTURE (version 4.1.1)[34] (n = 434 individuals, n = 4,816,764 SNPs).

(C) Ancestry proportions from ADMIXTURE analysis (version 1.3.0)[35] at the optimal K-value of 7 (n = 434 individuals, n = 172,267 SNPs).

which has been associated with lower tractability in Thoroughbreds.[49] However, neither Baron nor Ronald carried an excess of derived alleles for variants associated with anxiety or agonistism.[48] Overall, the imputation methodology developed here and implemented for the first time in historical horses showcases how the range of characterizable phenotypes can be advantageously extended beyond what measurable from their skeletal anatomy and described in historical sources.

Genomic makeup of historical horses

Neighbor-joining analysis of SNP variation confirmed previous phylogenetic reconstructions, in which PH and domestic horses formed two deep, distinct clades (Figure 2A). Admixture analyses also indicated strong genetic differences between these clades, which appeared already differentiated when considering two main genetic components (K = 2; Figures 2C, and S4). This was true for all horses considered, except KB7903, a known first-generation hybrid between a PH and a domestic horse that we used as control,[11] which showed balanced genetic contributions of both groups, as expected ((Figure 2C), denoted by an asterisk, *). In line with previous findings, the two historical PH specimens clustered together with their

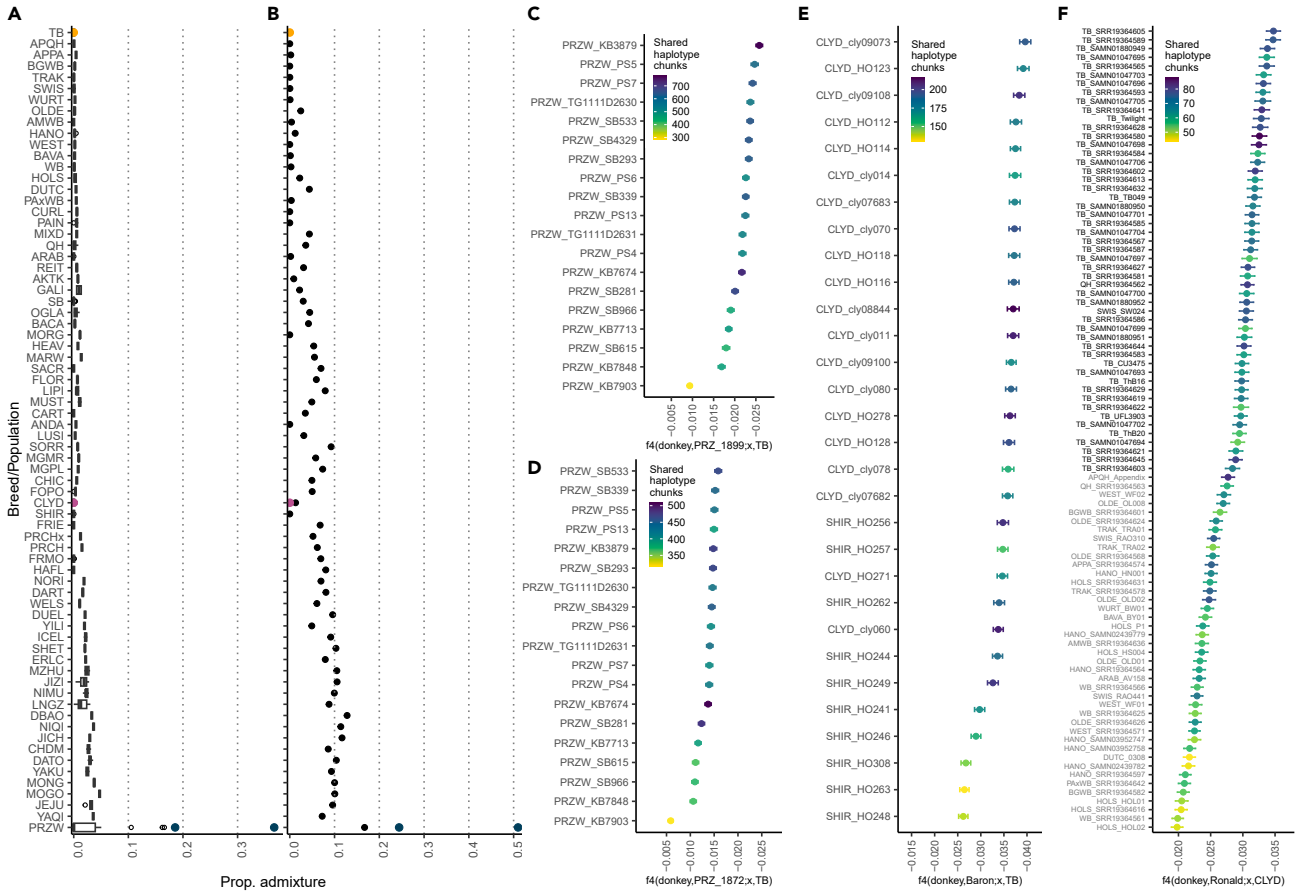


Figure 3. Przewalski Horse (PH) admixture and relationship between modern and ancient horses

The proportion of PH genetic admixture is shown for all domestic breeds and subpopulations, while the proportion of genetic admixture with domestic horses is shown for PH individuals (PRZW). The colored points represent the four historical samples, which are grouped with the rest of their breed. The number of individuals in each breed/subpopulation is detailed in Table S5.

(A) Admixture proportions were estimated by ADMIXTURE (version 4.1.1)[34] with 100 bootstrap pseudo-replicates (n = 432 individuals, n = 172,267 SNPs). B) Proportion of admixture as estimated by sourceFIND (version 2)[51] from 10 replicate runs (n = 432 individuals, n = 4,816,764 SNPs).

(C–F) Genetic sharedness between each ancient sample and modern individuals represented as f4-statistics (data points) estimated using the "qpDstat" function from Admixtools (version 751),[52,53] and colored according to the number of haplotype shared chunks identified by fineSTRUCTURE (version 4.1.1)[34] (n = 8,687,237 SNPs). The Thoroughbreds are shown in black on the y axis and the Warmblood/Quarter Horses in gray in panel F.

modern counterparts, with the 1872 Holotype branching out first and the 1899 Paratype branching second, as expected from their time difference. Furthermore, Baron clustered in the middle of modern Clydesdales, and showed a similar profile of genetic ancestry. Our analyses also show Clydesdales branching off from Shire horses, confirming previous reports of a close genetic relationship between the two breeds.[30] Finally, Ronald placed basal to all modern Thoroughbreds considered in this study, with 3 Quarter Horses and a Swiss Warmblood (Figure 2A). This reflects the open breeding structure of these two breeds, allowing introgression from Thoroughbred bloodlines.[30] The same genetic affinities were found using haplotype-based clustering analysis with fineSTRUCTURE[34] (Figure 2B). Combined, these analyses support the known breed and subpopulation structure of the four historical specimens analyzed, further confirming the validity of the genome imputation.

Evolutionary history of PH horse

Next, we further explored the horse genetic makeup, especially admixture between PH and domestic horses. We first found evidence for gene flow between both groups, representing 0.52%–4.92% of PH genetic ancestry in domestic horses when considering genotypes (Figure 3A), or 0.01%–11.28% when considering haplotype sharedness (Figure 3B). This agrees with previous reports of ~4%, 12%–13.7%, and ~6.8%, based on genome projections,[11] F4-ratios,[11] and f4-statistics,[12] respectively. Asian breeds showed higher

amounts of PH ancestry relative to non-Asian breeds (Wilcoxon rank-sum test, W = 24586, p-value <0.001), in line with PH subpopulations extending historically through Mongolia and Northern China.[50] Interestingly, the PH genetic contribution was heterogeneous across Asian breeds, and maximized in Mongolian breeds (Inner (MOGO) and outer (MONG) Mongolia). It remained substantial in breeds originating from China (Debao ponies, Ningqiang ponies, Jianchang ponies), Korea (Jeju horses), and Northern Siberia (Yakutian horses), but was also variable across Tibetan breeds, such as Langkazi, Jiangzi, and Datong horses.

In addition to tracking PH genetic ancestry in modern domestic horses, our genome panel also allowed us to directly infer the genetic contribution of the two historical PH specimens in the modern captive stock. The four modern PH previously reported to carry substantial portions of modern domestic horse ancestry showed lower genetic sharedness with both historical specimens (KB7848, SB615, KB7713, and SB966; Figures 3C and 3D, Table S6). While the 1872 Holotype appeared to have a similar level of genetic sharedness to all modern PH, the 1899 Paratype had differential genetic relatedness to contemporary lineages. This contribution was maximized in KB3879 (SB274) considering both genotype and haplotype sharedness, and remained relatively high in 3 specimens belonging to the purest lineage in the pedigree (SB4329, SB293, and SB533), with a previous study characterizing their genomes as "virtually devoid of [domestic] admixture."[11] Furthermore, no long runs of homozygosity (ROH) (>8 MB) were found in the two historical PH specimens (Figure 4), ruling out the presence of recent inbreeding at the time of their discovery in the 19th century. However, long ROH was a common feature of the modern PH specimens analyzed, in line with their documented history of inbreeding during captivity. Similarly, the total length of short and medium ROH increased in the modern PH captive stock relative to the two historical specimens, in agreement with their demographic collapse during the 20th century. The presence of short ROH in both historical PH genomes suggests, however, small effective population sizes over a long period of time prior to the 19th century.

Genomic consequences of modern breeding

The historical Clydesdale and Thoroughbred specimens present in our dataset revealed insights into the history of those two important breeds. This study also provided insights into the genetic structure of modern horses, particularly that no PH genetic contribution was found within Clydesdales and Thoroughbreds as well as across other modern domestic breeds showing no recorded history of PH admixture (Figures 3A and 3B).

Additionally, we identified three modern Clydesdale horses genetically closest to Baron: cly09073 and HO123, considering genotype sharedness, and cly08844 considering haplotype sharedness (Figure 3E). These three modern Clydesdales likely exhibit similar physical and behavioral traits to Baron as they show highly similar genotypes at the causative loci detailed in Figure 1B (Table S7). Notably, although all Clydesdales in this study were sampled from the United States, these three individuals have close ties with British bloodlines, particularly cly09073, who was directly imported to the United States from Scotland. Both HO123 and cly08844 were sired by the same British stallion with their maternal lineages also deriving from British bloodlines, suggesting that Baron's genetic influence is maximized in modern British Clydesdales compared to subpopulations found on other continents.

The extensive panel of 49 Thoroughbreds examined here show a continuum of genetic proximity to Ronald, which drops substantially in animals belonging to Quarter Horse (n = 3) and Warmblood breeds (n = 40), such as Oldenburgers, Trakeners, and Holsteiners (Figure 3F). This not only confirms Ronald as a Thoroughbred, despite his basal phylogenetic placement, but also that measurable amounts of Thoroughbred genetic variation are present in the Warmblood population, in line with studbook records. Importantly, our genetic analyses show that it is possible to identify modern individuals with higher genetic sharedness to important historical specimens. Similar to Clydesdales, the four Thoroughbreds with the highest genetic sharedness to Ronald also showed similar genotype frequencies at loci associated with traits of interest (Table S7). Applying this approach to other historical Thoroughbred champions, such as Triple Crown winners, could identify and incentivize modern bloodlines with the greatest genetic similarity to these individuals. This framework could also be applied to other sport horse breeds including for racing, eventing, and polo.

Finally, the amount of short (1–4 MB), medium (4–8 MB), and long (>8 MB) ROH was greater in modern Clydesdales and Thoroughbreds compared to Baron and Ronald, respectively (Figure 4). Pedigree records of Baron indicate two common ancestors in the both the 4th and 5th generation, where Ronald's shows four common ancestors in the 5th generation. The relatively limited presence of recent inbreeding reported in

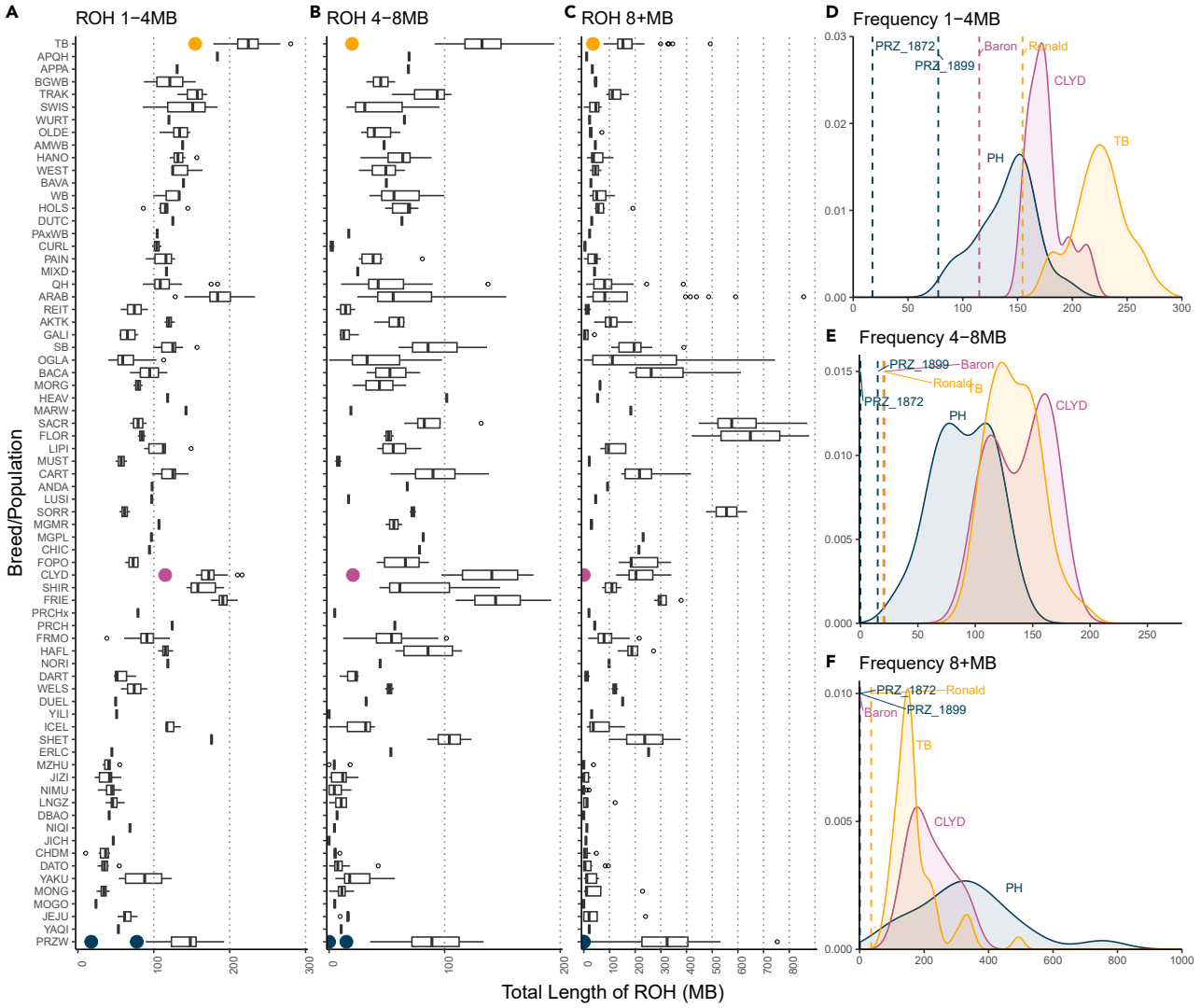


Figure 4. Inbreeding size distribution

(A–C) The total length of runs of homozygosity (ROH) in megabases (MB) estimated using PLINK (version 1.9)[54] and divided into three size classes: 1–4 MB, 4–8 MB, and 8+MB for each breed (n = 433 individuals, n = 2,802,133 transversion SNPs). The ROH in the ancient individuals are represented with colored points. The number of individuals in each breed/subpopulation is detailed in Table S5.

(D–F) Frequency distributions of ROH in the three size classes segregated into: Przewalski horse, Clydesdale, and Thoroughbred. The dotted lines represent the four ancient samples.

both pedigrees agrees with the low levels of genomic ROH, particularly long ROH which are indicative of recent inbreeding events. Higher levels of ROH in modern Clydesdales and Thoroughbreds indicate larger effective population sizes at the beginning of the 20th century than today in both breeds, a consequence of both modern breeding practices involving selection for desirable traits and inbreeding that is inherent due to closed stud books. Further analysis of a larger panel of historical horses will reveal the exact timing and context into which breeding practices most impacted the genomic makeup of breeds.

Conclusion

In this study, we reported an extensive panel of 8.7 million phased SNPs for 434 horses across 73 breeds and subpopulations. We used this panel to impute the genomes of four low-coverage historical specimens to high-coverage haplotypes. This included the newly sequenced genome of the famous historical Clydesdale Baron O'Buchylvie, and the publicly available genomes of the famous Thoroughbred Dark Ronald, the 1899 PH holotype and the 1872 PH paratype. Imputed genotypes provided us with insights into

behavioral and performance traits of Baron and Ronald, beyond the morphoanatomical, phenotypes measured from their remains. Imputation also provided fine-grained resolution into genetic ancestry and revealed the impact of changing breeding practices on admixture and inbreeding over the past century. It also helped characterize the genetic legacy of important historical specimens into modern subpopulations, which can provide new selection strategies for breeders and conservation biologists. The approach presented in this study, which is based on low-depth sequencing of historical specimens and genome imputation, can complement pedigree analysis to reconstruct the whole history of modern breeding in horses and other domestic animals.

Limitations of the study

Four historical horse genomes were used in this study which cannot fully capture the genetic changes across the many diverse breeds/populations through the past century. Sequencing and analyzing of additional historical horses in the future would add further insights into genetic changes in other horse breeds over this time period.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Sample collection, DNA extraction and sequencing of modern horses
  - Sample collection, DNA extraction and sequencing of ancient horses
  - Read alignment, rescaling, and trimming
  - Variant calling, quality control and phasing filtering of modern horses
  - Imputation of Baron and Ronald
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107104.

ACKNOWLEDGMENTS

The Genoscope sequencing platform is partly funded by the France Génomique National infrastructure, is part of the BUCEPHALE project and was funded by the French Government "Investissement d'Avenir" managed by Agence Nationale pour la Recherche (ANR-10-INBS-09).

This work was supported by CNRS and Université Paul Sabatier (AnimalFarm International Research Program, IRP), the France Génomique "Grands Projets" program (BUCEPHALE and MARENGO), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement 681605 - PEGASUS). Dr Yvette Running Horse Collin was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 890702 (MethylRIDE).

The bone sample of Baron O'Buchylverie was collected with thanks to BBC Scotland, The Canadian Film for Video Production Tax Credit, The Canadian Media Fund, Clyde Vet Group, Creative Scotland, Glasgow Life, Graven, Infield Fly Productions, Ontario Creates, Rogers Cable Network Fund, and Stream Scotland.

AUTHOR CONTRIBUTIONS

Conceptualization, L.O.; Methodology, E.T.T., L.O.; Formal Analysis, E.T.T., L.O.; Visualization, E.T.T.; Resources, A.F., R.S., Y.R.H.C., A.P., J-M.A., C.E., O.B., C.D., P.W., T.K., J.L.P., L.O.; Writing- Original Draft, L.O., E.T.T., Writing – Review and Editing, L.O., E.T.T. with input from J.L.P. and Y.R.H.C. and all co-authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 3, 2023
Revised: April 25, 2023
Accepted: June 8, 2023
Published: June 14, 2023

REFERENCES

1. Librado, P., Khan, N., Fages, A., Kusliy, M.A., Suchan, T., Tonasso-Calvière, L., Schiavinato, S., Alioglu, D., Fromentier, A., Perdereau, A., et al. (2021). The origins and spread of domestic horses from the Western Eurasian steppes. Nature *598*, 634–640. https://doi.org/10.1038/s41586-021-04018-9.

2. Kelekna, P. (2009). The Horse in Human History (Cambridge University Press).

3. Hendricks, B.L. (2007). International Encyclopedia of Horse Breeds (University of Oklahoma Press).

4. McShane, C., and Tarr, J.A. (2007). The Horse in the City: Living Machines in the Nineteenth Century (Johns Hopkins University Press).

5. FAOSTAT Food and Agriculture Organization. License: CC BY-NC-SA 3.0 IGO. Extracted from:. http://data.un.org/Data.aspx?d=FAO&f=itemCode%3A1096.

6. Fages, A., Hanghøj, K., Khan, N., Gaunitz, C., Seguin-Orlando, A., Leonardi, M., McCrory Constantz, C., Gamba, C., Al-Rasheid, K.A.S., Albizuri, S., et al. (2019). Tracking five millenia of horse management with extensive ancient genome time series. Cell *177*, 1419–1435.e31. https://doi.org/10.1016/j.cell.2019.03.049.

7. Orlando, L., and Librado, P. (2019). Origin and evolution of deleterious mutations in horses. Genes *10*, 649. https://doi.org/10.3390/genes10090649.

8. Bailey, E., Petersen, J.L., and Kalbfleisch, T.S. (2022). Genetics of Thoroughbred Racehorse Performance *10*, 131–150. https://doi.org/10.1146/annurev-animal-020420-035235.

9. Todd, E.T., Ho, S.Y.W., Thomson, P.C., Ang, R.A., Velie, B.D., and Hamilton, N.A. (2018). Founder-specific inbreeding depression affects racing performance in Thoroughbred horses. Sci. Rep. *8*, 6167. https://doi.org/10.1038/s41598-018-24663-x.

10. Hill, E.W., Stoffel, M.A., McGivney, B.A., MacHugh, D.E., and Pemberton, J.M. (2022). Inbreeding depression and the probability of racing in the Thoroughbred horse. Proc. Biol. Sci. *289*, 20220487. https://doi.org/10.1098/rspb.2022.0487.

11. Der Sarkissian, C., Ermini, L., Schubert, M., Yang, M.A., Librado, P., Fumagalli, M., Jónsson, H., Bar-Gal, G.K., Albrechtsen, A., Vieira, F.G., et al. (2015). Evolutionary genomics and conservation of the endangered Przewalski's horse. Curr. Biol. *25*, 2577–2583. https://doi.org/10.1016/j.cub.2015.08.032.

12. Gaunitz, C., Fages, A., Hanghøj, K., Albrechtsen, A., Khan, N., Schubert, M., Seguin-Orlando, A., Owens, I.J., Felkel, S., Bignon-Lau, O., et al. (2018). Ancient genomes revisit the ancestry of domestic and Przewalski's horses. Science *360*, 111–114. https://doi.org/10.1126/science.aao3297.

13. Ludwig, A., Reissmann, M., Benecke, N., Bellone, R., Sandoval-Castellanos, E., Cieslak, M., Fortes, G.G., Morales-Muñiz, A., Hofreiter, M., and Pruvost, M. (2015). Twenty-five thousand years of fluctuating selection on leopard complex spotting and congenital night blindness in horses. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20130386. https://doi.org/10.1098/rstb.2013.0386.

14. Frantz, L.A.F., Bradley, D.G., Larson, G., and Orlando, L. (2020). Animal domestication in the era of ancient genomics. Nat. Rev. Genet. *21*, 449–460. https://doi.org/10.1038/s41576-020-0225-0.

15. Bower, M.A., McGivney, B.A., Campana, M.G., Gu, J., Andersson, L.S., Barrett, E., Davis, C.R., Mikko, S., Stock, F., Voronkova, V., et al. (2012). The genetic origin and history of speed in the Thoroughbred racehorse. Nat. Commun. *3*, 643. https://doi.org/10.1038/ncomms1644.

16. 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. USA *111*, E5661–E5669. https://doi.org/10.1073/pnas.1416991111.

17. Librado, P., Gamba, C., Gaunitz, C., Der Sarkissian, C., Pruvost, M., Albrechtsen, A., Fages, A., Khan, N., Schubert, M., Jagannathan, V., et al. (2017). Ancient genomic changes associated with domestication of the horse. Science *356*, 442–445. https://doi.org/10.1126/science.aam5298.

18. Librado, P., Der Sarkissian, C., Ermini, L., Schubert, M., Jónsson, H., Albrechtsen, A., Fumagalli, M., Yang, M.A., Gamba, C., Seguin-Orlando, A., et al. (2015). Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. Proc. Natl. Acad. Sci. USA *112*, E6889–E6897. https://doi.org/10.1073/pnas.1513696112.

19. Orlando, L. (2020). The Evolutionary and Historical Foundation of the Modern Horse: Lessons from Ancient Genomics. Annu. Rev. Genet. *54*, 563–581. https://doi.org/10.1146/annurev-genet-021920-011805.

20. Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P.W., Ávila-Arcos, M.C., Fu, Q., Krause, J., Willerslev, E., Stone, A.C., and Warinner, C. (2021). Ancient DNA analysis. Nat. Rev. Methods Primers *1*, 14. https://doi.org/10.1038/s43586-020-00011-0.

21. Suchan, T., Chauvey, L., Poullet, M., Tonasso-Calvière, L., Schiavinato, S., Clavel, P., Clavel, B., Lepetz, S., Seguin-Orlando, A., and Orlando, L. (2022). Assessing the impact of USER-treatment on hyRAD capture applied to ancient DNA. Mol. Ecol. Resour. *22*, 2262–2274. https://doi.org/10.1111/1755-0998.13619.

22. Suchan, T., Kusliy, M.A., Khan, N., Chauvey, L., Tonasso-Calvière, L., Schiavinato, S., Southon, J., Keller, M., Kitagawa, K., Krause, J., et al. (2022). Performance and automation of ancient DNA capture with RNA hyRAD probes. Mol. Ecol. Resour. *22*, 891–907. https://doi.org/10.1111/1755-0998.13518.

23. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499–503. https://doi.org/10.1038/nature16152.

24. Ausmees, K., Sanchez-Quinto, F., Jakobsson, M., and Nettelblad, C. (2022). An empirical evaluation of genotype imputation of ancient DNA. G3 *12*, jkac089. https://doi.org/10.1093/g3journal/jkac089.

25. Todd, E.T., Tonasso-Calvière, L., Chauvey, L., Schiavinato, S., Fages, A., Seguin-Orlando, A., Clavel, P., Khan, N., Pérez Pardal, L., Patterson Rosa, L., et al. (2022). The genomic history and global expansion of domestic donkeys. Science *377*, 1172–1180. https://doi.org/10.1126/science.abo3503.

26. Erven, J.A.M., Çakirlar, C., Bradley, D.G., Raemaekers, D.C.M., and Madsen, O. (2022). Imputation of Ancient Whole Genome Sus scrofa DNA Introduces Biases Toward Main Population Components in the Reference Panel. Front. Genet. *13*, 872486. https://doi.org/10.3389/fgene.2022.872486.

27. Eggertsson, H.P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K.E., Jonasdottir, A., Jonasdottir, A., et al. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. Nat. Genet. *49*, 1654–1660. https://doi.org/10.1038/ng.3964.

28. Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. *103*, 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015.

29. Beeson, S.K., Mickelson, J.R., and McCue, M.E. (2020). Equine recombination map updated to EquCab3.0. Anim. Genet. *51*, 341–342. https://doi.org/10.1111/age.12898.

30. Petersen, J.L., Mickelson, J.R., Cothran, E.G., Andersson, L.S., Axelsson, J., Bailey, E., Bannasch, D., Binns, M.M., Borges, A.S., Brama, P., et al. (2013). Genetic diversity in the modern horse illustrated from genome-wide SNP data. PLoS One *8*, e54997. https://doi.org/10.1371/journal.pone.0054997.

31. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190. https://doi.org/10.1371/journal.pgen.0020190.

32. 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. https://doi.org/10.1038/ng1847.

33. Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol. Biol. Evol. *32*, 2798–2800. https://doi.org/10.1093/molbev/msv150.

34. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. *8*, e1002453. https://doi.org/10.1371/journal.pgen.1002453.

35. Alexander, D.H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinf. *12*, 246. https://doi.org/10.1186/1471-2105-12-246.

36. Hui, R., D'Atanasio, E., Cassidy, L.M., Scheib, C.L., and Kivisild, T. (2020). Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. Sci. Rep. *10*, 18542. https://doi.org/10.1038/s41598-020-75387-w.

37. Metzger, J., Kreft, O., Sieme, H., Martinsson, G., Reineking, W., Hewicker-Trautwein, M., and Distl, O. (2021). Hanoverian F/W-line contributes to segregation of Warmblood fragile foal syndrome type 1 variant PLOD1:c.2032G>A in Warmblood horses. Equine Vet. J. *53*, 51–59. https://doi.org/10.1111/evj.13271.

38. Grillos, A.S., Roach, J.M., de Mestre, A.M., Foote, A.K., Kinglsey, N.B., Mienaltowski, M.J., and Bellone, R.R. (2022). First reported case of fragile foal syndrome type 1 in the Thoroughbred caused by PLOD1 c.2032G>A. Equine Vet. J. *54*, 1086–1093. https://doi.org/10.1111/evj.13547.

39. Zhang, X., Hirschfeld, M., Schafberg, R., Swalve, H., and Brenig, B. (2020). Skin exhibits of Dark Ronald XX are homozygous wild type at the Warmblood fragile foal syndrome causative missense variant position in lysyl hydroxylase gene PLOD1. Anim. Genet. *51*, 838–840. https://doi.org/10.1111/age.12972.

40. Corbin, L.J., Pope, J., Sanson, J., Antczak, D.F., Miller, D., Sadeghi, R., and Brooks, S.A. (2020). An Independent Locus Upstream of ASIP Controls Variation in the Shade of the Bay Coat Colour in Horses. Genes *11*, 606. https://doi.org/10.3390/genes11060606.

41. Wagner, H.-J., and Reissmann, M. (2000). New polymorphism detected in the horse MC1R gene. Anim. Genet. *31*, 289–290. https://doi.org/10.1046/j.1365-2052.2000.00655.x.

42. Bellone, R.R., Holl, H., Setaluri, V., Devi, S., Maddodi, N., Archer, S., Sandmeyer, L., Ludwig, A., Foerster, D., Pruvost, M., et al. (2013). Evidence for a retroviral insertion in TRPM1 as the cause of congenital stationary night blindness and leopard complex spotting in the horse. PLoS One *8*, e78280. https://doi.org/10.1371/journal.pone.0078280.

43. Makvandi-Nejad, S., Hoffman, G.E., Allen, J.J., Chu, E., Gu, E., Chandler, A.M., Loredo, A.I., Bellone, R.R., Mezey, J.G., Brooks, S.A., and Sutter, N.B. (2012). Four loci explain 83% of size variation in the horse. PLoS One *7*, e39929. https://doi.org/10.1371/journal.pone.0039929.

44. Gu, J., MacHugh, D.E., McGivney, B.A., Park, S.D.E., Katz, L.M., and Hill, E.W. (2010). Association of sequence variants in CKM (creatine kinase, muscle) and COX4I2 (cytochrome c oxidase, subunit 4, isoform 2) genes with racing performance in Thoroughbred horses. Equine Vet. J. *42*, 569–575. https://doi.org/10.1111/j.2042-3306.2010.00181.x.

45. Hill, E.W., Gu, J., McGivney, B.A., and MacHugh, D.E. (2010). Targets of selection in the Thoroughbred genome contain exercise-relevant gene SNPs associated with elite racecourse performance. Anim. Genet. *41* (*Suppl 2*), 56–63. https://doi.org/10.1111/j.1365-2052.2010.02104.x.

46. Tozaki, T., Miyake, T., Kakoi, H., Gawahara, H., Sugita, S., Hasegawa, T., Ishida, N., Hirota, K., and Nakano, Y. (2010). A genome-wide association study for racing performances in Thoroughbreds clarifies a candidate region near the MSTN gene. Anim. Genet. *41* (*Suppl 2*), 28–35. https://doi.org/10.1111/j.1365-2052.2010.02095.x.

47. Blott, S., Cunningham, H., Malkowski, L., Brown, A., and Rauch, C. (2019). A Mechanogenetic Model of Exercise-Induced Pulmonary Haemorrhage in the Thoroughbred Horse. Genes *10*, 880. https://doi.org/10.3390/genes10110880.

48. Staiger, E.A., Albright, J.D., and Brooks, S.A. (2016). Genome-wide association mapping of heritable temperament variation in the Tennessee Walking Horse. Gene Brain Behav. *15*, 514–526. https://doi.org/10.1111/gbb.12290.

49. Hori, Y., Tozaki, T., Nambo, Y., Sato, F., Ishimaru, M., Inoue-Murayama, M., and Fujita, K. (2016). Evidence for the effect of serotonin receptor 1A gene (HTR1A) polymorphism on tractability in Thoroughbred horses. Anim. Genet. *47*, 62–67. https://doi.org/10.1111/age.12384.

50. Boyd, L., and Houpt, K.A. (1994). Przewalski's Horse: The History and Biology of an Endangered Species (State University of New York Press).

51. Chacón-Duque, J.C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuña-Alonzo, V., Barquera, R., Quinto-Sánchez, M., Gómez-Valdés, J., Everardo Martínez, P., Villamil-Ramírez, H., et al. (2018). Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. Nat. Commun. *9*, 5388. https://doi.org/10.1038/s41467-018-07748-z.

52. 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. https://doi.org/10.1534/genetics.112.145037.

53. Peter, B.M. (2016). Admixture, population structure, and f-statistics. Genetics *202*, 1485–1501. https://doi.org/10.1534/genetics.115.183913.

54. 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., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

55. Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes *9*, 88. https://doi.org/10.1186/s13104-016-1900-2.

56. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

57. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

58. 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. Bioinformatics *29*, 1682–1684. https://doi.org/10.1093/bioinformatics/btt193.

59. Skoglund, P., Northoff, B.H., Shunkov, M.V., Derevianko, A.P., Pääbo, S., Krause, J., and Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc. Natl. Acad. Sci. USA *111*, 2229–2234. https://doi.org/10.1073/pnas.1318934111.

60. 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. https://doi.org/10.1093/bioinformatics/btp352.

61. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. BMC Bioinf. *15*, 356. https://doi.org/10.1186/s12859-014-0356-4.

62. Browning, B.l., and Browning, S.R. (2016). Genotype imputation with millions of reference samples. Am. J. Hum. Genet. *98*, 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020.

63. Sieck, R.L., Fuller, A.M., Bedwell, P.S., Ward, J.A., Sanders, S.K., Xiang, S.H., Peng, S., Petersen, J.L., and Steffen, D.J. (2020). Mandibulofacial Dysostosis Attributed to a Recessive Mutation of CYP26C1 in Hereford Cattle. Genes *11*, 1246. https://doi.org/10.3390/genes11111246.

64. Seguin-Orlando, A., Donat, R., Der Sarkissian, C., Southon, J., Thèves, C., Manen, C., Tchérémissinoff, Y., Crubézy, E., Shapiro, B., Deleuze, J.-F., et al. (2021). Heterogeneous hunter-gatherer and steppe-related ancestries in late Neolithic and bell beaker genomes from present-day France. Curr. Biol. *31*, 1072–1083.e10. https://doi.org/10.1016/j.cub.2020.12.015.

65. Gamba, C., Hanghøj, K., Gaunitz, C., Alfarhan, A.H., Alquraishi, S.A., Al-Rasheid, K.A.S., Bradley, D.G., and Orlando, L. (2016). Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. Mol. Ecol. Resour. *16*, 459–469. https://doi.org/10.1111/1755-0998.12470.

66. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20130624. https://doi.org/10.1098/rstb.2013.0624.

67. Kalbfleisch, T.S., Rice, E.S., DePriest, M.S., Walenz, B.P., Hestand, M.S., Vermeesch, J.R., O'Connell, B.L., Fiddes, I.T., Vershinina, A.O., Saremi, N.F., et al. (2018). Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun. Biol. *1*, 197. https://doi.org/10.1038/s42003-018-0199-z.

68. Poullet, M., and Orlando, L. (2020). Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes. Front. Ecol. Evol. *8*. https://doi.org/10.3389/fevo.2020.00105.

69. Bates, D.W., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. BMJ Qual. Saf. *24*, 1–3. https://doi.org/10.18637/jss.v067.i01.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Details in Table S8	This study	N/A
Critical commercial assays		
Gentra Puregene Blood Kit	Qiagen	CAT# 158467
User Enzyme	New England BioLabs	Cat# M5505L
Deposited data		
Raw and analysed files	This study	ENA: PRJEB61662
VCF file of modern and ancient variants	This study	https://data.mendeley.com/datasets/wc4tsmy36y/draft?a=02e73170-eaef-456f-b481-03a27552948c
Software and algorithms		
AdapterRemoval2 (version 2.3.0)	Schubert et al.[55]	https://github.com/MikkelSchubert/adapterremoval
PALEOMIX (version 1.2.13.2)	Schubert et al.[16]	https://github.com/MikkelSchubert/paleomix
Bowtie2	Langmead and Salzberg[56]	https://github.com/BenLangmead/bowtie2
GATK (version 4.0.8.1)	McKenna et al.[57]	https://github.com/broadinstitute/gatk
mapDamage2	Jónsson et al.[58]	https://ginolhac.github.io/mapDamage/
PMDtools (version 0.60)	Skoglund et al.[59]	https://github.com/pontussk/PMDtools
Graphtyper (version 2.5.1)	Eggertsson et al.[27]	https://github.com/DecodeGenetics/graphtyper
BCFtools (version 1.8)	Li et al.[60]	https://github.com/samtools/bcftools
BEAGLE (version 5.1)	Browning et al.[28]	https://faculty.washington.edu/browning/beagle/b5_1.html
ANGSD (version 0.930)	Korneliussen et al.[61]	http://www.popgen.dk/angsd/index.php/ANGSD
BEAGLE (version 4.0)	Browning and Browning[62]	https://faculty.washington.edu/browning/beagle/b4_0.html
EIGENSOFT (version 6.1.4)	Patterson et al.,[31] Price et al.[32]	https://github.com/DReichLab/EIG
Tabix (version 1.8)	Li et al.[60]	http://www.htslib.org/doc/tabix.html
PLINK (version 1.9)	Purcell et al.[54]	https://www.cog-genomics.org/plink/
FastMe (version 2.1.4)	Lefort et al.[33]	http://www.atgc-montpellier.fr/fastme/binaries.php
fineSTRUCTURE (version 4.1.1)	Lawson et al.[34]	https://people.maths.bris.ac.uk/~madjl/finestructure/finestructure_info.html
ADMIXTURE (version 1.3.0)	Alexander and Lange[35]	https://dalexander.github.io/admixture/
SOURCEFIND (version 2)	Chacón-Duque et al.[51]	https://github.com/hellenthal-group-UCL/sourcefindV2
Admixtools (version 751)	Patterson et al.[52], Peter[53]	https://github.com/DReichLab/AdmixTools
Other		
Illumina HiSeq 4000	Illumina	
Illumina NovaSeq S4	Illumina	
Illumina Miniseq	Illumina	

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ludovic Orlando (ludovic.orlando@univ-tlse3.fr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- FASTQ data have been deposited at the European Nucleotide Archive under the project name PRJEB61662 and are publicly available as of the date of publication. All other genomes are available on public databases. Accession numbers are listed in the key resources table.
- The VCF with phased and imputed haplotypes for modern and ancient genomes was deposited on Mendeley data: https://data.mendeley.com/datasets/wc4tsmy36y/draft?a=02e73170-eaef-456f-b481-03a27552948c.
- Original code and any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Details of the historical samples used in this study can be found in the Results and Discussion section. We also downloaded 400 horse and 1 donkey genome from publicly available databases (key resources table). Due to no genomes of Shire or Clydesdale horses being publicly available, we sampled and sequenced 10 and 20 genomes, respectively from stud farms across the United States.

METHOD DETAILS

Sample collection, DNA extraction and sequencing of modern horses

We extracted and sequenced DNA from Blood samples of 20 Clydesdale and 10 Shire horses. All protocols were approved by the University of Nebraska-Lincoln's Institutional Animal Care and Use Committee. The DNA was extracted from EDTA blood samples using a modification of the Gentra Puregene Blood Kit (Qiagen) following the protocol described in.[63] The tubes were centrifuged at 2000x g (15 mins, 4 °C) to obtain the buffy coat. Next, 900 μL red blood cell lysis solution and 250 μL of buffy coat were combined, then vortexed. After incubation (5 mins, 22 °C), the samples were centrifuged at 13,000x g for 2 minutes and the supernatant discarded. A further 450 μL of red blood cell lysis solution was added to the pellet, then the tubes were again vortexed and incubated (5 mins, 22 °C). Samples were centrifuged (13,000x g, 2 mins), then the supernatant discarded before adding 6 μL Proteinase K and 900 μL lysis solution. The tubes were vortexed and incubated, then cooled on ice to room temperature. Protein precipitation solution was added to each tube (200 μL), before vortexing and storing on ice. The tubes were centrifuged (13,000x g, 2 mins) and the supernatant poured into a new tube with 800 μL of 100% isopropanol. The tubes were inverted 50 times, centrifuged (8,000x g, 2 mins), then the supernatant discarded, the pellet dried for 1 minute, and washed with 300 μL of 70% ethanol. The DNA pellet was dried for 15 minutes before rehydration in 100 μL of DNA hydration solution at (22 °C) overnight and then stored at 4 °C.

Library preparation and DNA sequencing of 6 Clydesdale samples was performed at the University of Minnesota Genomics Center (Minneapolis, MN, USA), using 125bp paired-end sequencing across nine lanes of an Illumina HiSeq 4000 platform. KAPA library preparation and 150bp paired-end sequencing of the other Clydesdales and all Shires was completed at Admera Health (south Plainfield, NJ, USA) on an Illumina NovaSeq S4 instrument.

Additionally, publicly available FASTQ files for 382 modern horses, 18 PH and 1 donkey were downloaded from the National Library of Medicine, the Genome Sequence Archive database and the European Nucleotide Archive (key resources table, Table S1), resulting in a total of 431 individuals in our modern dataset.

Sample collection, DNA extraction and sequencing of ancient horses

Publicly available FASTQ files for Ronald, a PH captured in 1872 (Holotype) and a PH captured in 1899 (Paratype) were downloaded from the European Nucleotide Archive (key resources table). Additionally, DNA of Baron was extracted from bone fragments and sequenced.

A total of six calcaneus bone fragments from Baron's skeleton were sampled at the Kelvingrove Art Gallery and Museum in Glasgow, Scotland using sterile personal protective equipment and a portable Dremel rotary device. The DNA extraction, library construction and shallow sequencing of Baron's DNA followed the procedures outlined by Seguin-Orlando and colleagues[64] and Librado and colleagues,[1] after a thin vernish layer was abrased mechanically. Drilling and extraction of DNA from the osseous material was carried out in the CAGT ancient DNA facilities (Toulouse, France). A total of 260-300mg of osseous material was powdered using the Mixel Mill MM200 (Retsch) Micro-dismembrator. DNA was then extracted using a procedure described by Gamba and colleagues,[65] which is tailored to facilitate the recovery of even the shortest DNA fragments. DNA extracts were treated with the USER (NEB) enzymatic cocktail to eliminate a fraction of postmortem DNA damage,[66] then DNA libraries were constructed using double-stranded DNA templates. The DNA library construction method relied on the ligation of two adapters containing unique internal indexes of 7 nucleotides, as presented by Rohland and colleagues.[66] One external index of 6 nucleotides was added during the PCR amplification of the library. The amplified and triple-indexed DNA libraries were purified and quantified, then pooled for shallow sequencing on a Miniseq Illumina instrument (high-output 80PE mode) at CAGT. A total of five DNA libraries were validated by sequencing from 2 independent bone fragments and were further sequenced on an Illumina HiSeq4000 instrument at the Genoscope (Evry, France) (75PE mode).

Read alignment, rescaling, and trimming

The sequences of the modern and ancient equids were trimmed, mapped, and filtered using the methods outlined by Librado and colleagues.[1] For each raw FASTQ file, AdapterRemoval2 (version 2.3.0) was used to demultiplex, collapse and trim the sequencing reads.[55] Paired end reads were collapsed if they showed sufficient sequence overlap and trimmed (truncated) if the ends showed insufficient qualities. All reads that were collapsed, truncated, and reads that were not collapsed (paired) were then run through the PALEOMIX pipeline (version 1.2.13.2),[16] which mapped the reads using Bowtie2[56] against the EquCab3 reference sequence.[67] Optimal parameters outlined by Poullet and Orlando[68] were used for mapping, and alignments were locally realigned around indels using IndelRealigner from GATK (version 4.0.8.1).[57] Finally, reads with mapping quality score inferior to 25, and sequence alignments shorter than 25 nucleotides, and/or representing PCR duplicates were removed.

For the ancient genomes, the presence of nucleotide mi-incorporation profiles characteristic of ancient DNA were checked using mapDamage2,[58] by randomly selecting 100,000 reads at the library level. There was an increase of C to T (G to A) mis-incorporation rates at read starts (read ends), which is expected in ancient samples due to postmortem cytosine deamination. The genomic positions proceeding read starts where higher in cytosines, in line with the excision of deaminated cytosines by the sequential activities of Uracil DNA glycosylase and Endonuclease VIII enzymes present in the USER mix.

The aligned reads that likely contain postmortem DNA damage were separated from those that did not using PMDtools (version 0.60)[59] using the parameters "—threshold 1; DAM" and "—upperthreshold 1; NODAM", respectively. The NODAM reads were directly trimmed for 5bp was directly at their ends. The DAM reads were rescaled using mapDamage, with all transitions penalized, then trimmed for 10bp at both ends and merged with the NODAM trimmed reads, as outlined by Librado and colleagues.[1]

Variant calling, quality control and phasing filtering of modern horses

Following the procedures from Todd and colleagues,[25] we called variants (single nucleotide polymorphisms (SNPs) and insertions or deletions of bases (INDELs)) in the mapped genomes present in our modern database. This procedure consisted in running Graphtyper (version 2.5.1)[27] in parallel for each individual chromosome (n = 45,543,765 variants). We then applied the recommended variant filters using the "vcffilter" function from Vcflib (version 1.0) (50): ABHet <0.0, ABHet > 0.33, BHom < 0.0, ABHom > 0.97, MaxAASR > 0.4, MQ > 30, INFO/LOGF > 0.5. We used GATK and BCFtools (version 1.8)[60] to apply the following genotype filters: Phred score > 20, minor allele frequency (MAF) $\geq$ 0.01, Hardy-Weinberg equilibrium *P*-value $\geq$ 0.001 and genotype missingness $\leq$ 0.2. We then removed the unassembled contigs, the

X chromosome, and INDELs variants leaving a total of 14,707,275 SNPs on 31 autosomes for further analysis.

We then phased the filtered variants in our modern dataset using BEAGLE (version 5.1),[28] with the recombination map generated by Beeson et al[29] (n= 1,737,839 positions).

Imputation of Baron and Ronald

We imputed the four ancient genomes using the panel of modern variants with a MAF $\geq$ 0.05 (n = 8,687,237 SNPs), and the methods developed by Hui and colleagues,[36] which were extensively tested by Todd and colleagues[25] (Figure S1).

To impute the ancient genomes, we genotyped the four individuals at all variants found in the modern reference panel using ANGSD (version 0.930)[61] with the following parameters: "-doMajorMinor 3 -GL 1 -doMaf 1 -snp_pval 1e-6 -doGeno 4 -doPost 1 -postCutoff 0.99 -remove_bads 1 -C 50 -minMapQ 25 -minQ 30 -uniqueOnly 1 -baq 1". We applied a pre-imputation filter of "GP $\geq$ 0.99" using BEAGLE (version 4.0)[62] to our ancient variant panel, using default parameters. We then imputed the genotypes of our ancient individuals with BEAGLE (version 5.1),[28] using only the filtered variants, the reference panel of modern equids and the recombination map generated by Beeson and colleagues[29] with all other parameters as default. We reapplied the filter "GP $\geq$ 0.99" post-imputation and merged the variants from ancient and modern individuals into a single file using the "merge" function in BCFtools. We ran this imputation pipeline using the BAM files both before and after trimming and rescaling. We found a high level in consistency of the imputed variants between the two datasets (Table S3, Figure S3). However, we used the imputed variants obtained from the rescaled and retrimmed files for further analysis as a conservative measure.

We tested the accuracy of this imputation pipeline using this reference panel by downscaling and re-imputing 9 high-coverage modern genomes from the same breeds/subpopulations as the historical samples (3 PH, 3 Clydesdales and 3 Thoroughbreds). We chose the 3 individuals from each breed that had the highest coverage and providing a representation of the diversity within the breed (i.e. the three samples were not placed side by side on the phylogenetic tree). The variants for each individual were downsampled to 50, 80, 90, 92, 94, 96 and 98% missingness using custom R scripts. The 9 individuals were then removed from the reference panel and the variants re-imputed following the same pipeline as described above. Imputation accuracy was assessed as the number of identical variant calls for all variants, heterozygous variants and homozygous variants (Table S2, Figure S2).

We then pseudo-haploidized the four ancient bam alignment files, conditioning on transversions found in the modern variant panel with a MAF $\geq$0.05 (n = 2,802,133) using the following parameters in ANGSD: "--dohaplocall 1 -doCounts 1 -doMajorMinor 3". We projected the ancient individuals onto the PCA of modern horses and imputed ancient genomes using the "lsqproject" function in the smartpca program from the EIGENSOFT package (version 6.1.4).[31,32] We found that the pseudo-haploidized samples and imputed samples were similarly placed on the PCA, indicating that their genomes had been imputed reliably and could be used for further analysis (Figure 1A).

QUANTIFICATION AND STATISTICAL ANALYSIS

We shortlisted a total of 105 SNP variants known to be causative or associated with appearance, racing performance, optimal distance, and behaviour in Clydesdale and Shire horses. Forty two of these variants were found in our panel and extracted using tabix (version 1.8)[60] (Table S4). Genotype frequencies for these variants in Thoroughbreds (n=49) and Clydesdales (n=20) were calculated using the "-hardy" parameter in PLINK (version 1.9)[54] (Figure 1B).

We constructed a neighbour joining tree by first calculating pairwise distances between all horses (n=434 individuals, n= 8,687,237 SNPs), using PLINK (version 1.9),[54] with the parameter "–distance square 1-ibs flat-missing". We then retrieved the tree topology by the bioNJ algorithm in FastME (version 2.1.4).[33] Node supports for the tree were estimated for 100 bootstrap pseudo-replicates with the topology refinement parameter (-n) using custom scripts (Figure 2A).

Next, we used fineSTRUCTURE (version 4.1.1)[34] to construct a haplotype-based clustering matrix of all modern and ancient PH and domestic horses (n=434 individuals, excluding the donkey). We converted

the variants in the VCF file present in all individuals (n= 4,816,764 SNPs) to the required input file formats using custom R scripts and the provided perl scripts from the fineSTRUCTURE package. After running fineSTRUCTURE, the co-ancestry matrix and genetic sharedness was plotted using the chunkcounts output file (Figures 2B and 3C–3F).

We estimated ancestry proportions using ADMIXTURE (version 1.3.0),[35] first thinning the variants (n = 8,687,237 SNPs) in PLINK using the parameter "–indep-pairwise 500 10 0.2", then calculating admixture proportions for models with K values between 2 and 8 using ADMIXTURE (n = 434 individuals (excluding the donkey) and n = 172,267 SNPs, after LD pruning). An optimal K value of 7 was estimated by comparing the cross-validation values of the different models (Figures 2C, and S4). We conducted 100 bootstrap pseudo-replicates at the optimal K to estimate the amount of ancestral component maximised in PH horses in domestic horse breeds (Figure 3A). The Thoroughbred horse used to create the reference panel (Twilight), the donkey, and the F1 PH/domestic hybrid was removed from the final plot (n = 432 individuals). We compared the proportion of this ancestral component between Asian and non-Asian breeds (excluding PH), using a Wilcoxon rank sum test in R (*P*-value significance threshold <0.05, see Results and Discussion section).

To estimate the PH ancestry in domestic horses, we painted the chromosomes of each individual using the Chromopainter function in fineSTRUCTURE using the same input files described above. We used six donor populations (PH, Thoroughbred, Cartujano, Clydesdale, Arabian, Quarter Horse). We used the chunklength file from Chromopainter for input into SOURCEFIND (version 2),[51] with the parameters: self.copy.ind=1, num. surrogates=6, exp.num. surrogates=1, num.slots=1000, num.iterations=200000, num.burnin=50000, and num.thin=5000. We ran SOURCEFIND 10 times, and then took the weighted mean of the highest posterior probability from each run for each breed/subpopulation to obtain final PH admixture values (Figure 3B, Table S5).

We calculated f4-statistics using the qpDstat function Admixtools (version 751)[52,53] to estimate the genetic similarity between Baron and modern Clydesdales and Shires (x) on the one hand (Easi, Baron; x , PH) (Figure 3E), and between Ronald and modern Thoroughbreds/Warmbloods (x) on the other hand (Easi, Ronald; x, PH) (Figure 3F). To estimate the genetic similarity between the 1899 Paratype (PRZ_1899) and the modern PH specimens (x), we estimated f4 statistics in the form of (Easi, PRZ_1899; x, TB) (Figure 3C, Table S6). We estimated the genetic similarity between the 1872 Holotype (PRZ_1872) and the modern PH specimens (x) using f4 statistics in the form of (Easi, PRZ_1872; x, modern, TB) (Figure 3D, Table S6). The correlation between the f4-statistics and haplotype sharing from fineSTRUCTURE were estimated using fitted linear models in R. The adjusted R^2 value for the models were 0.506, 0.668, 0.354, and 0.414 for PRZ_1872, PRZ_1899, Baron, and Ronald respectively (*P*-value significance threshold <0.05, the *P*-value was <0.001 for all four models).

Finally, we estimated the proportion of the genome in runs of homozygosity for each individual in our dataset using the "–homozyg" function in PLINK. To account for imputation errors leading the inaccurate calculations of ROH in low-coverage ancient samples, we allowed for up to 5 heterozygous variants in each sliding window and accounted for transversions only (n= 2,802,133 variants), with all other parameters as default. We then divided the ROHs for each individual into size distributions of: 1-4MB (distant inbreeding), 4-8MB (moderate inbreeding), 8+MB (recent inbreeding) (Figure 4). The squared correlation between ROH estimated with different numbers of heterozygotes was estimated using the lme package in R (Table S9).[69] We also calculated heterozygosity for each individual using the "—het" function in PLINK (Figure S5). The donkey and the F1 hybrid were removed from the final plots (n = 233 individuals). We calculated the ROH for the down-sampled and re-imputed modern genomes to estimate the accuracy of these methods after imputation of low-coverage genomes (Figure S6).