Supplementary information

Prehistoric genomes reveal the genetic foundation and cost of horse domestication

Mikkel Schubert[a,1], Hákon Jónsson[a,1], Dan Chang[b,1], Clio Der Sarkissian[a], Luca Ermini[a], Aurélien Ginolhac[a], Anders Albrechtsen[c], Isabelle Dupanloup[d,e], Adrien Foucal[d,e], Bent Petersen[f], Matteo Fumagalli[g], Maanasa Raghavan[a], Andaine Seguin-Orlando[a,h], Thorfinn Korneliussen[a], Amhed M.V. Velazquez[a], Jesper Stenderup[a], Cindi A. Hoover[i], Carl-Johan Rubin[j], Ahmed H. Alfarhan[k], Saleh A. Alquraishi[k], Khaled A.S. Al-Rasheid[k], David E. MacHugh[l,m], Ted Kalbfleisch[n], James N. MacLeod[o], Edward M Rubin[i], Thomas Sicheritz-Ponten[f], Leif Andersson[j], Michael Hofreiter[p], Tomas Marques-Bonet[q,r], M Thomas P Gilbert[a], Rasmus Nielsen[s], Laurent Excoffier[d,e], Eske Willerslev[a], Beth Shapiro[b], Ludovic Orlando[a,2].

[a]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350K Copenhagen, Denmark; [b]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA 95064; [c]The Bioinformatics Centre, Department of Biology, University of Copenhagen, 2200N Copenhagen, Denmark; [d]Institute of Ecology and Evolution, University of Berne, 3012 Berne, Switzerland; [e]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; [f]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark; [g]UCL Genetics Institute, Department of Genetics, Evolution, and Environment, University College London, London WC1E 6BT, United Kingdom; [h]National High-Throughput DNA Sequencing Center, University of Copenhagen, 1353K Copenhagen, Denmark; [i]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598; [j]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, SE-751 23 Uppsala, Sweden; [k]Zoology Department, College of Science, King Saud University, Riyadh 11451, Saudi Arabia; [l]Animal Genomics Laboratory, UCD School of Agriculture and Food Science, University College Dublin, Belfield, Dublin 4, Ireland; [m]UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin 4, Ireland; [n]Biochemistry and Molecular Biology, School of Medicine, University of Louisville, Louisville, KY 40292; [o]Department of Veterinary Science, Gluck Equine Research Center, University of Kentucky, Lexington, KY 40546; [p]Institute for Biochemistry and Biology, Faculty for Mathematics and Natural Sciences, University of Potsdam, 14476 Potsdam, Germany; [q]Institutició Catalana de Recerca i Estudis Avançats, Institut de Biologia Evolutiva (Universitat Pompeu Fabra/Consejo Superior de Investigaciones Cientificas), 08003 Barcelona, Spain; [r]Centro Nacional de Análisis Genómico, 08028 Barcelona, Spain; and [s]Departments of Integrative Biology and Statistics, University of California, Berkeley, CA 94720

[1] These authors contributed equally to this work.
[2] To whom correspondence should be addressed. E-mail: Lorlando@snm.ku.dk

Table of contents

Author contributions

- LO initially conceived and headed the project**.**
- BS provided samples.
- AHA, SAA, KAR, DEM, TMB, LEx, TG, EW, BS and LO provided reagents and material.
- MR, ASO, JS, CAH and LO did ancient DNA extractions and constructed Illumina ancient DNA libraries.
- MS, AG and BP performed read mapping, variant calling and generated genome alignments.
- LO did phylogenetic reconstruction and dating on mitochondrial genomes.
- HJ and AG estimated levels of inbreeding.
- AA estimated sequencing error rates and performed admixture tests based on the D-statistics.
- AG did Pairwise-Sequential Markovian Coalescent inference analyses, and examined DNA damage patterns, with input from MS.
- MS did phylogenomic reconstructions and dating on nuclear and Y-chromosomal DNA, and aggregated selection scan results.
- MS did PAML analyses, with input from HJ.
- HJ did TreeMix analyses.
- LO and AG estimated population divergence using F-statistics.
- MF performed the ∂a∂i analyses.
- AG and MS did the functional assessment of SNPs, and carried out functional enrichment.
- CDS and LEr did metagenomic profiling.
- MS, HJ, AF, ID, LEx performed genetic load analyses, with input from LO.
- AMVV did principal component analyses.
- HJ, AG and TK did the selection scans based on theta-Watterson estimates, with input from RN.
- DC and BS did the selection scans based on HMM and SNP BeadChip genotypes.
- MS, HJ, DC, CDS, LEr, AG, BS and LO wrote the Supplemental Note.
- LO wrote the manuscript, with input from MS, HJ, BS, DC, EW, KAR and all co-authors.

Supplementary Figures

Supplementary Tables

S1 Genome sequencing

S1.1 Sample information

The samples used in this study (listed in Supplementary Table S1) were previously described in Orlando *et al. 2013* (1), with the exception of the sample labeled “Icelandic (P5782)”, previously described in Andersson *et al.* 2012 (2). The domestic Thoroughbred horse (Twilight) corresponds to the individual originally sequenced for the assembly of the horse reference genome EquCab2.0 (3). It was subsequently deep-sequenced on Illumina platforms, representing an additional sequence dataset equivalent to 20.71× genome coverage (1).

In the present study, we characterize the genomic sequence of two ancient horse specimens, CGG10022 and CGG10023, at 24.27× and 7.36× coverage, respectively. The ancient specimens were excavated in Krasnoyarsk (Taymyr peninsula), Russia, and radiocarbon dated to 42,692 ± 891 (UBA-16478) and 16,099 ± 192 cal BP (UBA-16479) respectively, following calibration using Calib rev6.0.0 (4). These ages are roughly equivalent to 43 kyr and 16.5 kyr BP, in line with the ages used in the following section and reported in the main text. Dating was carried out at the 14Chrono Centre, Queen's University in Belfast (1).

We refer to samples by breed, followed by the associated name in parentheses, if any. Our ancient samples pre-date the earliest known evidence of horse domestication 5.5 kyr BP (5) and are therefore not associated with any domestic breed. These are referred to using the sample name alone (*i.e.* CGG10022 and CGG10023). To differentiate between the Icelandic horse sequenced by Orlando *et al.* 2013 (1) and the Icelandic horse sequenced by Andersson *et al.* 2012 (2), we refer to these as “Icelandic (unnamed)” and “Icelandic (P5782)”, respectively (Supplementary Table S1). The term “pre-domesticated” is used to refer to the samples CGG10022 and CGG10023, while the term “wild” is used to refer to all undomesticated horses, including the two pre-domesticated horses (CGG10022 and CGG10023), as well as the Przewalski’s horse, which has never been domesticated (1).

	Gender	Age (BCE)	Domestic	Coverage (×)	Reference
Arabian	Female	Modern	Yes	10.44	(1)
CGG10022	Female	42,012-40,094	No	24.27	(1)
CGG10023	Male	14,900-14,044	No	7.36	(1)
Domestic donkey (Willy)	Male	Modern	Yes	11.82	(1)
Icelandic (P5782)	Male	Modern	Yes	32.66	(2)
Icelandic (unnamed)	Male	Modern	Yes	8.10	(1)
Norwegian Fjord	Female	Modern	Yes	7.44	(1)
Przewalski's horse	Male	Modern	No	9.09	(1)
Standardbred	Male	Modern	Yes	11.58	(1)
Thoroughbred (Twilight)	Female	Modern	Yes	20.71	(1)

Supplementary Table S1. Samples used in this study

Where available, the name of the sample is shown in parentheses following the name of the breed, except for the two pre-domesticated samples, where only the sample name is shown; the age is given based on the calibrated radiocarbon dates. Coverage is given relative to the EquCab2.0 reference nuclear genome (see section S1.3).

S1.2 DNA sequencing

The pre-domesticated specimens (CGG10022 and CGG10023) had previously been sequenced to an average depth-of-coverage of 1.78× and 0.18× respectively (1), based on alignments against the EquCab2.0 horse reference genome (3); this sequencing information resulted from the deep-sequencing of one A-tailed and one blunt-end library for CGG10022, and one blunt-end library for CGG10023 (1).

For this study, new Illumina DNA libraries were built for these two samples and shotgun sequenced in order to increase the depth-of-coverage. We used previously described protocols (1, 6, 7), except that 1) 500nM of adapters were used for ligation and that 2) DNA libraries included one, two or three unique adapters. For CGG10022, a total of 3 new blunt end libraries were built using a method modified from Kircher and Meyer 2010 (8), where either regular or modified adapters were used. Modified adapters included of a unique 7-mer index located downstream of a block including Illumina sequencing primers and 5 random bases. Therefore, when one of such adapters was used, the first read mate started with the sequence of the 5 random bases and was followed by the 7-mer that could be used, together with the standard index read performed during Illumina multiplex runs, for identifying the library that was sequenced. The block of 5 random bases was included in order to enable proper base calling. When two of such adapters were used, the second read mate could be also used for library identification. More specifically, a first adapter was prepared following Kircher and Meyer 2010 (8) by mixing IS1F_03 5'-A*C*A*C*TCTTTCCCTACACGACGCTCTTCCGATCTNNNNNAAC*T*G*G*C-3' and IS3F_03 5'-G*C*C*A*GTTNNNNNAGATCGGAA*G*A*G*C-3', where * corresponds to a PTO bond and N to any base. The second adapter was prepared using the same procedure by mixing IS2R_54 5'-G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNTGG*C*A*T*G-3' and IS3R_54 5'-C*A*T*G*CCANNNNNAGATCGGAA*G*A*G*C-3'. Reads generated from such DNA libraries were identified requiring a strict sequence identity to the indexes selected and were further trimmed for their first 12 base positions.

We also built a series of six new DNA libraries for CGG10022 using the modified TruSeq DNA library building procedure described in Pedersen et al. 2014 (9). Two of those DNA libraries were PCR amplified using one amongst three possible PCR amplification conditions. The first amplification conditions consisted of a 25 µl volume reaction with 10 µl of DNA library, and 5 units Ampli*Taq* Gold (Life Technologies), 1× Gold Buffer, 4 mM $MgCl_2$, 1 mg/ml BSA, 62.5 µM of each dNTP, 0.3 µM of Primer 1.0 (5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC -3') and 0.3 µM of Primer 2.0 (5'-CAA GCA GAA GAC GGC ATA CGA GAT-3'). PCR cycling conditions consisted of initial denaturation for 10 min at 92°C, followed by 15 cycles of 30 sec denaturation at 92°C, 30 sec annealing at 65°C and 3 min elongation at 72°C. There was a final 7 min elongation step at 72°C. The second amplification conditions were similar, except that Accuprime *Pfx* was used instead of *Taq* Gold. A total of 3.125 enzyme units and 1 mM of $MgSO_4$ were used in the reaction mix. PCR cycling conditions were 2 min at 95°C, 15 cycles of 15 seconds at 95°C, 30 seconds at 60°C and 40 seconds at 68°C, followed by a final elongation of 7 min at 68°C. Finally, the third amplification conditions were identical to the second ones, except that the amplification was performed in an emulsion by mixing the PCR preparation to 150 µl of a mixture consisting of Tegosoft DEC (73% vol.; Evonik), mineral oil (20% vol.; Sigma) and ABIL WE 09 (7% vol; Evonik). Following vortexing for 2 min at maximum speed, the emulsion was then divided into two PCR tubes for performing PCR amplification. Post-amplification, the tubes from each sample were pooled again before the emulsion was broken by the addition of 1ml isobutanol to each tube. Amplified DNA libraries were then purified using Qiagen Minelute and adding 250 µl of buffer PB. The final elution was performed in 20 µL EB following 15 min incubation at 37°C.

A total of 8 new A-tailed libraries, together with one new blunt-end library, were built for CGG10023 (LOd, see Supplementary Table S3), following the procedures described in Orlando et al. 2013 (1). Indexed libraries were sequenced using either the Illumina HiSeq2000 platform or the Solexa GAIIx (detailed in Supplementary Table S2 and Supplementary Table S3). DNA contamination from the laboratory and reagents were monitored through mock extractions and amplification blanks. All controls were negative.

The sequencing data generated for this study is available from the European Nucleotide Archive under accession number PRJEB7537.

Identifier	Library	PCR	# Lanes	# Raw reads	# Filtered reads	# Collapsed pairs
ACTTGA *(1)*	KM	TG	1SE HiSeq	50,078,895	49,976,516	
CGTAGT *(1)*	AT	PL	1SE GA	18,999,000	17,711,195	
CGTAGT *(1)*	AT	PL	1SE HiSeq	28,229,336	26,545,981	
CTTGTA	TS	AP	1SE HiSeq	62,408,084	60,233,779	
CTTGTA	TS	APem	1SE HiSeq	28,567,181	27,607,083	
CTTGTA	TS	TG	1PE HiSeq	68,766,184	66,836,830	29,340,618
CTTGTA	TS	TG	8SE HiSeq	385,968,696	372,028,432	
TGACCA	TS	APem	1SE HiSeq	33,367,380	32,343,270	
TGACCA	TS	AP	1PE HiSeq	54,490,950	53,232,764	23,631,043
TGACCA	TS	AP	8SE HiSeq	338,500,525	326,790,205	
TGCAGG	KM	TG	1PE HiSeq	64,079,317	63,373,822	63,093,868
ACTGCC	KM^{1}	TG	1PE HiSeq	42,672,686	42,451,869	20,320,853
GCAACG	KM^{2}	TG	1PE HiSeq	41,767,892	41,577,160	18,679,590
Totals				1,217,896,126	1,180,708,906	155,065,972

Supplementary Table S2. Sequencing information concerning CGG10022

For the sequenced lanes, SE stands for single-end and PE for paired-end; the first 2 libraries were generated by Orlando et al. *2013, and the second was sequenced both on a GAIIx Illumina platform and a HiSeq 2000 (1); see Supplementary Section S1.2 for a break-down. Reads were quality filtered (# Filtered) and PE pairs were collapsed (# Collapsed) using AdapterRemoval (Supplementary Section S1.3). AT = A-tailing DNA library building procedure described in Orlando* et al. *2013 (1). TS = modified TruSeq DNA library building procedure described in Pedersen et al. 2014* (9). *KM = DNA library building procedure from Meyer and Kircher 2010* (8). *KM^{1} = same as KM, except that adapters IS1F_03 and IS3F_03 were used. KM^{2} = same as KM^{1}, except that adapters IS2R_54 and IS3R_54 were also used. TG = DNA library PCR amplification with Taq Gold. PL = DNA library PCR amplification with Platinum Taq DNA polymerase Hifi (Life Technologies). AP = DNA library PCR amplification with AccuPrime. APem = same as AP, except that the amplification was performed in emulsion.*

Identifier	Library	PCR	# Lanes	# Raw reads	# Filtered reads	# Collapsed pairs
Lib1	AT	PL	2 SE GA, 7 SE HiSeq	885,248,698	820,985,205	
Lib1	AT	PL	9 PE HiSeq	3,736,172,758	3,631,029,592	1,691,109,372
Lib2	AT	PL	2 SE GA	51,121,936	49,132,328	
Lib3	AT	PL	2 SE GA	55,407,117	52,786,584	
Lib4	AT	PL	2 SE GA	57,995,741	55,112,384	
Lib9	AT	PL	1 SE GA	33,774,214	20,514,037	
Lib10	AT	PL	1 SE GA, 16 SE HiSeq	1,523,064,414	1,380,426,783	
Lib11	AT	PL	1 SE GA	20,186,952	19,796,775	
Lib12	AT	PL	7 SE GA 8 SE HiSeq	899,425,110	877,664,959	
Lib12	AT	PL	9 PE HiSeq	1,858,699,964	1,840,358,776	896,290,345

LOb *(1)*	KM	TG	2 PE HiSeq	448,763,138	434,297,959	211,231,343
LOc *(1)*	KM	TG	2 PE HiSeq	423,988,764	396,247,048	192,241,677
LOd	KM	TG	2 PE HiSeq	781,736,652	766,567,690	379,264,045
Totals				10,775,585,458	10,344,920,120	3,370,136,782

Supplementary Table S3. Sequencing information concerning CGG10023

For the sequenced lanes, SE stands for single-end and PE for paired-end; the libraries LOb and LOc reflect the combined sequencing effort by this study and the study by Orlando et al. *2013 (1); see Supplementary Section S1.2 for a break-down. Reads were quality filtered (# Filtered) and PE pairs were collapsed (# Collapsed) using AdapterRemoval (Supplementary Section S1.3). See caption for* Supplementary Table S2*, for further details about DNA libraries and PCR amplification conditions.*

S1.3 Read alignment against reference genomes

The alignment procedure was based on previously described methods [Orlando *et al. 2013,* Supplementary Information section 4.2.b] (1), as implemented by the PALEOMIX pipeline (10), with small modifications.

For the pre-domesticated specimens CGG10022 and CGG10023, adapter sequences were trimmed from raw Illumina reads using AdapterRemoval (11) v1.2, using the same parameters as described in Orlando *et al.* 2013 (1). For the samples based on previously published sequence data, we used reads that had already been trimmed for the previous studies (1, 2). In both cases, paired-ended reads in which the sequences overlapped by at least 11bp (with a maximal number of one mismatch, or one third of the overlapping region when overlapping for more than 11 bp) were collapsed into a single consensus sequence (1). For the two pre-domesticated horses, paired-ended reads that did not overlap (and where thus not collapsed by AdapterRemoval) were excluded as likely modern contamination.

Trimmed reads were mapped against the horse reference genome EquCab2.0 (3) excluding the mitochondrial genome, but unlike previously (1) also including chromosome Un (unplaced contigs) to account for reads that could match several genomic locations. In addition, mapping was carried out separately for the horse reference mitochondrial genome (Accession Nb. NC_001640). In both cases, mapping was carried out using BWA (12) v0.5.10 with default parameters, except that the seed-region was disabled for the two pre-domestic specimens (13).

Unmapped reads and reads with a mapping quality less than to 25 were discarded. Duplicates were identified and removed per library using the 5'-end mapping coordinate in the case of singleton reads (SE reads, and PE reads where one mate was discarded due to low quality), and using both external coordinates for paired (PE) and collapsed reads. Singleton and PE reads were filtered using *MarkDuplicates* from the Picard Tools suite (http://picard.sourceforge.net), and collapsed reads were identified and filtered using a modified version of the *FilterUniqueBAM.py* script kindly provided by Martin Kircher, included with the PALEOMIX pipeline (10). Final BAM files were realigned around indels using the Genome Analysis Toolkit (14) and processed using SAMTools *(15) calmd* to update edit distances and MD tags.

Raw sequencing reads for the sample "Icelandic (P5782)" were kindly provided by Pr. Leif Andersson and Dr. Calle Rubin. Following the procedure outlined above, an average depth-of-coverage of 33.22× was obtained for this sample, in agreement with the coverage reported in the original publication (2).

Mapping results are reported in Supplementary Table S4. Note that the number of processed reads given corresponds to the number of reads following filtering and collapsing of mate pairs. For CGG10022 and CGG10023, only SE reads and collapsed PE reads are included. The amount of endogenous DNA content was estimated for each library for the two pre-domesticated horses (Supplementary Table S5 and Supplementary Table S6).

Following mapping, read length distributions were determined for the two pre-domesticated specimens based on the length of aligned bases for collapsed reads (Supplementary Figure S1). These reads represent DNA inserts sequenced over their full length, thus giving access the length distribution of the ancient DNA fragments (1). A ca. 10 bp periodicity was observed for both specimens, especially striking for the older sample CGG10022. This pattern is likely to result from the DNA helix complete turn (10 bp) and the corresponding nucleosome protection footprint, as recently suggested by Pedersen *et al* 2014 (9). Interestingly, both size distribution profiles appear in phase and the ca. 10 bp periodicity vanishes for insert sizes longer than 160 bp, which is close to the length of the DNA protected by the core nucleosome.

	# Processed reads	# Reads mapped	# Bases aligned	Coverage (X)
Arabian	584,787,811	316,578,043	25,933,539,961	10.44
CGG10022	1,087,238,002	683,106,199	60,299,305,610	24.27
CGG10023	6,974,783,338	229,138,553	18,295,001,670	7.36
Domestic donkey (Willy)	710,057,244	390,728,563	29,376,890,451	11.82
Icelandic (P5782)	1,012,659,095	837,658,548	81,160,940,284	32.66
Icelandic (unnamed)	825,399,262	213,360,593	20,135,482,802	8.10
Norwegian Fjord	641,198,340	227,416,322	18,496,722,991	7.44
Przewalski's horse	560,502,966	254,870,164	22,592,812,840	9.09
Standardbred	577,987,318	340,320,656	28,782,364,817	11.58
Thoroughbred (Twilight)	545,128,052	473,669,917	51,442,515,528	20.71

Supplementary Table S4. Sequence mapping statistics for modern and ancient samples

Trimmed sequences were mapped to the EquCab2.0 reference genome using BWA v0.5.10. The number of bases aligned and the coverage is given relative to the EquCab2.0 reference genome, including chrUn, but excluding chrM. Hits were tabulated subsequent to quality filtering and removal of PCR duplicates (see Supplementary Section S1.3). Processed reads are reported after quality filtering and collapsing of PE reads; for the pre-domesticated horses, only SE reads and collapsed PE reads are used.

Identifier	Library	PCR	# Lanes	# Processed reads	# Mapped reads	% Mapped
ACTTGA *(1)*	KM	TG	1SE HiSeq	49,976,516	28,449,687	0.569
CGTAGT *(1)*	AT	PL	1SE GA, 1SE HiSeq	44,257,176	25,979,640	0.587
CTTGTA	TS	AP	1SE HiSeq	60,233,779	42,656,693	0.708
CTTGTA	TS	APem	1SE HiSeq	27,607,083	19,787,353	0.717
CTTGTA	TS	TG	1PE HiSeq, 8SE HiSeq	409,524,644	263,105,917	0.642
TGACCA	TS	APem	1SE HiSeq	32,343,270	22,822,280	0.706
TGACCA	TS	AP	1PE HiSeq, 8SE HiSeq	356,391,926	231,166,727	0.649
TGCAGG	KM	TG	1PE HiSeq	64,359,271	29,843,595	0.464
ACTGCC	KM	TG	1PE HiSeq	21,464,154	10,443,470	0.487
GCAACG	KM	TG	1PE HiSeq	21,080,183	8,850,837	0.420

Totals				1,087,238,002	683,106,199	0.628

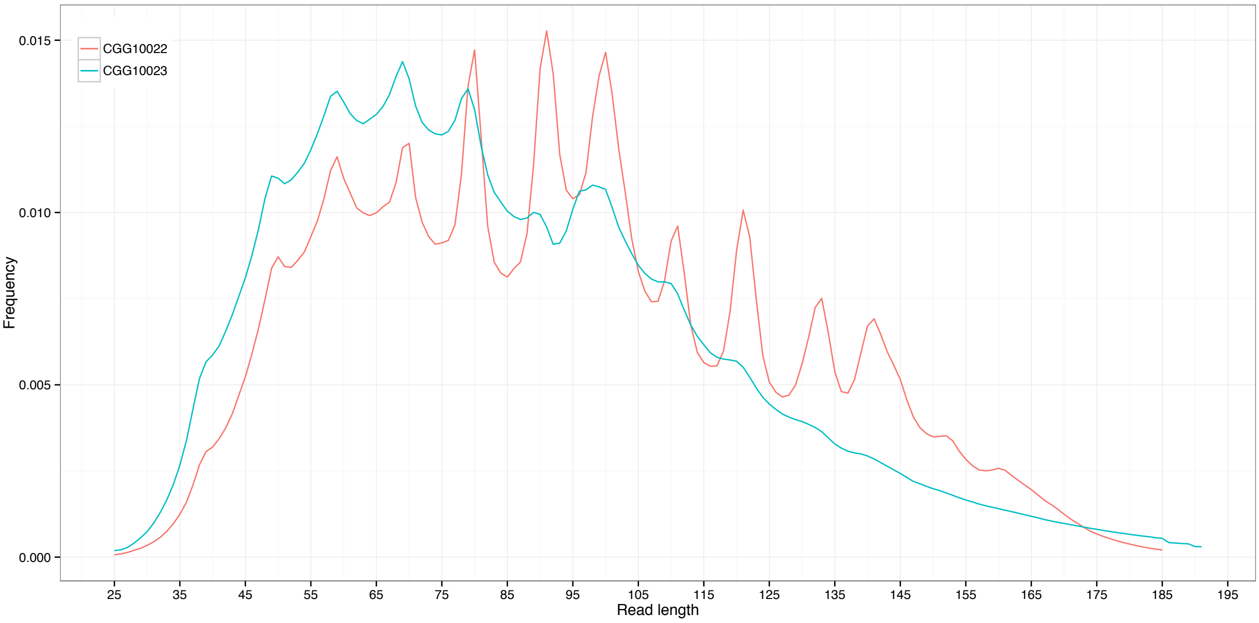
Supplementary Table S5. Estimation of endogenous DNA content of libraries for CGG10022

The endogenous DNA content was estimated as the fraction of high quality, collapsed reads that mapped to the EquCab2.0 reference genome with a minimum mapping quality of 25, excluding reads identified as PCR duplicates (see Supplementary Section S1.3). Thus this number provides a conservative number of the fraction of informative reads in the libraries.

Identifier	Library	PCR	# Lanes	# Processed reads	# Mapped reads	% Mapped
Lib1	AT	PL	2 SE GA, 7 SE HiSeq, 9 PE HiSeq	2,760,905,425	64,836,298	0.023
Lib2	AT	PL	2 SE GA	49,132,328	4,224,340	0.086
Lib3	AT	PL	2 SE GA	52,786,584	4,407,409	0.083
Lib4	AT	PL	2 SE GA	55,112,384	5,797,519	0.105
Lib9	AT	PL	1 SE GA	20,514,037	1,001,756	0.049
Lib10	AT	PL	1 SE GA, 16 SE HiSeq	1,380,426,783	21,571,027	0.016
Lib11	AT	PL	1 SE GA	19,796,775	350,937	0.018
Lib12	AT	PL	7 SE GA 8 SE HiSeq, 9 PE HiSeq	1,821,733,390	34,390,599	0.019
LOb *(1)*	KM	TG	2 PE HiSeq	434,297,959	18,618,500	0.043
LOc *(1)*	KM	TG	2 PE HiSeq	396,247,048	4,064,389	0.010
LOd	KM	TG	2 PE HiSeq	766,567,690	69,875,779	0.091
Totals				7,757,520,403	229,138,553	0.030

Supplementary Table S6. Estimation of endogenous DNA content of libraries for CGG10023

The endogenous DNA content was estimated as the fraction of (collapsed) quality / PCR filtered reads that mapped to the EquCab2.0 reference genome with a minimum mapping quality of 25 (see Supplementary Section S1.3). Thus this number provides a conservative number of the fraction of informative reads in the libraries.



Supplementary Figure S1. Length distribution of mapped reads from pre-domesticated horses

The length-distributions of aligned sequences (excluding clipped bases) for collapsed PE reads mapped to the EquCab2.0 reference genome, for pre-domesticated horses CGG10022 and CGG10023.

S1.4 Sample-wise error rates

Error rates were determined for each sample as described in Orlando et al. 2013 [see their Supplementary section S4.4](1), excluding the Thoroughbred (Twilight), as it represents the individual from which the reference sequence was built (see Supplementary Section S1.1). Briefly, the samples were mapped to the EquCab2.0 reference genome, a single high quality horse was selected to represent the “perfect genome”, and the outgroup was used to polarize alleles. As each horse should be equidistant from the outgroup, an increase in derived alleles relative to the “perfect genome” can be interpreted as an increase in the error rate.

For a given sample s let A_s and $a_s$denote the true and observed number of ancestral alleles respectively. Similary let D_s and $d_s$denote the true and observed number of derived alleles. Given an error rate ϵ_s the expected number of derived allele is

$$\mathbb{E}[d_s] \ = D_s(1-\epsilon_i) + A_s\epsilon_s.$$

In order to obtain estimates for the true number of ancestral and derived alleles we use the “perfect genome” p such that $\widehat{A_s} = a_p$ and $\widehat{D_s} = d_p$. An estimate of the overall error rate can then be obtained as

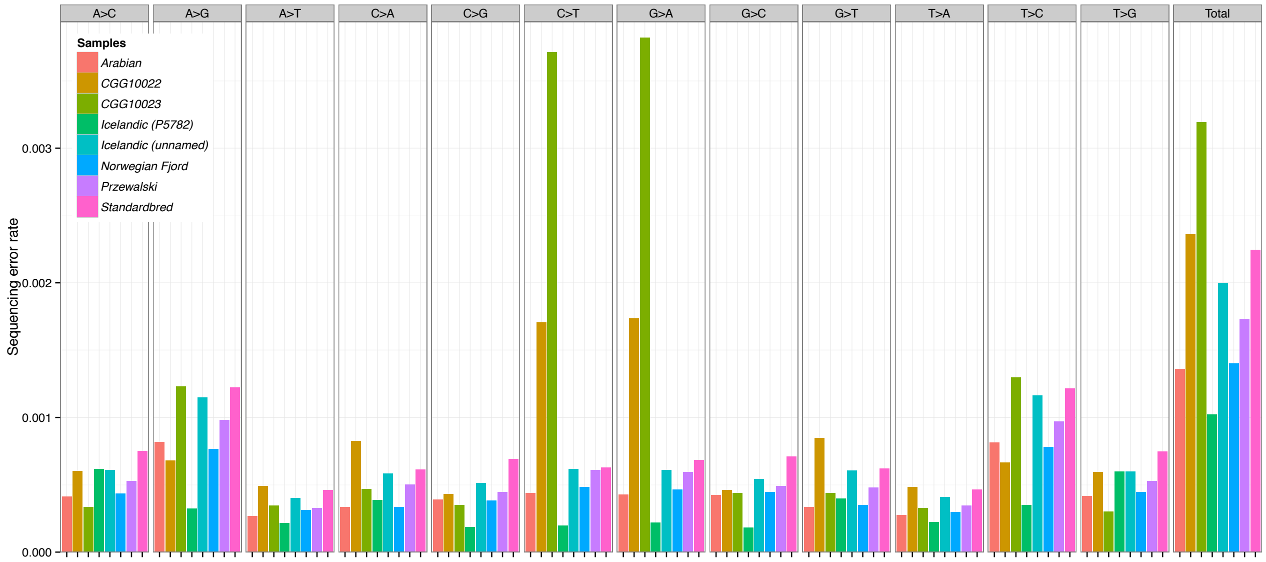
$$\hat{\epsilon}_i = {}^{d_i - d_p}\!\big/_{a_p - d_p}.$$

The type specific error rates are estimated based on a maximum likelihood model based on similar assumptions, which are described in details in Orlando *et al*. 2013 (1).

The Icelandic (P5782) was chosen as the “perfect genome”, as it is the sample with the highest coverage, and the domestic donkey (Willy) was used to determine the ancestral allele (see above).

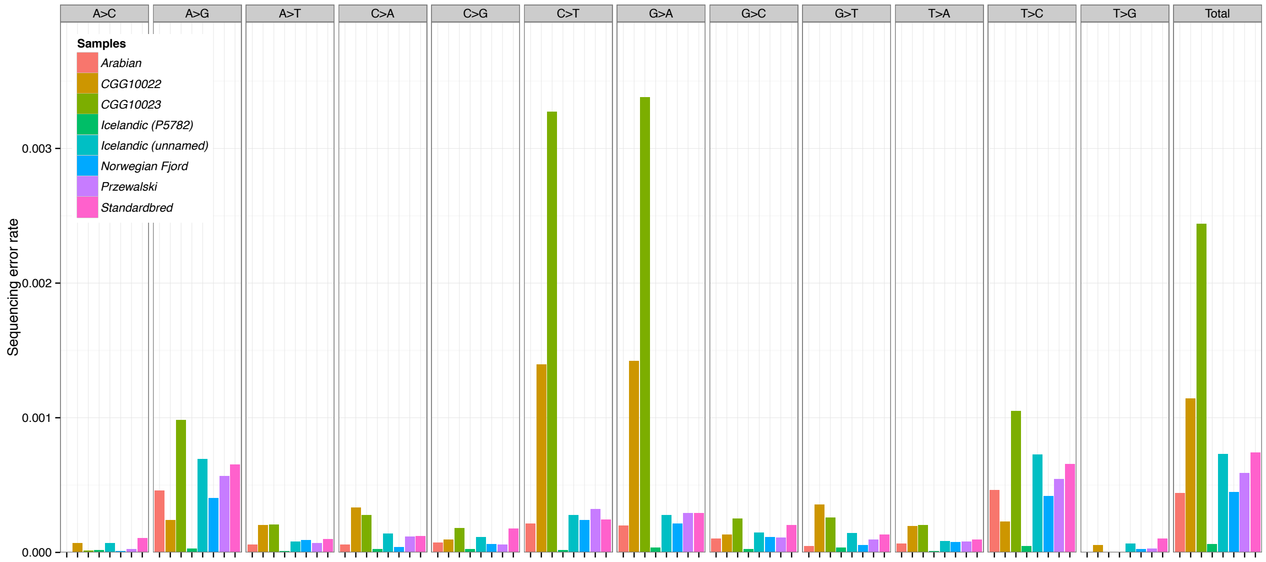
A single base was sampled at each position for each sample in order to derive the estimates; for this purpose, the Icelandic (P5782) and the domestic donkey (Willy) were filtered using the criteria that the Phred encoded mapping score must be at least 30, and that the Phred encoded base quality must be at least 35 (Supplementary Figure S2). Based on the error rates estimated based on these criteria, the remaining samples were filtered using a minimum mapping quality of 30, and a minimum base quality of 25 (Supplementary Figure S3).

Following quality filtering, we observed error rates below 0.1% per base for all mutation types, except for samples CGG10022 and CGG10023 where C→T and G→A error rates were found to be 0.14% and 0.14% for CGG10022, and 0.34 and 0.33% for CGG10023, respectively. Those substitutions correspond to nucleotides that were misincorporated during DNA library amplification, due to the presence of cytosine residues deaminated into uracil residues *post-mortem* (16) (see also Supplemental section S1.5). Overall, we estimate that the two ancient genomes characterized in this study have an average error rate of 0.24% (CGG10023) and 0.11% (CGG10022) per base.



Supplementary Figure S2. Error rates of wild and domestic horses without quality filtering

Type specific error rates estimated relative to the Icelandic (P5782) sample, assuming that this sample represents a "perfect genome", and using the domestic donkey (Willy) to determine the ancestral alleles. The "perfect genome" and the outgroup was filtered using a minimum mapping score of 30, and a minimum base-quality score of 35 (both Phred-scaled); the remaining samples were not quality-filtered.

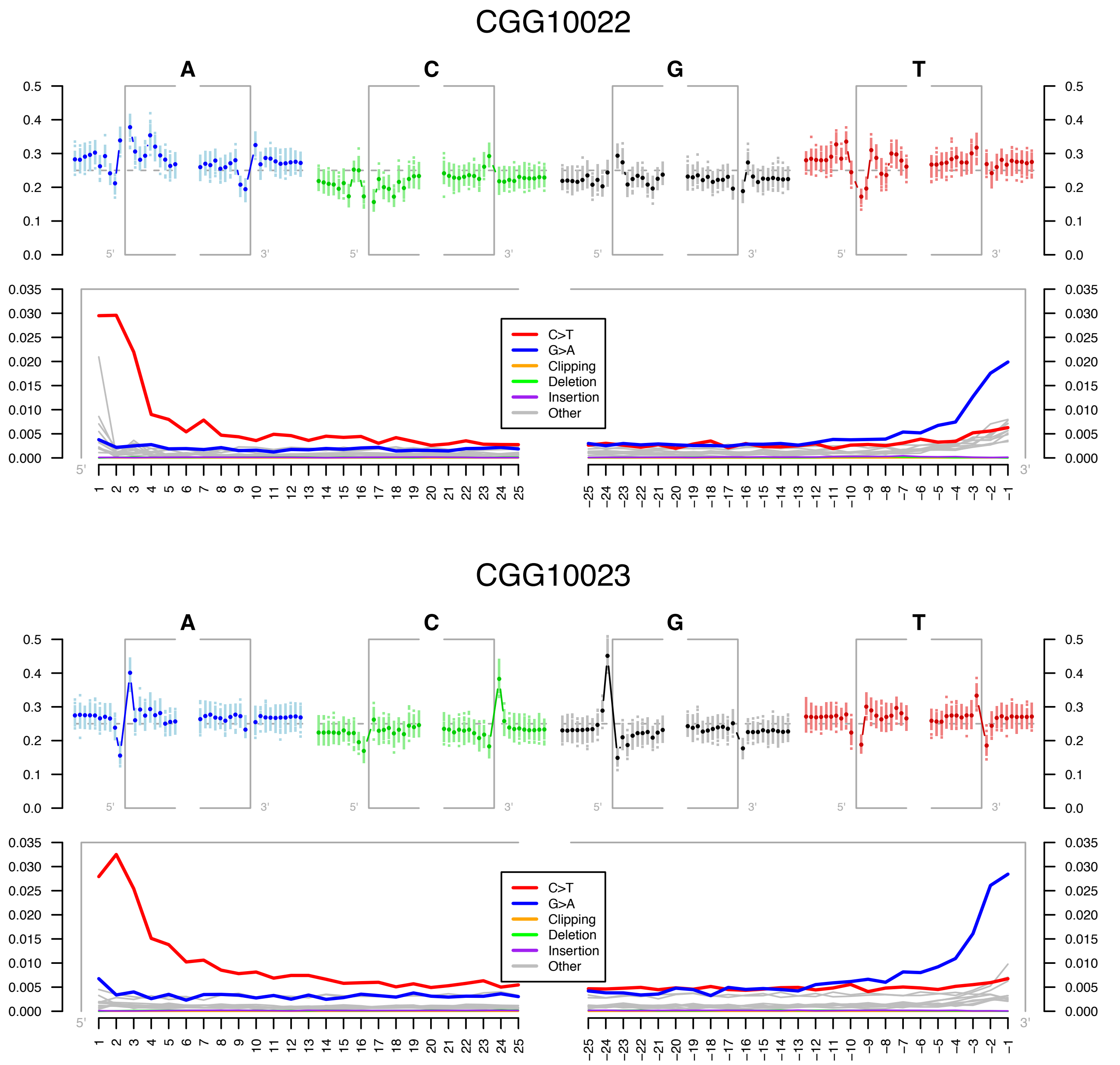


Supplementary Figure S3. Error rates of wild and domestic horses with quality filtering

Type specific error rates estimated relative to the Icelandic (P5782) sample, assuming that this sample represents a "perfect genome", and using the Domestic Donkey (Willy) to determine the ancestral alleles. The "perfect genome" and the outgroup was filtered using a minimum mapping score of 30, and a minimum base-quality score of 35 (both Phred-scaled); the remaining samples were filtered using a minimum mapping quality of 30, and a minimum base quality score of 25 (Phred scaled).

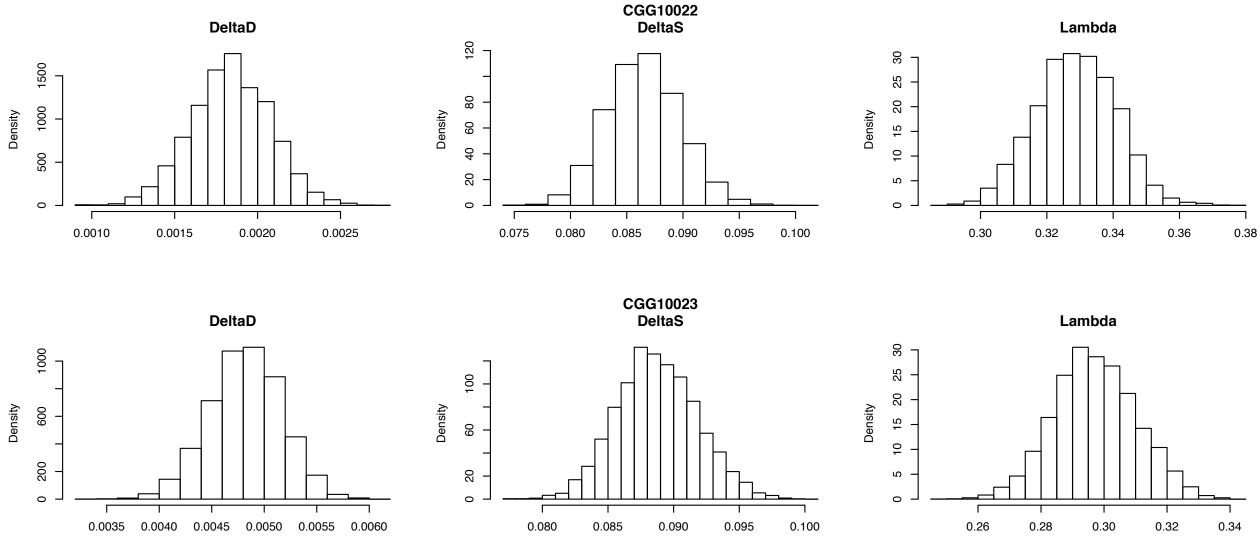
S1.5 *Post-mortem* damage of pre-domesticated samples

Sample-wide *post-mortem* DNA damage and fragmentation patterns for the pre-domesticated horses were estimated using mapDamage2.0(17), and plotted using a modified version of the mapDamage2.0 R-script (Supplementary Figure S4). Estimates of parameters were based on 100,000 randomly selected sequence alignments (using option "–n 100000") for each sample (Supplementary Figure S5). The plots reveal the expected pattern of *post-mortem* damage in the form of C>T substitutions at the 5' termini, and the complementary G>T substitutions at the 3' termini. The excess of purines observed near read-termini furthermore supports fragmentation driven by depurination (16).



Supplementary Figure S4. Post-mortem DNA damage and fragmentation patterns of pre-domesticated horses

DNA composition around read-termini (top plots), and DNA misincorporation errors relative to the 5' and 3' read termini for the pre-domestic samples (bottom plots); the two distributions for post mortem damage signatures (C>T and G>A) are shown in red and blue respectively, while other types of substitutions are shown in gray. Nucleotide frequencies are shown for 10 bases upstream and downstream of the 5' and 3' read termini.



Supplementary Figure S5. Model parameters estimated by mapDamage2.0

MCMC estimated posterior distribution for the model parameters. Lambda is the probability of termating an overhang and DeltaD/DeltaS is the probability of cytosine deamination in a double and single strand context, respectively.

S1.6 Average genomic-wide heterozygosity estimates

The average heterozygosity was estimated for each horse described in this study, by calculating the Watterson estimator ($\hat{\theta}_w$) (18) for 50 kb overlapping regions with a step-size of 10 kb of the genome as described in Supplementary Section S4.2. Priors for the autosomes were constructed using the site frequency spectrum observed for chromosome 22, while the site frequency spectrum observed for chrX was used for that chromosome. Windows in which less than 45 kb were covered were excluded from the analyses. The $\hat{\theta}_w$ for the pre-domesticated horses (CGG10022 and CGG10023) was calculated with or without transitions, in order to control for the effect of *post-mortem* DNA damage on the estimates. To examine the effect of *post-mortem* DNA damage, the average genomic heterozygosity of the samples were estimated with and without transitions (Supplementary Table S7).

	With transitions		**Without transitions**	
Sample	**Autosomes**	**chr X**	**Autosomes**	**chrX**
Arabian	4.116	3.558	3.168	2.663
CGG10022	4.582	4.131	3.684	3.356
CGG10023	5.272		4.080	
Norwegian Fjord	4.314	3.850	3.348	2.981
Icelandic (unnamed)	4.323		3.378	
Icelandic (P5782)	4.431		3.380	
Przewalski's horse	4.284		3.303	
Standardbred	4.161		3.150	

Thoroughbred (Twilight)	4.031	3.994	3.004	3.035
Domestic donkey (Willy)	4.041		3.100	

Supplementary Table S7. Average log($\widehat{\theta}_w$) for horse autosomes and chrX

The heterozygosity of chrX is only estimated for female individuals.

S1.7 Functional categorization of SNP variation

Genotyping was carried out as described in Supplementary Section S2.5 for each sample, excluding chromosome X and “Un”, due to the variable ploidy of chromosome X, and due to the highly repetitive nature of chrUn, leading to a high risk of SNPs called due to paralogous sequences; the filtered VCF records were annotated using v72 of the Ensembl “Variant Effect Predictor” (VEP) script (19). The results are tabulated in Supplementary Table S8; in cases were a variant was assigned multiple classifications (e.g. when adjacent to multiple genes / transcripts), the variant was counted once (and only once) for each type of classification. Genes are reported only once for each associated classification.

The following categories of variants are listed:

1. Outside genes:
   1.1. *Intergenic,* variant located *between* genes, but not in 1.2 or 1.3
   1.2. *Upstream*, variant located less than 5kb upstream of the 5’-termini of a gene
   1.3. *Downstream,* variant located less than 5kb downstream of the 3’-termini of a gene
2. Inside genes
   2.1. *Intron*, variant located in intronic region of gene
   2.2. *Non-coding Exon*, variant located in a non-coding exon (e.g. RNA)
   2.3. *5′ UTR*, variant located in the 5’ un-translated region of a gene
   2.4. *3′ UTR*, variant located in the 3’ un-translated region of a gene
   2.5. Splice Site, variant located in splice region, within 1-3 bp of the exon, or within 3-8 bp of the intron
   2.6. *Mature miRNA*, variant located within the sequence of a mature miRNA
   2.7. *Coding exon*, variants located within coding regions
      2.7.1. *Frameshift*, variant resulting in a frameshift of the amino acid sequence
      2.7.2. *Synonymous*, variant not resulting in a change in the amino acid sequence
      2.7.3. *Non-synonymous*, variant resulting in a change in the amino acid sequence, while preserving the length of the sequence
      2.7.4. *Stop gain*, variant resulting in the gain of a stop codon
      2.7.5. *Stop loss*, variant resulting in the loss of a stop codon

A Venn diagram showing the intersections between the filtered set of SNPs observed for the domestic horses, and the filtered sets of SNPs observed for each of the wild horses is plotted in Supplementary Figure S6. VCF files for the domestic horses were merged using the bcftools ‘merge’ command (included with SAMTools (15)), and intersections were determined using the BEDTools (20) ‘intersect’ command with options –r –f 1.0. Final plotting was performed using the R package ‘VennDiagram’ (http://cran.r-project.org/web/packages/VennDiagram/index.html).

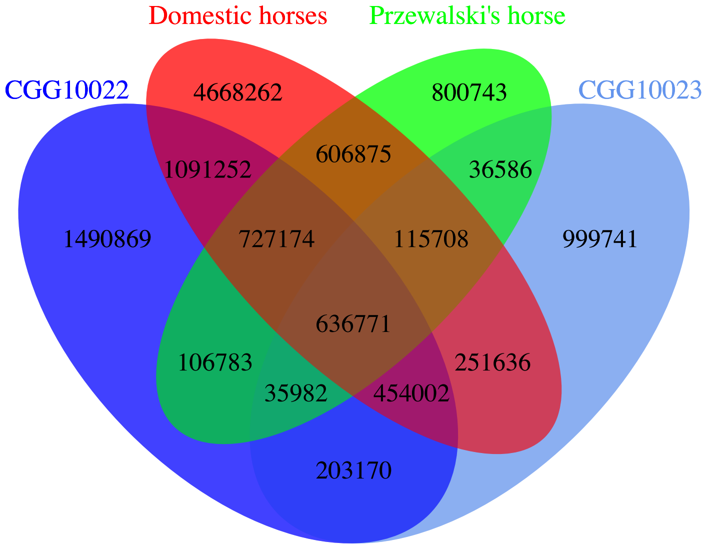
	Arabian		**CGG10022**		**CGG10023**		**Domestic donkey (Willy)**		**Icelandic (unnamed)**	
Variant class	**Genes**	**Variants**	**Genes**	**Variants**	**Genes**	**Variants**	**Genes**	**Variants**	**Genes**	**Variants**
All variants	24,552	2,812,453	25,218	5,035,345	25,068	2,849,804	25,220	23,507,788	24,911	2,375,124

Not genic		2,039,000		3,687,011		2,057,609		17,004,370		1,745,273
Intergenic		1,759,228		3,201,627		1,768,611		14,946,772		1,521,989
Downstream		152,505		262,642		155,160		1,079,688		118,223
Upstream		140,845		244,891		148,213		1,065,181		114,763
Genic	17,690	821,239	19,497	1,427,564	17,926	843,234	21,702	6,841,884	17,613	665,664
Intron	14,663	789,985	15,522	1,378,439	15,204	807,013	15,986	6,683,770	14,619	644,959
Non-coding Exon	1,213	2,297	1,809	3,969	954	1,895	3,264	13,575	1,314	2,302
5' UTR	954	1,351	1,472	2,267	940	1,292	2,390	4,666	443	594
3' UTR	1,352	1,876	2,126	3,281	1,526	2,173	5,241	13,796	1,039	1,356
Splice Site	2,566	3,402	3,969	5,733	2,967	3,980	9,164	19,683	1,850	2,239
Mature miRNA	19	19	23	23	9	9	52	62	10	10
Coding Exon	8,022	16,630	10,524	25,748	8,784	18,789	16,261	87,159	6,197	10,538
Frameshift	676	742	1,073	1,263	526	563	1,144	1,303	333	345
Synonymous	7,505	15,077	9,937	23,242	8,357	17,367	16,089	84,619	5,854	9,801
Non-synonymous	6,045	11,753	8,148	18,045	7,539	14,745	12,770	51,663	4,657	7,696
Stop gain	70	71	150	156	337	348	392	414	72	74
Stop loss	8	8	7	8	11	11	15	15	6	6

	Icelandic (P5782)		**Norwegian Fjord**		**Przewalski's Horse**		**Standardbred**		**Thoroughbred (Twilight)**	
Variant class	**Genes**	**Variants**	**Genes**	**Variants**	**Genes**	**Variants**	**Genes**	**Variants**	**Genes**	**Variants**
All variants	25,123	5,907,958	24,821	2,509,192	25,119	3,215,366	24,751	3,355,119	21,396	2,455,855
Not genic		4,371,085		1,813,593		2,347,764		2,457,020		1,805,385
Intergenic		3,813,775		1,572,944		2,040,091		2,142,366		1,569,123
Downstream		298,032		128,505		163,902		167,293		126,152
Upstream		282,859		123,309		157,818		161,096		120,125
Genic	20,140	1,624,524	17,546	736,637	18,484	919,227	18,132	949,769	15,132	687,787
Intron	15,425	1,575,555	14,814	709,390	15,155	887,730	14,803	919,019	12,481	666,265
Non-coding Exon	2,490	6,717	1,032	1,780	1,447	2,785	1,461	2,799	1,226	2,704
5' UTR	1,148	1,783	703	936	732	980	725	959	570	838
3' UTR	2,232	3,481	1,216	1,623	1,535	2,098	1,477	2,076	1,076	1,491
Splice Site	3,909	5,670	2,273	2,897	2,620	3,363	2,611	3,411	2,037	2,625
Mature miRNA	28	29	15	16	25	25	19	19	13	13
Coding Exon	10,041	23,560	7,587	14,876	8,334	16,345	8,010	15,867	5,399	10,458
Frameshift	1,040	1,243	396	414	469	491	515	556	678	753
Synonymous	9,441	21,193	7,248	13,933	7,934	15,260	7,557	14,664	4,647	8,931
Non-synonymous	7,838	18,032	5,702	10,236	6,402	11,907	6,038	11,554	4,066	8,016
Stop gain	141	148	85	87	103	105	76	78	74	74
Stop loss	8	8	3	3	3	3	10	10	6	7

Supplementary Table S8. Classification of functional variants across samples

*Classification carried out using the Ensembl VEP script; see Supplementary Section S1.7 for a description of the various classes. The column "**Genes**" gives the number of unique Ensembl annotated genes intersecting with one or more classes of variants; the column "**Variants**" gives the number of unique variants which are assigned a given class.*



Supplementary Figure S6. Intersection of SNPs called for domestic and wild horses

S1.8 Microbial profiling of ancient horse DNA extracts

S1.8.1 Background

We performed metagenomic analyses on a subset of shotgun sequenced DNA libraries for the two samples CGG10022 and CGG10023, as well as for ancient horse shotgun datasets previously described in Orlando et al., 2013 (1) (Supplementary Table S9). Comparative analyses also included the microbial profiles of six ancient horses characterized in Der Sarkissian et al., 2014 (21).

All datasets were profiled using MetaPhlAn (Metagenomic Phylogenetic Analysis version 1.7.7, February 2013) (22) as implemented in the PALEOMIX pipeline (10, 21). Shotgun sequencing reads resulting from each DNA library were mapped to the markers of the MetaPhlAn database using the Bowtie2 v2.1.0 aligner (23) with default parameters and a sensitive global alignment strategy (default --end-to-end mode). For those published datasets sequenced on the GA Illumina analyzer the Bowtie2 option --solexa-quals was used. Before running MetaPhlAn, PCR duplicates were removed from each DNA library. For the paired-end sequenced samples CGG10022 and CGG10023, we used a modified version of the filtering program used for the Neanderthal draft genome (*FilterUniqueBAM.py*) (24). Uncollapsed PE reads were not used. For all remaining single-end datasets, PCR duplicates were first identified using the MarkDuplicates function of Picard Tools version 1.82 (http://picard.sourceforge.net/) and then removed with the SAMtools “view” command (15) as described in Der Sarkissian et al. 2014 (21).

In order to identify potential biases, we compared microbial profiles characterized from three different types of DNA library procedures: A-tailing DNA libraries (AT) (1, 25), modified TruSeq DNA libraries (TS) (9) and blunt ended DNA libraries (KM) (8) (Supplementary Table S9).

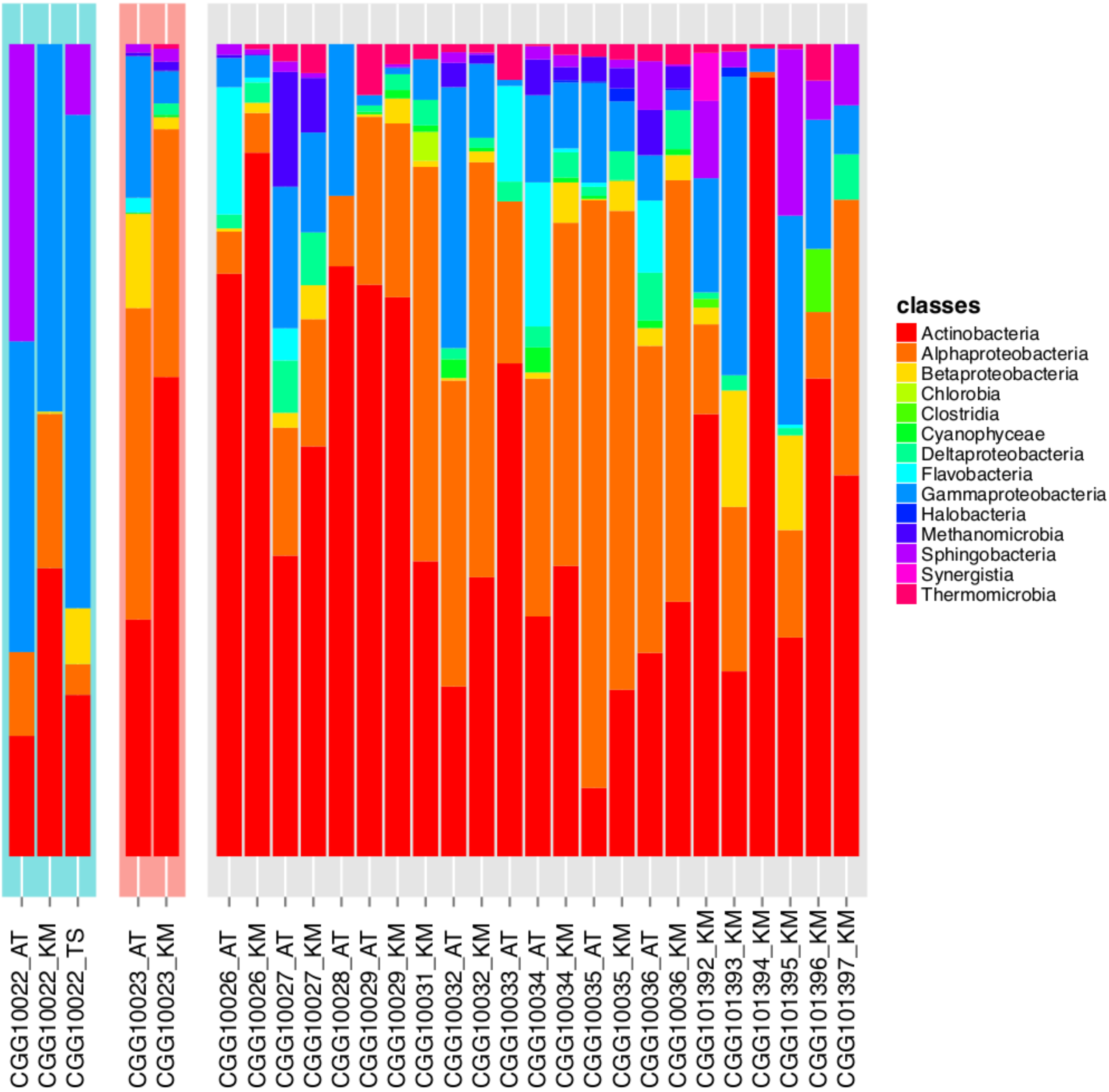
We compared taxon abundances among DNA libraries using a suite of analyses in the statistical environment R version 3.0.1 (26) as described previously (10, 21). In order to reduce the rate of false positives, we excluded low-abundance taxa (less than 1%). Shannon diversity indices were computed from relative abundance data using the function *diversity* of the vegan package in R (http://cran.r-project.org/package=vegan). A two-sample t-test was carried out to compare the diversity of A-tailing (AT) libraries versus blunt ended (KM) DNA libraries from the distribution of Shannon indices. TruSeq (TS) libraries were excluded from this test as only one TS library was available. Principal Coordinate Analysis (PCoA, R function *pcoa*) was performed using Bray-Curtis distances among profiles. We carried out hierarchical clustering using the R package *pvclust*, based on the Manhattan metric and average linkage clustering method (27), with 10,000 bootstrap iterations to estimate cluster support (Approximately Unbiased *p-values* and Bootstrap Probabilities).

The CGG10022 and CGG10023 ancient horse sample microbial profiles were also compared with profiles of soil (28) and human samples (29) by PCoA based on Bray-Curtis distances. The soil and human comparative MetaPhlAn profiles are publicly available (21, 29).

S1.8.2 Results

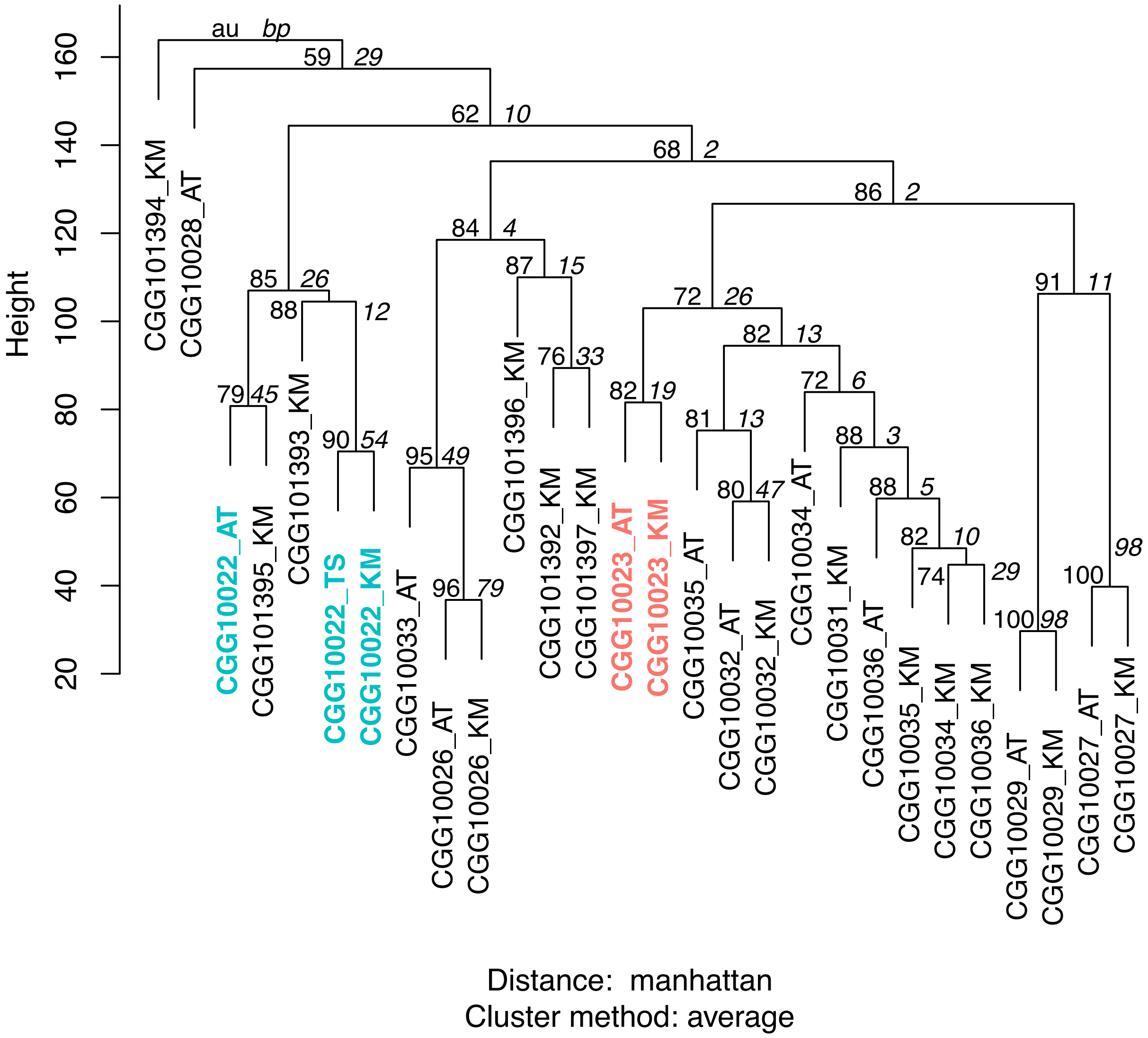
Actinobacteria (8-96%) is the dominant class across the ancient horse microbial profiles (Supplementary Figure S7). It was also detected at high frequencies in CGG10022 (15-35%) and CGG10023 (29-59%). The *Alphaproteobacteria* (1-72%) and the *Gammaproteobacteria* (1-36%) classes were broadly found in the ancient horses extracts, as well as in CGG10022 (3-19% and 45-61%) and CGG10023 (30-38% and 4-17%). The *Betaproteobacteria* (0-7%) and *Sphingobacteria* (9-36%) classes have also been detected in CGG10022, and CGG10023's microbial profiles showed occurrence of members of the *Flavobacteria* (0-2%), *Sphingobacteria* (1-2%), *Betaproteobacteria* (1-12%) and *Deltaproteobacteria* (0-1%) classes.

On the basis of genus-level relative abundances, ancient horse microbial profiles segregated into two main clusters (Supplementary Figure S8), with each of CGG10022 and CGG10023 belonging to one cluster, and the samples CGG101394 and CGG10028 grouping out. This structure of the microbial diversity is in line with previous results described in Der Sarkissian et al. 2014 (21). Hierarchical clustering and PCoA (Supplementary Figure S8 and Supplementary Figure S9) show that the variability in the profiles obtained between DNA libraries of a given sample is smaller than that between any two samples from distinct clusters, thus suggesting a negligible impact of the DNA library building method on the distribution of the microbial diversity among samples. In line with this result, Shannon diversity indices calculated from AT-libraries (mean: 1.9, standard deviation: 0.53) and from KM-libraries (mean: 2.2, standard deviation: 0.36) showed no statistically significant differences on the basis of the t-test (p-value: (p-value: 0.14). In addition, the presence of different groups of ancient horse microbial profiles does not support an extensive contamination of all the ancient horse extracts, which would have masked any difference among samples. In addition PCoA support low levels of human-derived contamination (Supplementary Figure S10), as the CGG10022 and CGG10023 microbial profiles are found to be distinct from human-associated microbiomes, but similar to soil samples. Overall, the dominant microorganisms detected in the ancient horse samples (*Mycobacterium*: 2-63%, *Pseudomonas*: 0-60%, *Rhodopseudomonas*: 0-60%, and *Arthrobacter*: 0-43%) indeed belong to genera commonly found in soils (28). This is in accordance with previous results suggesting that the diversity observed in microbe-derived DNA reads sequenced in the ancient horse remains derives from *post-mortem* colonisation by micro-organisms of the depositional environment (21).



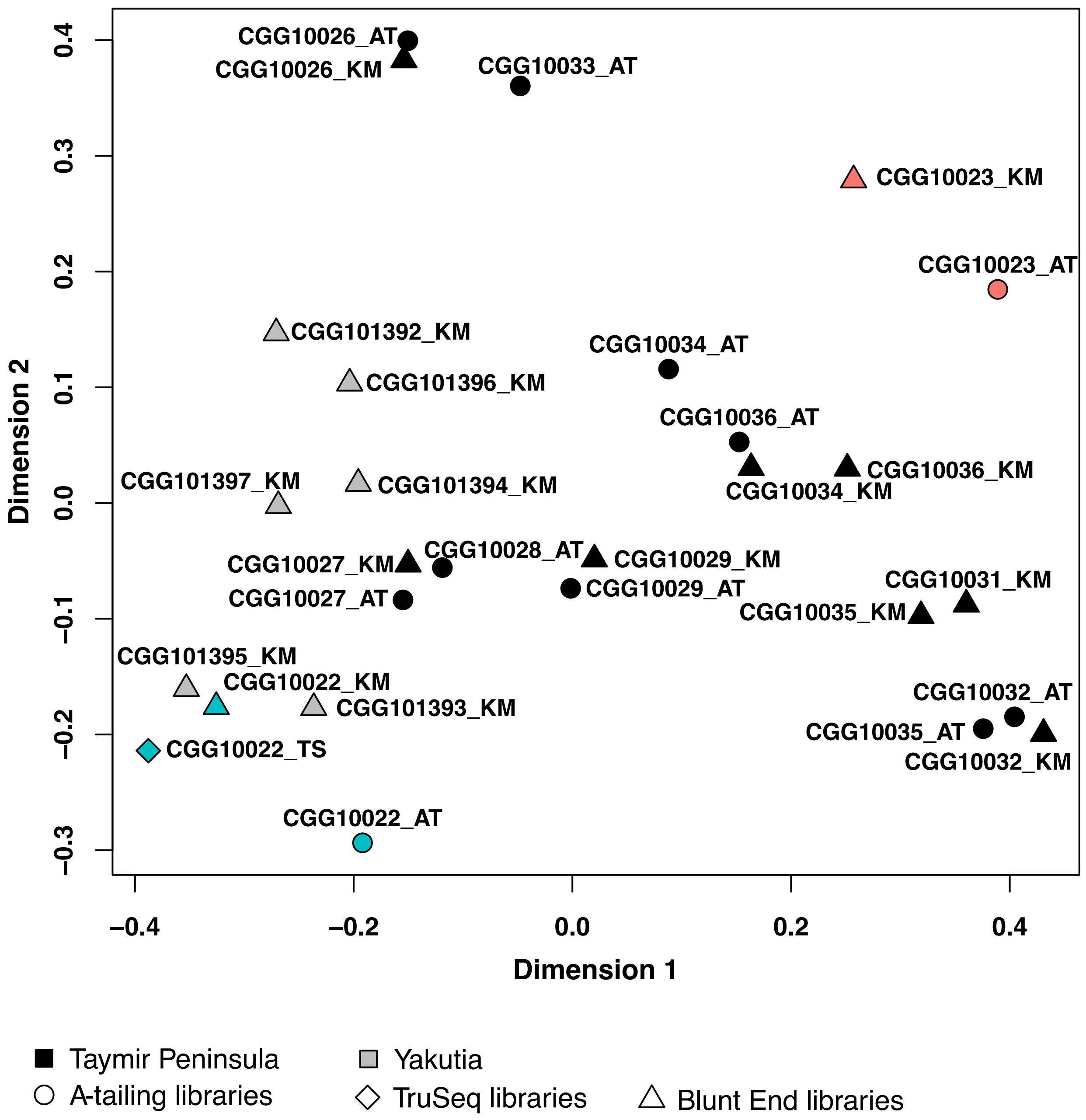
Supplementary Figure S7. Relative abundances of microbial classes in ancient horse DNA extracts

AT=A-tailing DNA library procedure described in Orlando et al 2013 (1). *TS=Modified TruSeq DNA library building procedure described in Pedersen et al. 2014* (9). *KM=DNA library building procedure from Meyer and Kircher 2010* (8).



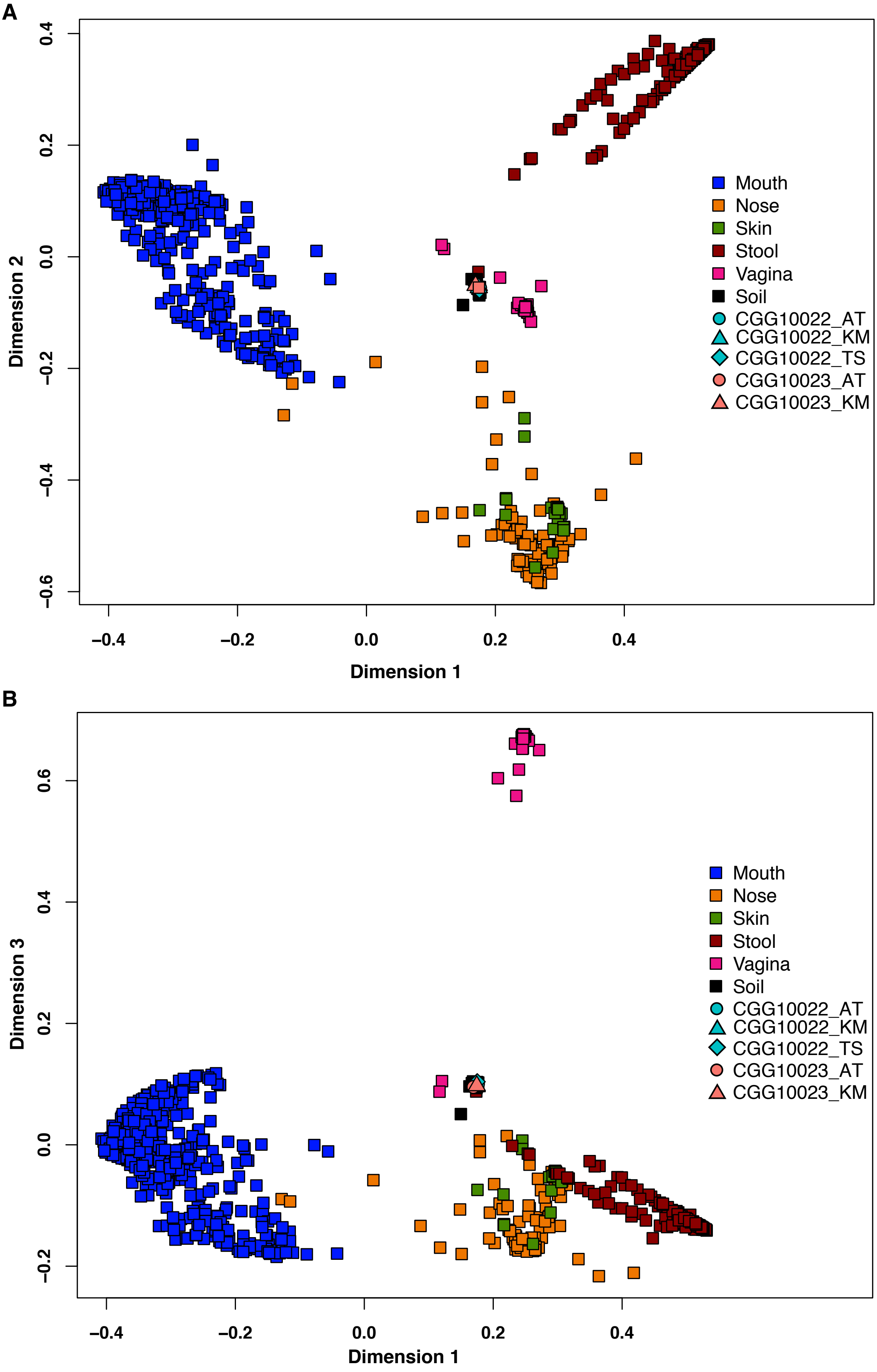
Supplementary Figure S8. Hierarchical clustering of Manhattan distances between microbial DNA profiles at the genus level in ancient horse extracts (10,000 bootstraps).

AT=A-tailing DNA library procedure described in Orlando et al 2013 (1). *TS= modified TruSeq DNA library building procedure described in Pedersen et al. 2014* (9). *KM= DNA library building procedure from Meyer and Kircher 2010* (8)*; "au", approximately unbiased p-value; "bp", bootstrap probability.*



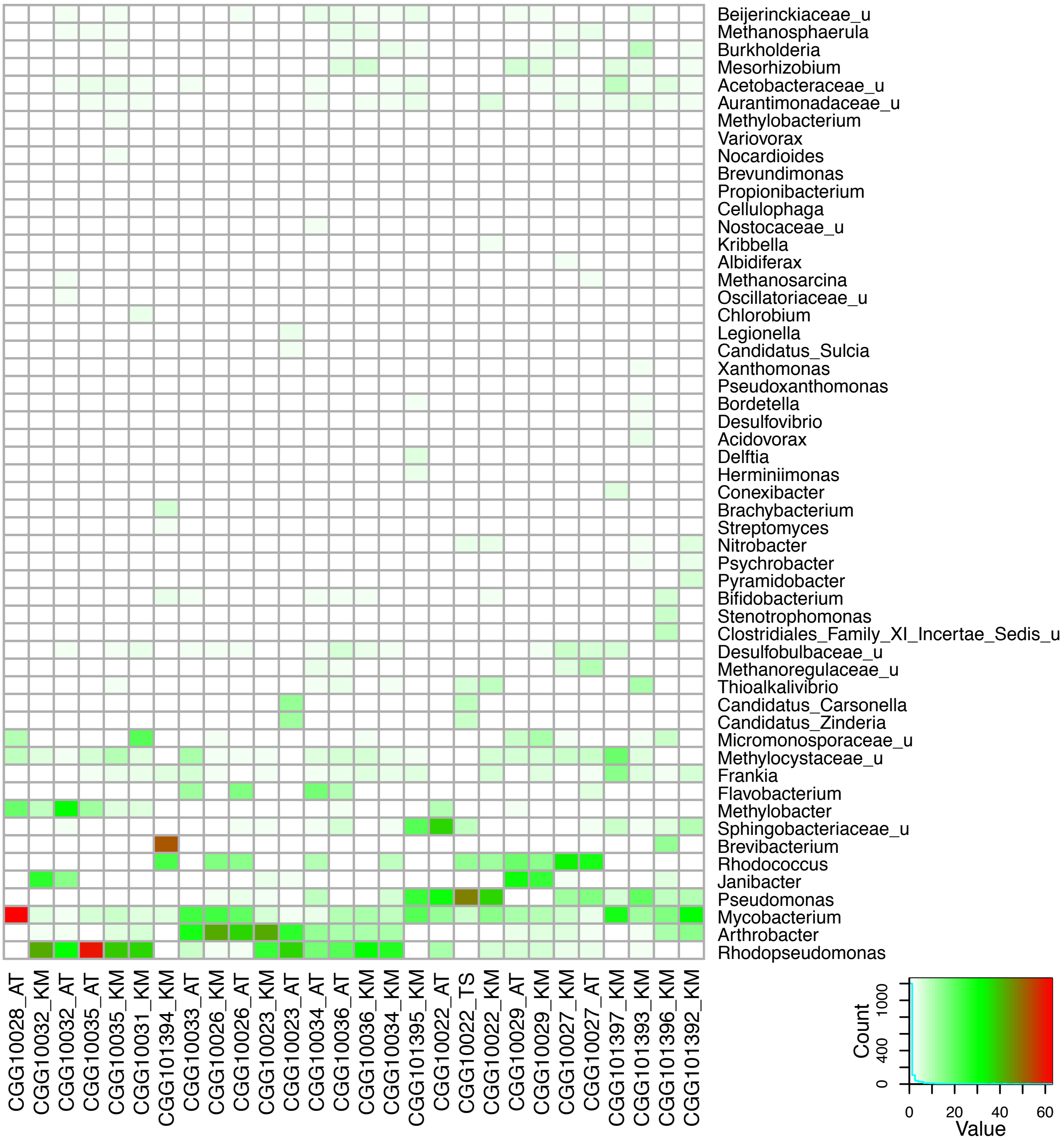
Supplementary Figure S9. Principal Coordinate Analysis of Bray-Curtis distances between microbial DNA profiles at the genus level in ancient horse extracts.

AT=A-tailing DNA library procedure described in Orlando et al 2013 (1)*. TS= modified TruSeq DNA library building procedure described in Pedersen et al. 2014* (9)*. KM= DNA library building procedure from Meyer and Kircher 2010* (8)*.*


A
Dimension 2
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
Dimension 1
Mouth
Nose
Skin
Stool
Vagina
Soil
CGG10022_AT
CGG10022_KM
CGG10022_TS
CGG10023_AT
CGG10023_KM
B
Dimension 3
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
Dimension 1
Mouth
Nose
Skin
Stool
Vagina
Soil
CGG10022_AT
CGG10022_KM
CGG10022_TS
CGG10023_AT
CGG10023_KM

Supplementary Figure S10. Principal Coordinate Analysis of Bray-Curtis distances between microbial DNA profiles at the genus level in various human associated microbiomes, soils, and the ancient horse extracts CGG10022 and CGG10023

AT=A-tailing DNA library procedure described in Orlando et al 2013 (1)*. TS= modified TruSeq DNA library building procedure described in Pedersen et al. 2014* (9)*. KM= DNA library building procedure from Meyer and Kircher 2010* (8)*. A. Dimensions 1 and 2. B. Dimensions 1 and 3.*



Supplementary Figure S11. Heatmap showing relative abundances of microbial genera in ancient horse DNA extracts

"_u", unclassified, AT=A-tailing DNA library procedure described in Orlando et al 2013 (1)*. TS= modified TruSeq DNA library building procedure described in Pedersen et al. 2014* (9)*. KM= DNA library building procedure from Meyer and Kircher 2010* (8)*.*

Sample	Library building method	Identifier / Platform	#Trimmed reads	#Genera	Shannon diversity
CGG10022	KM	ACTTGA[a]	49,964,648	11	2.0
	AT	CGTAGT[a]	26,525,706	6	1.5
		CGTAGT[a]	17,699,174		
	TS	CTTGTA[a]	37,496,212	6	1.7
		TGACCA[a]	29,601,721		
CGG10023	AT	Lib1[a]	1,939,920,220	5	1.7
		Lib12[a]	387,303,645		
	KM	LOb[a]	223,066,616	9	1.9
		LOc[a]	204,005,371		
		LOd[a]	387,303,645		
CGG10026	AT	CGG10026_s[a]	27,529,781	10	1.8
	KM	CGG10026_TAGCTT[a]	19,951,256	9	1.8
CGG10027	AT	CGG10027_s[a]	17,945,473	18	2.5
		CCG10027_[a]	87,405,232		
		CGG10027_GA[b]	11,966,997		
	KM	CGG10027_GGCTAC[a]	30,132,473	15	2.4
CGG10028	AT	CGG10028_s[a]	9,823,131	4	1.0
CGG10029	AT	CGG10029_s[a]	27,277,717	13	2.1
	KM	CGG10029_CTTGTA[a]	21,064,242	11	2.5
CGG10031	KM	CGG10031_CTATCA[a]	50,436,657	13	2.1
CGG10032	AT	CGG10032_s[a]	14,597,524	14	1.9
		CGG10032_GA[b]	15,781,540		
	KM	CGG10032_CGCTAT[a]	50,306,194	8	1.8
CGG10033	AT	CGG10033_s[a]	5,113,925	9	1.9
		CGG10033_GA[b]	11,439,325		
CGG10034	AT	CGG10034_s[a]	29,667,651	17	2.7
	KM	CGG10034_CGTATA[a]	76,350,016	20	2.7
CGG10035	AT	CGG10035_AT[a]	23,567,250	10	1.5
	KM	CGG10035_TGATCG[a]	44,525,214	17	2.4
CGG10036	AT	CGG10036_s[a]	29,740,941	14	2.7
	KM	CGG10036_GTGTAT[a]	73,775,803	18	2.5
CGG101392	KM	OSRE[a]	16,400,000	20	2.26
CGG101393	KM	BaBRE[a]	9,800,000	31	2.84
CGG101394	KM	YaRE[a]	7,200,000	16	1.54
CGG101395	KM	TyRE[a]	12,200,000	19	2.2
CGG101396	KM	TaRE[a]	29,400,000	18	2.56
CGG101397	KM	TuRE[a]	15,100,000	10	2.07

Supplementary Table S9. Samples used for metagenomic analyses

For Platform, a) indicates Illumina HiSeq 2000, and b) indicates Illumina Genome Analyzer IIx.

S2 Phylogenetic and demographic analyses

S2.1 Phylogenetic inference from mitochondrial sequences

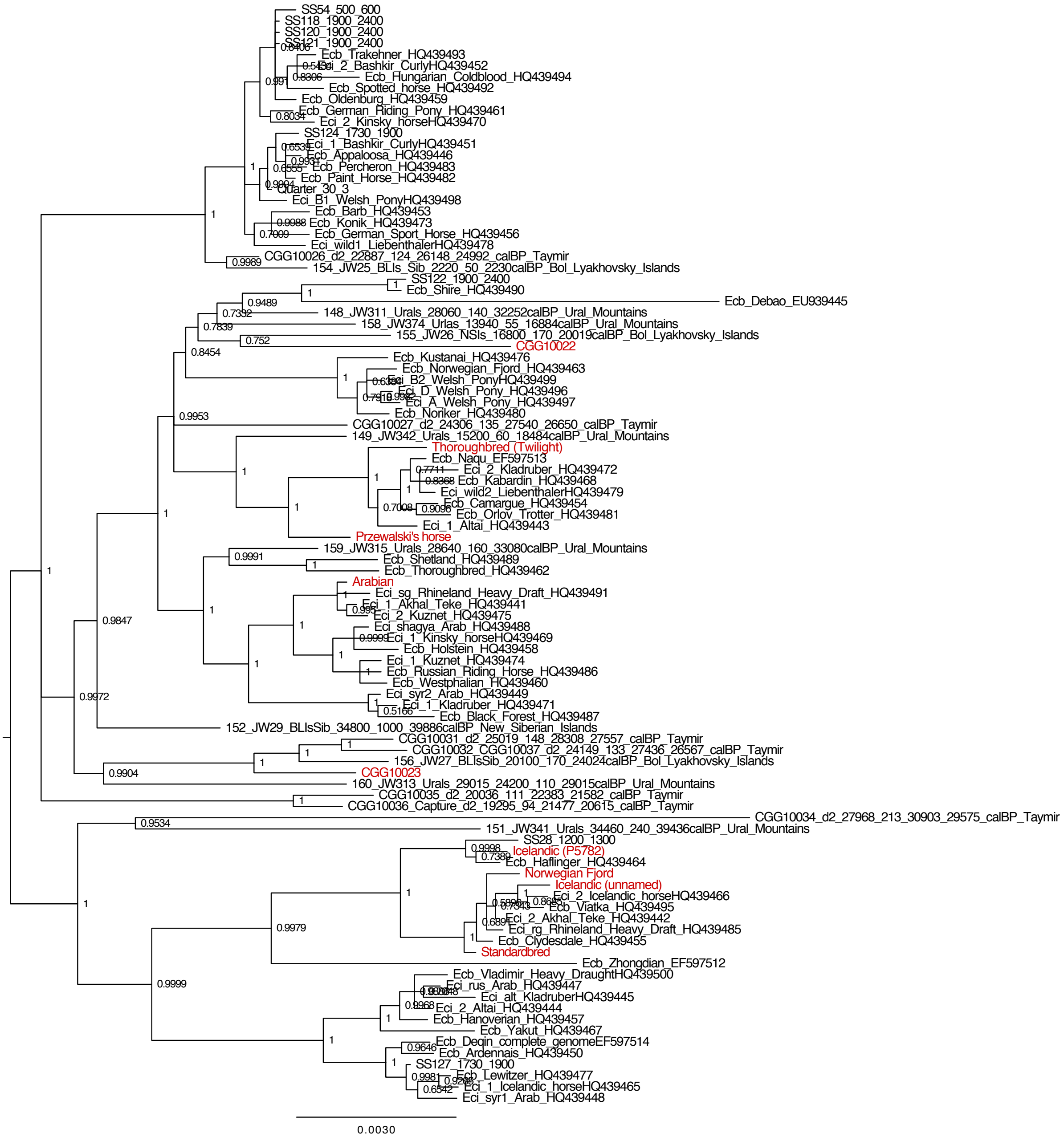
Following mapping against the horse reference mitochondrial genome (Accession Nb. NC_001640; see section S1.3), we prepared the mitochondrial genome sequence of samples CGG10022 and CGG10023 based on a majority rule at each covered position. Only positions covered by a minimum number of three independent unique reads showing base qualities greater or equal to 30 were called. The mitochondrial consensus sequences were then aligned to a total of 25 ancient horse mitochondrial sequences previously reported(1, 30) as well as to those from the eight modern horse specimens investigated here (see section S1.1) and to an additional number of 64 mitochondrial sequences collected from modern domestic breeds (see the labels of external branches on Supplementary Figure S12 for a list of Accession numbers). We partitioned the alignment into six main regions, namely: ribosomal RNA, tRNA, Control Region, and the first, second and third codon positions for CDS. We selected the best mutational model for each of those partitions using ModelGenerator v851 (31) and eight rate categories. The partitions and their corresponding mutational models were used for Bayesian phylogenetic inference with MrBayes (32), running two analyses in parallel, each with four MCMC chains. The final tree topology was recovered following a total number of 300,000,000 generations, sampling 1 every 1,000 generations and disregarding the first 25% as burn-in. The resulting tree, as drawn with MEGA v5.0 (33), is shown in Supplementary Figure S12.

The six partitions were also used as input for Bayesian skyride analyses in BEAST v1.8.0 (33), where we also used the radiocarbon dates (or stratigraphic context information) of the ancient specimens for tip calibration (when a range of dates was available, we chose the center of the proposed range as a date). Two types of analyses were run: In the first series of analysis, a strict clock was assumed whereas the second series of analyses assumed a log-Uncorrelated relaxed clock. For each type of analysis, we also applied three possible demographic models: we assumed that the population size remained constant, a Bayesian Skyride model or a Bayesian Skyline model. The analyses were run for 200 million generations, sampling 1 every 50,000 generations, and the first 10% were disregard as burn-in. As Bayes Factor calculation supported the Bayesian Skyline model assuming log-Uncorrelated relaxed clock (7.764 ≤ log BF ≤ 44.985), we reconstructed the past demographic profile of horses using the Bayesian Skyline analysis based on this clock assumption (Supplementary Figure S14). We also used TreeStatv1.8.0 in order to recover the posterior distribution for the mitochondrial TMRCA of horses (Supplementary Table S10) and the final tree topology was recovered using FigTree, where node support is provided by the node posterior probability (Supplementary Figure S13).

	5%	Median	95%
Mutation Rate (per site per myr)	3.15E-08	**4.68E-08**	6.36E-05
TMRCA (yr BP)	104638	**145617**	214694

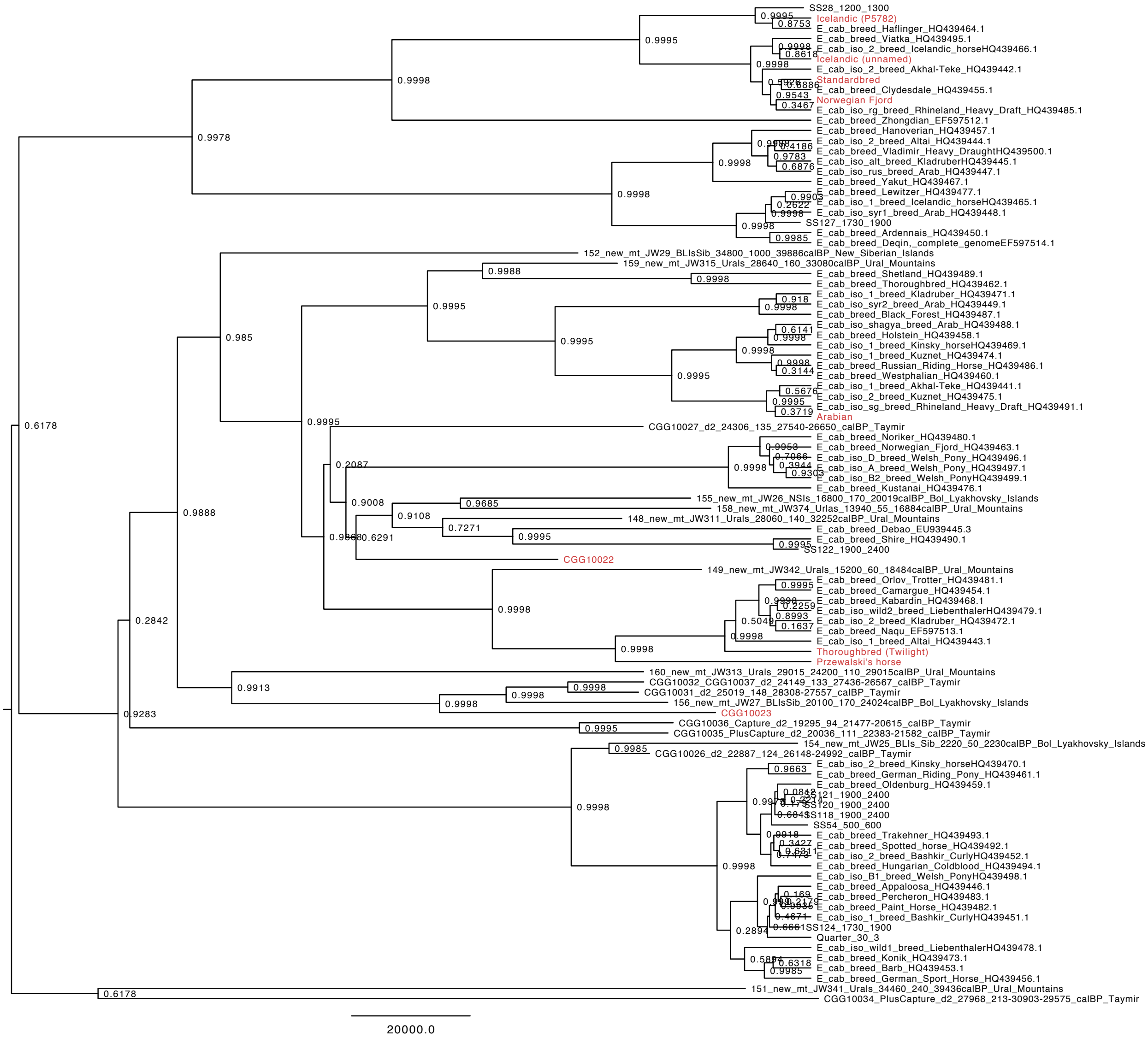
Supplementary Table S10. Posterior estimates of mt mutation rate and TMRCA of horses

We provide median, 5%- and 95%-quantile values of the posterior distribution sampled in BEAST using 10% as burn-in.


SS54_500_600
SS118_1900_2400
SS120_1900_2400
SS121_1900_2400
Ecb_Trakehner_HQ439493
Eci_2_Bashkir_CurlyHQ439452
Ecb_Hungarian_Coldblood_HQ439494
Ecb_Spotted_horse_HQ439492
Ecb_Oldenburg_HQ439459
Ecb_German_Riding_Pony_HQ439461
Eci_2_Kinsky_horseHQ439470
SS124_1730_1900
Eci_1_Bashkir_CurlyHQ439451
Ecb_Appaloosa_HQ439446
Ecb_Percheron_HQ439483
Ecb_Paint_Horse_HQ439482
Quarter_30_3
Eci_B1_Welsh_PonyHQ439498
Ecb_Barb_HQ439453
Ecb_Konik_HQ439473
Ecb_German_Sport_Horse_HQ439456
Eci_wild1_LiebenthalerHQ439478
CGG10026_d2_22887_124_26148_24992_calBP_Taymir
154_JW25_BLIs_Sib_2220_50_2230calBP_Bol_Lyakhovsky_Islands
SS122_1900_2400
Ecb_Shire_HQ439490
Ecb_Debao_EU939445
148_JW311_Urals_28060_140_32252calBP_Ural_Mountains
158_JW374_Urlas_13940_55_16884calBP_Ural_Mountains
155_JW26_NSIs_16800_170_20019calBP_Bol_Lyakhovsky_Islands
CGG10022
Ecb_Kustanai_HQ439476
Ecb_Norwegian_Fjord_HQ439463
Eci_B2_Welsh_PonyHQ439499
Eci_D_Welsh_Pony_HQ439496
Eci_A_Welsh_Pony_HQ439497
Ecb_Noriker_HQ439480
CGG10027_d2_24306_135_27540_26650_calBP_Taymir
149_JW342_Urals_15200_60_18484calBP_Ural_Mountains
Thoroughbred (Twilight)
Ecb_Naqu_EF597513
Eci_2_Kladruber_HQ439472
Ecb_Kabardin_HQ439468
Eci_wild2_LiebenthalerHQ439479
Ecb_Camargue_HQ439454
Ecb_Orlov_Trotter_HQ439481
Eci_1_Altai_HQ439443
Przewalski's horse
159_JW315_Urals_28640_160_33080calBP_Ural_Mountains
Ecb_Shetland_HQ439489
Ecb_Thoroughbred_HQ439462
Arabian
Eci_sg_Rhineland_Heavy_Draft_HQ439491
Eci_1_Akhal_Teke_HQ439441
Eci_2_Kuznet_HQ439475
Eci_shagya_Arab_HQ439488
Eci_1_Kinsky_horseHQ439469
Ecb_Holstein_HQ439458
Eci_1_Kuznet_HQ439474
Ecb_Russian_Riding_Horse_HQ439486
Ecb_Westphalian_HQ439460
Eci_syr2_Arab_HQ439449
Eci_1_Kladruber_HQ439471
Ecb_Black_Forest_HQ439487
152_JW29_BLIsSib_34800_1000_39886calBP_New_Siberian_Islands
CGG10031_d2_25019_148_28308_27557_calBP_Taymir
CGG10032_CGG10037_d2_24149_133_27436_26567_calBP_Taymir
156_JW27_BLIsSib_20100_170_24024calBP_Bol_Lyakhovsky_Islands
CGG10023
160_JW313_Urals_29015_24200_110_29015calBP_Ural_Mountains
CGG10035_d2_20036_111_22383_21582_calBP_Taymir
CGG10036_Capture_d2_19295_94_21477_20615_calBP_Taymir
CGG10034_d2_27968_213_30903_29575_calBP_Taymir
151_JW341_Urals_34460_240_39436calBP_Ural_Mountains
SS28_1200_1300
Icelandic (P5782)
Ecb_Haflinger_HQ439464
Norwegian Fjord
Icelandic (unnamed)
Eci_2_Icelandic_horseHQ439466
Ecb_Viatka_HQ439495
Eci_2_Akhal_Teke_HQ439442
Eci_rg_Rhineland_Heavy_Draft_HQ439485
Ecb_Clydesdale_HQ439455
Standardbred
Ecb_Zhongdian_EF597512
Ecb_Vladimir_Heavy_DraughtHQ439500
Eci_rus_Arab_HQ439447
Eci_alt_KladruberHQ439445
Eci_2_Altai_HQ439444
Ecb_Hanoverian_HQ439457
Ecb_Yakut_HQ439467
Ecb_Deqin_complete_genomeEF597514
Ecb_Ardennais_HQ439450
SS127_1730_1900
Ecb_Lewitzer_HQ439477
Eci_1_Icelandic_horseHQ439465
Eci_syr1_Arab_HQ439448
0.0030

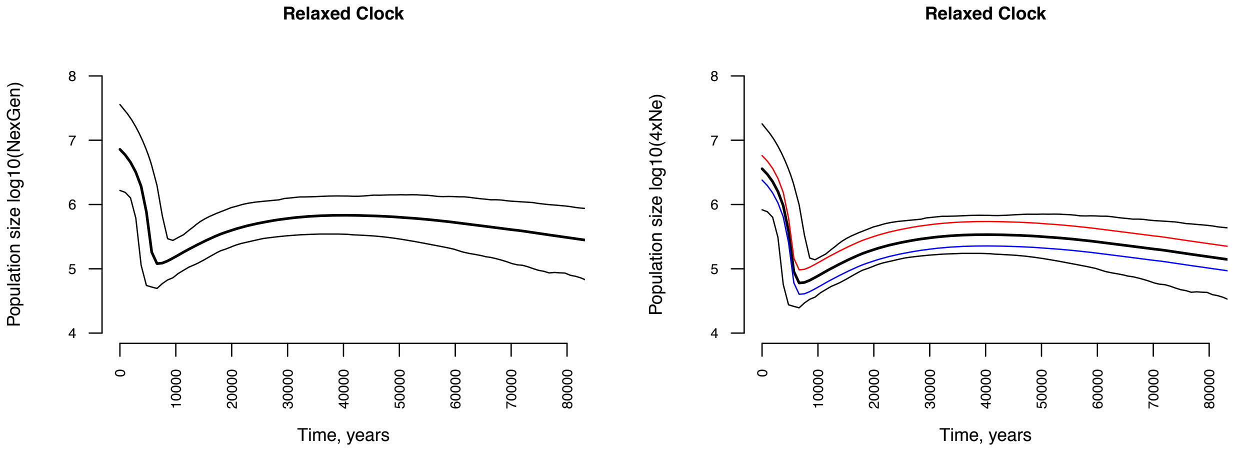
Supplementary Figure S12. Bayesian phylogenetic reconstruction using MrBayes

The labels of each external branch refer to the accession numbers of previously released modern mitochondrial genome sequences. Labels without an accession number refer to samples described in this study (those labeled in red; see Supplementary Section S1.1) or in the study by Orlando et al. 2013 (1). *The tree is rooted on the midpoint.*


SS28_1200_1300
Icelandic (P5782)
E_cab_breed_Haflinger_HQ439464.1
E_cab_breed_Viatka_HQ439495.1
E_cab_iso_2_breed_Icelandic_horseHQ439466.1
Icelandic (unnamed)
E_cab_iso_2_breed_Akhal-Teke_HQ439442.1
Standardbred
E_cab_breed_Clydesdale_HQ439455.1
Norwegian Fjord
E_cab_iso_rg_breed_Rhineland_Heavy_Draft_HQ439485.1
E_cab_breed_Zhongdian_EF597512.1
E_cab_breed_Hanoverian_HQ439457.1
E_cab_iso_2_breed_Altai_HQ439444.1
E_cab_breed_Vladimir_Heavy_DraughtHQ439500.1
E_cab_iso_alt_breed_KladruberHQ439445.1
E_cab_iso_rus_breed_Arab_HQ439447.1
E_cab_breed_Yakut_HQ439467.1
E_cab_breed_Lewitzer_HQ439477.1
E_cab_iso_1_breed_Icelandic_horseHQ439465.1
E_cab_iso_syr1_breed_Arab_HQ439448.1
SS127_1730_1900
E_cab_breed_Ardennais_HQ439450.1
E_cab_breed_Deqin,_complete_genomeEF597514.1
152_new_mt_JW29_BLIsSib_34800_1000_39886calBP_New_Siberian_Islands
159_new_mt_JW315_Urals_28640_160_33080calBP_Ural_Mountains
E_cab_breed_Shetland_HQ439489.1
E_cab_breed_Thoroughbred_HQ439462.1
E_cab_iso_1_breed_Kladruber_HQ439471.1
E_cab_iso_syr2_breed_Arab_HQ439449.1
E_cab_breed_Black_Forest_HQ439487.1
E_cab_iso_shagya_breed_Arab_HQ439488.1
E_cab_breed_Holstein_HQ439458.1
E_cab_iso_1_breed_Kinsky_horseHQ439469.1
E_cab_iso_1_breed_Kuznet_HQ439474.1
E_cab_breed_Russian_Riding_Horse_HQ439486.1
E_cab_breed_Westphalian_HQ439460.1
E_cab_iso_1_breed_Akhal-Teke_HQ439441.1
E_cab_iso_2_breed_Kuznet_HQ439475.1
E_cab_iso_sg_breed_Rhineland_Heavy_Draft_HQ439491.1
Arabian
CGG10027_d2_24306_135_27540-26650_calBP_Taymir
E_cab_breed_Noriker_HQ439480.1
E_cab_breed_Norwegian_Fjord_HQ439463.1
E_cab_iso_D_breed_Welsh_Pony_HQ439496.1
E_cab_iso_A_breed_Welsh_Pony_HQ439497.1
E_cab_iso_B2_breed_Welsh_PonyHQ439499.1
E_cab_breed_Kustanai_HQ439476.1
155_new_mt_JW26_NSIs_16800_170_20019calBP_Bol_Lyakhovsky_Islands
158_new_mt_JW374_Urlas_13940_55_16884calBP_Ural_Mountains
148_new_mt_JW311_Urals_28060_140_32252calBP_Ural_Mountains
E_cab_breed_Debao_EU939445.3
E_cab_breed_Shire_HQ439490.1
SS122_1900_2400
CGG10022
149_new_mt_JW342_Urals_15200_60_18484calBP_Ural_Mountains
E_cab_breed_Orlov_Trotter_HQ439481.1
E_cab_breed_Camargue_HQ439454.1
E_cab_breed_Kabardin_HQ439468.1
E_cab_iso_wild2_breed_LiebenthalerHQ439479.1
E_cab_iso_2_breed_Kladruber_HQ439472.1
E_cab_breed_Naqu_EF597513.1
E_cab_iso_1_breed_Altai_HQ439443.1
Thoroughbred (Twilight)
Przewalski's horse
160_new_mt_JW313_Urals_29015_24200_110_29015calBP_Ural_Mountains
CGG10032_CGG10037_d2_24149_133_27436-26567_calBP_Taymir
CGG10031_d2_25019_148_28308-27557_calBP_Taymir
156_new_mt_JW27_BLIsSib_20100_170_24024calBP_Bol_Lyakhovsky_Islands
CGG10023
CGG10036_Capture_d2_19295_94_21477-20615_calBP_Taymir
CGG10035_PlusCapture_d2_20036_111_22383-21582_calBP_Taymir
154_new_mt_JW25_BLIs_Sib_2220_50_2230calBP_Bol_Lyakhovsky_Islands
CGG10026_d2_22887_124_26148-24992_calBP_Taymir
E_cab_iso_2_breed_Kinsky_horseHQ439470.1
E_cab_breed_German_Riding_Pony_HQ439461.1
E_cab_breed_Oldenburg_HQ439459.1
SS121_1900_2400
SS120_1900_2400
SS118_1900_2400
SS54_500_600
E_cab_breed_Trakehner_HQ439493.1
E_cab_breed_Spotted_horse_HQ439492.1
E_cab_iso_2_breed_Bashkir_CurlyHQ439452.1
E_cab_breed_Hungarian_Coldblood_HQ439494.1
E_cab_iso_B1_breed_Welsh_PonyHQ439498.1
E_cab_breed_Appaloosa_HQ439446.1
E_cab_breed_Percheron_HQ439483.1
E_cab_breed_Paint_Horse_HQ439482.1
E_cab_iso_1_breed_Bashkir_CurlyHQ439451.1
SS124_1730_1900
Quarter_30_3
E_cab_iso_wild1_breed_LiebenthalerHQ439478.1
E_cab_breed_Konik_HQ439473.1
E_cab_breed_Barb_HQ439453.1
E_cab_breed_German_Sport_Horse_HQ439456.1
151_new_mt_JW341_Urals_34460_240_39436calBP_Ural_Mountains
CGG10034_PlusCapture_d2_27968_213-30903-29575_calBP_Taymir
20000.0

Supplementary Figure S13. BEAST phylogenetic reconstruction

The labels of each external branch refer to the accession numbers of previously released modern mitochondrial genome sequences. Labels without an accession number refer to samples described in this study (those labeled in red; see Supplementary Section S1.1) or in the study by Orlando et al. 2013 (1)*.*



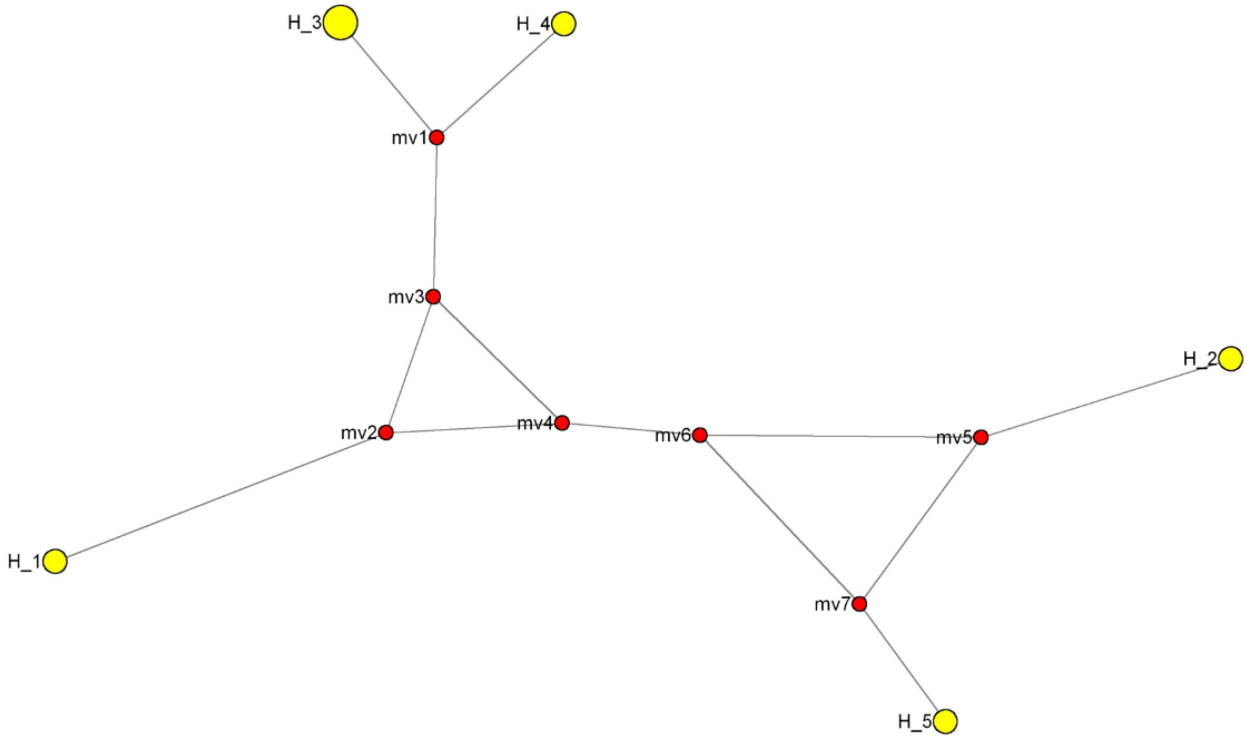
Supplementary Figure S14. Bayesian skyline demographic profile

***Left:** The y-axis provides a log10-transformed measure of the product of the effective size and the generation time. Median values are indicated as thick lines, in contrast to the 5%-95% confidence range is indicated with plain thin lines. **Right:** Same as in the left panel, except that the y-axis now refers to the log10-transformed equivalent of the population size (i.e. 4 times the mitochondrial effective size), assuming a generation time of 8 years. Red and blue lines provide median estimates for the effective population size, assuming 5 years and 12 years as a generation time.*

S2.2 Median graph of chromosome Y sequences

Samples were mapped to the chrY contigs identified by Wallner *et al.* (34) (represented by sample HT1) and Lippold *et al.* (35) (excluding sequence G72337.1 due to overlap with the Wallner sequences), yielded a total of 193,857 bp of chrY sequences. Initial identification of candidate sequence alignments was carried out as described above (see Supplementary Section S1.3) using an index containing just these chrY contigs, excepting that quality and PCR duplicate filtering was not performed. Subsequently, these hits were re-mapped against the complete nuclear genome, with the addition of the chrY contigs, and filtering was carried out as described previously. Genotyping of the chrY contigs was carried out as described in Supplementary Section S2.5, excepting that a minimum depth of 4 and a maximum depth of 50 was used.

Scaffolds were merged into an unpartitioned supermatrix. The supermatrix was converted to RDF format using DNASP (36) v5.10.1. Median joining (37) was carried out using Network v4.612. The eight samples reported by Wallner *et al.* (34) (two Przewalski's horses and six domestic horses) were not included, as including these resulted in two distinct groups representing the samples from this study and the study Wallner *et al.* (34), a likely consequence of the (differing) systematic biases introduced by the sequencing techniques employed in the two studies (1, 34). The resulting graph matched the topology observed for the whole-genome phylogeny (Supplementary Figure S15).

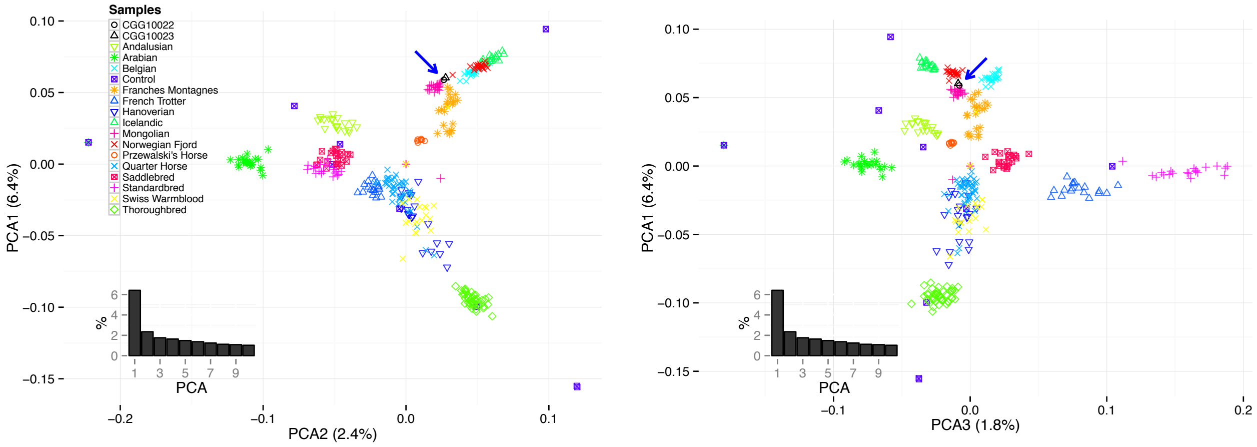


Supplementary Figure S15. Median joining network of chrY sequences

*Median joining network based on 193,875 bp of chrY sequences; individual mutations are not shown. **H_1**: Przewalski's horse; **H_2**: CGG10022; **H_3:** Icelandic horses (unnamed and P5782); **H_4:** Standardbred; **H_5:** Domestic donkey (Willy).*

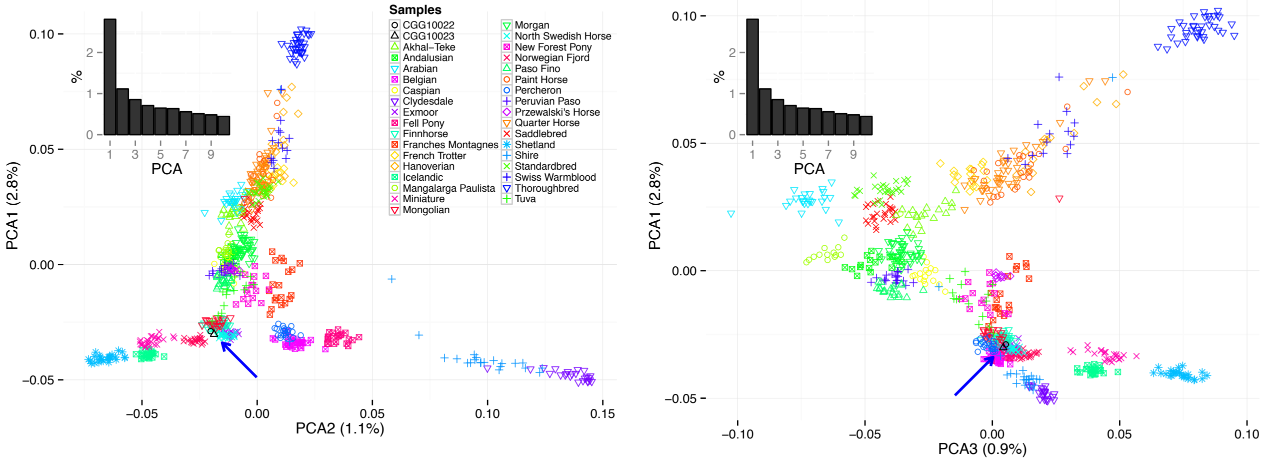
S2.3 Principal Component Analysis and Procrustes Analysis

BAM files containing the genomic coordinates of the sequenced reads mapped against EquCab2.0 were processed and converted to mpileup format by SAMTools (15). The mpileup files were used to call SNP variants following the quality filters described in Supplemental Section S2.5. We then considered the SNP variants that overlapped with the genomic coordinates covered by the equine SNP array and compared our ancient horses to the genetic diversity present amongst 9 Przewalski horses, as well as 14 (38) and 32 (39) domestic horse breeds, representing a total of 348 and 729 individuals, respectively. This resulted in the identification of 48,990 and 27,326 sites overlapping with the SNP dataset from McCue *et al.* 2012 for samples CGG10022 and CGG10023, respectively. A total of 46,672 and 25,977 sites were found in the SNP dataset from Petersen et al. 2013. Individual genotypes were converted into PLINK map and ped formats (40) and further analyzed using the software 'smartpca' of EIGENSOFT 4.0 (41). The first 10 eigen-vectors were calculated. Due to the significant lower number of sites covered for sample CGG10023, we combined our two ancient samples using a Procrustes transformation as implemented by the "proc" function of the CRAN library vegan (42). PCA plots were generated using R 2.12.2 (26) and were restricted to the first three principal components5. Supplementary Figure S16 shows PCA plots for the first three principal components following Procrustes transformation and the analysis of the reference dataset from McCue *et al.* 2012. The same analysis was performed on the dataset from Petersen *et al.* 2013 and is shown as Supplementary Figure S17.



***Supplementary Figure S16. PCA analysis of pre-domesticated and McCue et al*. horses**

Principal Component Analysis using 354 horses genotyped in McCue et al. 2012(38). The blue arrow indicates the position of the pre-domesticated samples CGG10022 and CGG10023. The barplot indicates the % of the variance explained by each of the principal components.



***Supplementary Figure S17. PCA analysis of pre-domesticated and Petersen et al*. horses**

Principal Component Analysis using 354 horses genotyped in Petersen et al. 2013 (39). The blue arrow indicates the position of the pre-domesticated samples CGG10022 and CGG10023. The barplot indicates the % of the variance explained by each of the principal components.

S2.4 Functional assessment of candidate SNPs

We screened the modern and ancient genomes investigated in this study for 50 loci associated with known diseases, coat coloration, and other phenotypical traits (Supplementary Table S11 and Supplementary Table S12), including phenotypes collected by Orlando *et al.* 2013 (1), Doan *et al.* 2012 (43), and the Mendelian Inheritance In Animals (OMIA) database (44). Genotyping was carried out as described in Supplementary Section S2.5; for sites in which the VCF record was filtered based on the criteria described in the same section, the total number of nucleotides observed at that site were tabulated as described in Orlando *et al.* 2013, *i.e.* excluding any nucleotide for which the base quality Phred score was < 35.

In the majority of cases, the two pre-domestic horses carry the reference allele; exceptions include alleles in the genes *ACN9*, *CKM*, and *COX4/1*, which are associated with racing performance, and which suggests that selection for racing performance on domestic horses acted on standing variation. While the derived allele for *CKM* and *COX4/1* is observed only for CGG10023, the derived allele for *ACN9* is observed for both CGG10022 and CGG10023.

Corroborating previous observations based on lower-coverage datasets for the pre-domesticated horse CGG10023 (1), this individual was found to be heterozygous for several loci involved body size (two loci in *PROP1*), known to be associated with dwarfism in domestic horses. Notable, this observation was further supported by the presence of derived alleles in CGG10022, which could not be included in its previous examination due to limited depth-of-coverage (1). It is noteworthy that both pre-domestic horses are heterozygous for SNPs located in *ZFAT*, all of which are associated with wither height in domestic horses. A fourth derived allele is observed only in CGG10022, for which the low coverage of CGG10022 (1 read) prevents genotyping.

The heterozygous allele observed in the *ZFAT* gene (chr9: 74,798,143) for the two pre-domesticated horses has been associated with a ~0.5 cm increase in height at the withers with regards to the deregressed estimated breeding values (dEBVs) (45). The exact effect of the other mutations observed in this gene has not been determined. None of the mutations known to be associated with various coat color phenotypes were observed for the pre-domesticated horses, preventing the determination of the coat color for these individuals. Nor were the mutations associated with spotting observed for these, nor in the Przewalski's horse, despite the fact that this phenotype is known to predate the domestication of horses (46, 47).

Ref	Chr	Coordinate	Gene	Phenotype	Mutation	ARA	CGG22	CGG23	FJO	ICE	P5782	PRZ	STD	TWI	DON
(48)	1	74,842,283	*ACTN2*	Racing performance	A>G	A/A	A/A	A4	A6	A3	A/A	A/A	A/A	A/G	A/A
(49)	1	108,249,293	*TRPM1*	Leopard complex spotting and cation channel congenital stationary night blindness	C>T	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
(50)	1	128,056,148	*PPIB*	Hereditary equine dermal isomerase B asthenia	G>A	G/G	G/G	G/G	G/G	G5	G/G	G6	G/G	G/G	G/G
(51)	1	138,235,715	*MYO5A*	Lavender foal syndrome	1 bp del	-	-	-	-	-	-	-	-	-	-
(52)	2	13,074,277	*TOE1*	Cerebellar abiotrophy	G>A	C/C	C/C	C/C	C/C	C/C	C/C	C5	C/C	C/C	C/C
(48)	3	32,772,871	*COX4/1*	Racing performance	C>T	C/C	C/C	C1,T5	C/T	C/C	T/T	C/T	C/T	C/C	C/C
(53)	3	36,259,552	*MC1R*	Chestnut coat color	C>T	C/T	C/C	C/C	C/T	C/T	T/T	C/C	C/T	C/C	C/C
(54)	3	36,259,554	*MC1R*	Chestnut coat color	G>A	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
(55)	3	77,735,520	*KIT*	Sabino spotting	A>T	A2	A/A	A1	A5	A/A	A/A	A1	A/A	A/A	A/A
(56)	3	77,740,163	*KIT*	Tobiano spotting pattern	G>A	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
(45, 57)	3	105,547,002	*LCORL / NCAPG*	Larger body size	T>C	T/T	T/T	T7	T/T	T/T	T/T	T/T	T/T	T/T	T/T
(48)	4	38,697,145	*PON1*	Racing performance	C>T	C/C	C/C	C/C	C/C	C/C	C/C	C/C	T/T	C/T	C/C
(48)	4	38,969,307	*PDK4*	Racing performance	C>A	C/C	C/C	C5	C3	C/C	C/C	C3	C/C	A/C	A/A
(48)	4	38,973,231	*PDK4*	Racing performance	G>A	G/G	G/G	G3	G3	G5	A/G	G7	G/G	A/G	A/A
(48)	4	40,279,726	*ACN9*	Racing performance	C>T	C/C	C/T	C/T	C/C	T/T	C/T	C/C	T/T	C/T	T/T
(58)	4	96,375,588	*CLCN1*	Congenital myotonia	A>C	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A/A	A6
(59)	5	20,256,789	*LAMC2*	Junctional epidermolysis bullosa	1 C ins	-	-	-	-	-	-	-	-	-	-
(60, 61)	6	11,429,753	*PAX3*	Splashed white coat	C>T	C/C	C/C	C/C	C7	C7	C/C	C/C	C/C	C/C	C5
(62)	6	73,665,304	*PMEL17*	Silver coat color	G>A	G/G	G/G	G/G	G/G	G2	G/G	G/G	G/G	G/G	G/G
(57)	6	81,481,065	*HMGA2*	Larger body size	C>T	T2	C2	C4	T2	T4	T/T	T4	C/T	C/T	C/C
(63)	8	45,603,643	*LAMA3*	Junctional epidermolysis bullosa	6,589 bp del	-	-	-	-	-	-	-	-	-	-
(64)	9	35,528,429	*DNAPK*	Severe combined immunodeficiency	5bp del	-	-	-	-	-	-	-	-	-	-
(45)	9	74,795,013	*ZFAT*	Wither height	C>T	T/T	C/T	C1,T1	C/C	T/T	C/C	C/T	C/C	C/C	C/C
(45)	9	74,795,089	*ZFAT*	Wither height	C>A	A/A	A/C	C1	C/C	A3	C/C	A/C	C/C	C/C	C/C
(45)	9	74,795,236	*ZFAT*	Wither height	G>A	A/A	A/G	A3,G4	G/G	A/A	G/G	A/G	G/G	G/G	C2
(45)	9	74,798,143	*ZFAT*	Wither height	G>A	A/A	A/G	A/G	G/G	A/A	G/G	A/G	G/G	G/G	G/G
(57)	9	75,550,059	*ZFAT*	Larger body size	C>T	C/C	C/C	C/C	C/C	C1	C/C	C/C	C/T	T/T	C/C

Supplementary Table S11. Functional assessment of SNPs in horses and the domestic donkey (chromosomes 1 – 9)

*The following abbreviations are used; **ARA**bian, Norwegian **FJO**rd, **ICE**landic (unnamed), Icelandic (**P5782**), **STA**ndardbred, **THO**roughbred (Twilight), **CGG**10**22**, **CGG**10**23**, **PRZ**ewalski, and domestic **DON**key (Willy).*

Ref	Chr	Coordinate	Gene	Phenotype	Mutation	ARA	CGG22	CGG23	FJO	ICE	P5782	PRZ	STD	TWI	DON
(65)	10	9,554,699	*RYR1*	Malignant hyperthermia	C>G	C7	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
(66)	10	15,884,567	*CKM*	Racing performance	G>A	A/G	G/G	A/G	A/G	A3,G1	A/A	G/G	G/G	A/G	G4
(38)	10	18,940,324	*GYS1*	Polysaccharide storage myopathy	C>T	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
(67)	11	15,500,439	*SCN4A*	Equine hyperkalemic periodic paralysis	C>T	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/C
(68)	11	19,184,674	*ITGA2B*	Glanzmann Thrombasthenia	Del 10 bp	-	-	-	-	-	-	-	-	-	-
(57)	11	23,259,732	*LASP1*	Larger body size	G>A	G/G	G/G	G4	G4	A6	G/G	A6	G/G	G1	G/G
(69)	14	3,761,254	*PROP1*	Dwarfism	G>C	G/G	C/G	C4,G2	C/G	G/G	C/G	G/G	G/G	G/G	G/G
(69)	14	3,761,355	*PROP1*	Dwarfism	T>C	T/T	C/C	C4,T2	C/T	T/T	C/T	T/T	T4	T/T	C/C
(69)	14	5,418,619	*ND*	Dwarfism	G>A	G/G	G/G	G/G	G/G	G/G	A/A	A/G	G/G	A/G	G/G
(70)	14	26,701,092	*SLC36A1*	Champagne dilution	G>C	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G	G/G
(71)	14	27,991,841	*SCL26A2*	Autosomal recessively inherited chondrodysplasia	A>G	G/G	G/G	G4	G/G	G/G	G/G	G/G	G/G	A/G	G/G
(60, 61)	16	20,103,081	*MITF*	Macchiato, hearing loss	T>C	T6	T/T	T/T	T/T	T3	T/T	T/T	T/T	T/T	T/T
(60, 61)	16	20,105,348	*MITF*	Splashed white coat	5 bp del	-	-	-	-	-	-	-	-	-	-
(60, 61)	16	20,117,302	*MITF*	Splashed white coat	11bp indel	-	-	-	-	-	-	-	-	-	-
(72)	17	50,624,658	*EDNRB*	Lethal white foal syndrome	GA>CT	-	-	-	-	-	-	-	-	-	-
(73, 74)	18	66,493,737	*MSTN*	Optimum racing distance	T>C	T7	T7	T7	T3	T4	T/T	T4	T4	T/T	T/T
(75)	21	30,666,626	*SLC45A2*	Cream coat color	G>A	G/G	G/G	G/G	G/G	G2	G/G	G/G	G/G	G/G	G/G
(66)	22	22,684,390	*COX4/2*	Racing performance	C>T	C/T	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C/T	C/C
(76)	22	25,168,567	*ASIP*	Black and bay color	Del 11 bp	-	-	-	-	-	-/+	-	-	-	-
(2)	23	22,999,655	*DMRT3*	Pattern of locomotion (altered gait)	C>A	C/C	C/C	C/C	C/C	A4	C/C	C/C	A/A	C/C	C/C
(77)	26	30,660,224	*SLC5A3*	Foal immunodeficiency syndrome	C>T	C/C	C/C	C/C	C7	C/C	C/C	C/C	C/C	C/C	C3
(78)	X	49,635,250	*AR*	Androgen insensitivity syndrome (AIS)	A>G	A/A	A/A	A/A	A7	A1	A/A	A1	A/A	A/A	A2
(79)	X	122,833,887	*IKBKG*	Incontinentia pigmenti	C>T	C/C	C/C	C/C	C/C	C/C	C/C	C/C	C4	C/C	C4

Supplementary Table S12. Functional assessment of SNPs in horses and the domestic donkey (chromosomes 10 – X)

*The following abbreviations are used; **ARA**bian, Norwegian **FJO**rd, **ICE**landic (unnamed), Icelandic (**P5782**), **STA**ndardbred, **THO**roughbred (Twilight), **CGG**10**022**, **CGG**10**023**, **PRZ**ewalski, and domestic **DON**key (Willy).*

S2.5 Phylogenetic inference using super-matrix of nuclear coding sequences

Phylogenetic inference was carried out using a partitioned supermatrix of the coding sequences of protein-coding genes from Ensembl v72 (80); the longest transcript was selected for each gene, excluding any transcript for which the CDS was not divisible by 3; consequently 55 of 20,449 genes were excluded. Subsequently, a small subset of very poorly covered genes (10 in total) found to cause failures during bootstrapping, due to sampling generating alignments in which one or more nucleotide types were entirely absent, were excluded. The final selection included 20,384 genes out of 20,449 protein-coding genes included in the Ensemble release.

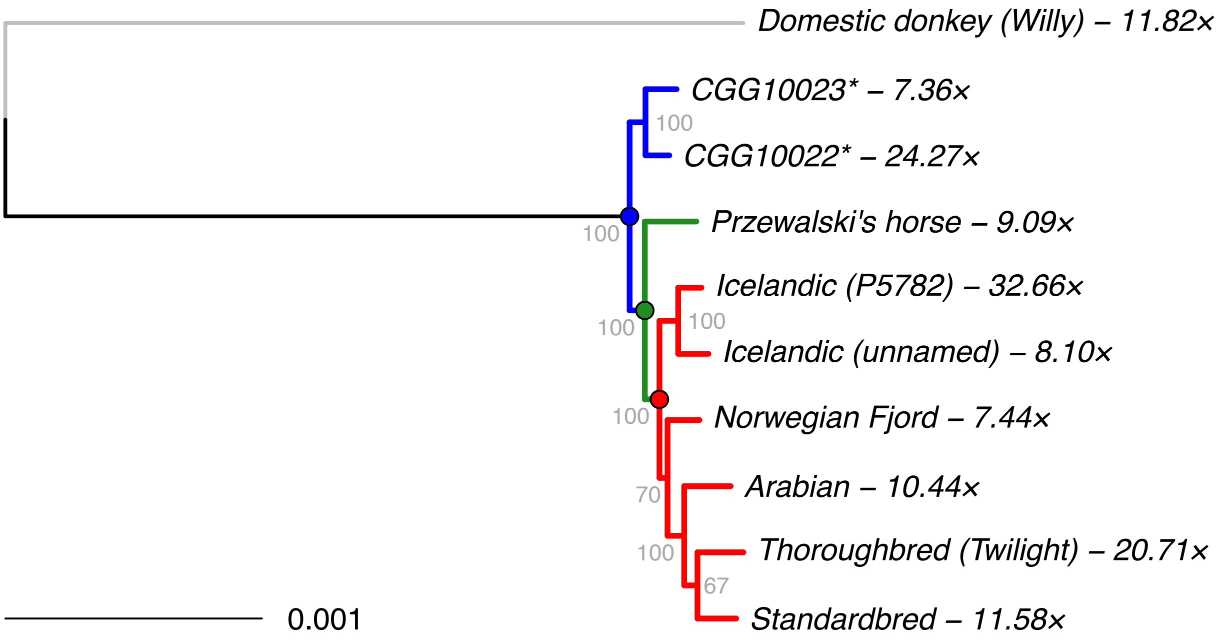
Genotyping was performed as described previously [Orlando *et al.* 2013, see their Supplementary Information, section S8.3] (1) using the PALEOMIX pipeline (10), with few modifications: Firstly, the maximum depth of coverage per site was set to > 0.995 of the coverage distribution (excluding sites with depth 0; Supplementary Table S13) for the sample being genotyped, and secondly, no filtering of singletons was done for the pre-domesticated horses. Variant, reference sites, and indels were called, but indels were only used to filter SNPs adjacent to indels, and not included in the final sequences, and no multiple-sequence alignment was done following genotyping. Sites containing heterozygous SNPs were represented using the standard IUPAC codes (81). The sequences were subsequently merged into a partitioned super-matrix, with two partitions for each gene: The first partition covered codon positions 1 and 2, and the second partition covered codon position 3, for a total of 31,220,478 columns in 40,768 partitions. RAxML (82) v7.3.2 was used to remove columns / partitions that consisted purely of uncalled bases (N), reducing the final super-matrix to 30,185,251 columns in 40,096 partitions.

Phylogenetic inference was carried out using the PALEOMIX pipeline (10); briefly 100 bootstrap pseudo-replicate alignments were generated from the super-matrix, and parsimony starting trees were generated using RAxML for both the original super-matrix and the bootstrap super-matrices. Phylogenetic inference was carried out for each super-matrix using ExaML v1.0.2 (http://sco.h-its.org/exelixis/software.html) under the "GAMMA" model of nucleotide substitutions (Supplementary Figure S18), and the starting trees generated above. The resulting phylogenetic trees were rooted on the midpoint.

Sample	Max	Sample	Max
Arabian	39	Icelandic (unnamed)	30
CGG10022	72	Norwegian Fjord	23
CGG10023	35	Przewalski's horse	30
Domestic donkey (willy)	28	Standardbred	37
Icelandic (P5782)	69	Thoroughbred (Twilight)	45

Supplementary Table S13. Per-sample maximum depths for SNP quality filtering

Maximum depth of coverage for quality filtering of SNPs, determined as the > 0.995 quantile of the per-site depth distribution.


Domestic donkey (Willy) – 11.82×
CGG10023* – 7.36×
100
CGG10022* – 24.27×
100
Przewalski's horse – 9.09×
Icelandic (P5782) – 32.66×
100
100
Icelandic (unnamed) – 8.10×
100
Norwegian Fjord – 7.44×
70
Arabian – 10.44×
100
Thoroughbred (Twilight) – 20.71×
67
Standardbred – 11.58×
0.001

Supplementary Figure S18. Phylogenetic tree of domesticated and wild horses

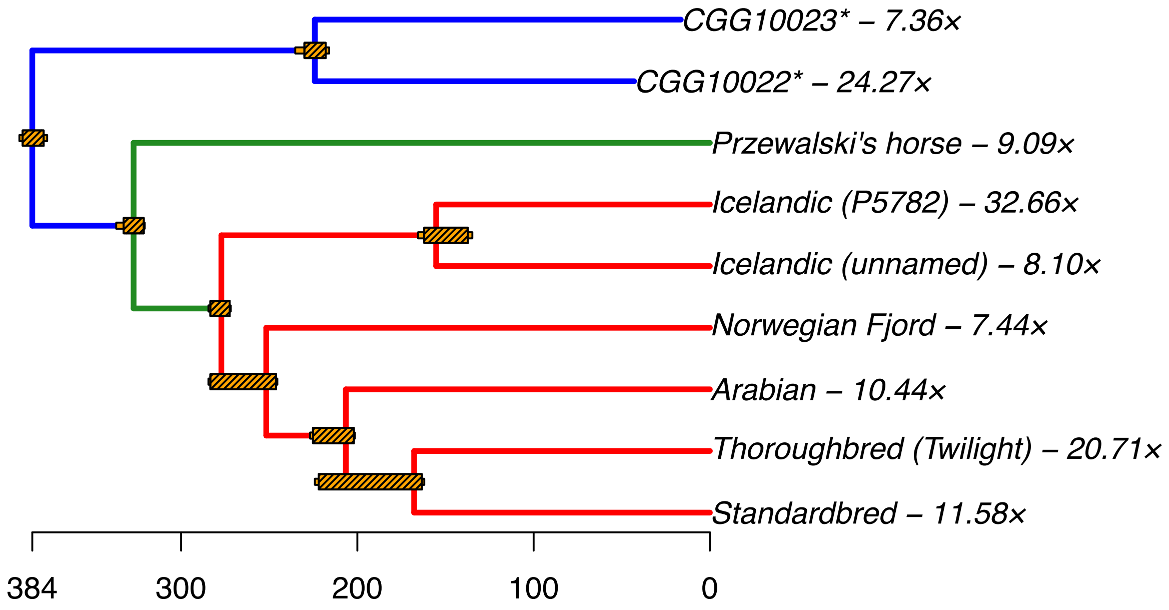
Pre-domesticated horses are marked in blue, modern, wild horses are marked in green, and domesticated horses are marked in red respectively; node labels show bootstrap support calculated using 100 bootstrap trees. Average depth-of-coverage relative to the EquCab2.0 reference genome is listed after the sample name. Samples marked with a star were generated for this study (see section S1.1).

S2.6 Dating of the most recent common ancestors

TMRCAs were estimated for the bootstrap trees generated during phylogenetic inference (see section S2.5) using r8s (83), and using the LF method (molecular clock, maximum likelihood) and the NPRS method (relaxed molecular clock; Non-Parametric Rate Smoothing), both using the POWELL algorithm. In addition the following parameters were set:

- A random seed for each run
- ftol = 10^{-9}
- num_time_guesses = 10
- num_restarts = 10
- maxiter = 2000

The date of the root node was constrained to 4.0-4.5 Mya following Orlando *et al*. 2013 (1), and the dates of CGG10022 and CGG10023 were fixed at 43 kyr and 16.5 kyr BP respectively, representing the midpoint of the calibrated radiocarbon dates (see Supplementary Section S1.1). To account for differing topologies in the bootstrap trees, the dating of the smallest clade forming a superset of a clade to be dated was determined for each bootstrap tree, and the median and CI determined based on these dates for both the molecular clock (Supplementary Figure S19 and Supplementary Table S14) and the relaxed clock (Supplementary Figure S20 and Supplementary Table S15).


CGG10023* – 7.36×
CGG10022* – 24.27×
Przewalski's horse – 9.09×
Icelandic (P5782) – 32.66×
Icelandic (unnamed) – 8.10×
Norwegian Fjord – 7.44×
Arabian – 10.44×
Thoroughbred (Twilight) – 20.71×
Standardbred – 11.58×
384
300
200
100
0

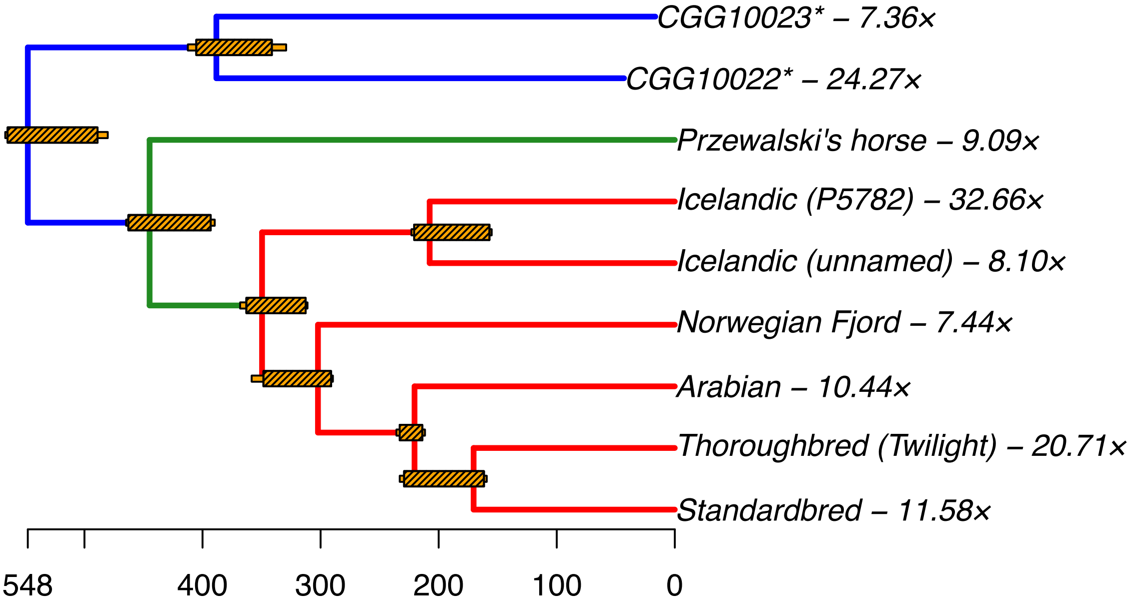
Supplementary Figure S19. Chronogram of wild and domesticated horses

Chronogram using the median ages estimated using r8s (Supplementary Section S2.6) under a molecular clock. Ages are shown in thousands of years. Pre-domesticated horses are marked in blue, modern, wild horses are marked in green, and modern, domesticated horses are marked in red respectively. Average depth-of-coverage relative to the EquCab2.0 reference genome is listed after the sample name. Samples marked with a star were generated for this study (see section S1.1). The dashed, orange bars indicate the 95% quantiles of the date estimates. The thin, orange bars indicate the minimum and maximum estimates.

	Min	Q(0.025)	Median	Q(0.975)	Max
Root	4000.38	4000.4	4000.53	4000.56	4000.56
Horses	375.99	377.98	384.44	389.91	391.76
Modern horses	320.66	321.18	326.97	332.62	336.83
Domestic horses	271.86	272.48	277.17	283.41	284.51
Arabian, Norwegian Fjord, Standardbred, and Thoroughbred (Twilight)	245.39	246.10	251.69	283.41	284.51
Arabian, Standardbred, and Thoroughbred (Twilight)	201.42	202.00	206.54	225.18	226.81
CGG10022 and CGG10023	215.97	218.01	224.15	230.04	235.23
Icelandic (P5782) and Icelandic (unnamed)	134.77	137.35	155.30	162.05	165.58
Standardbred and Thoroughbred (Twilight)	162.03	163.39	167.79	221.99	224.12

Supplementary Table S14. Times (kyr) to the most recent common ancestors

Divergence times were estimated using 'r8s' in thousands of years, under the assumption of a molecular clock, based on 100 bootstrap phylogenies generated from the CDS supermatrix (Supplementary Section S2.5). See Supplementary Figure S19 *for the list of domesticated and modern horses.*



Supplementary Figure S20. Relaxed chronogram of wild and domesticated horses

Chronogram using the median ages estimated using r8s (Supplementary Section S2.6) under a molecular clock. Ages are shown in thousands of years. Pre-domesticated horses are marked in blue, modern, wild horses are marked in green, and modern, domesticated horses are marked in red respectively. Average depth-of-coverage relative to the EquCab2.0 reference genome is listed after the sample name. Samples marked with a star were generated for this study (see section S1.1). The dashed, orange bars indicate the 95% quantiles of the date estimates. The thin, orange bars indicate the full range of estimates.

	Min	Q(0.025)	Median	Q(0.975)	Max
Root	4499.81	4499.81	4499.85	4499.91	4499.92
Horses	480.25	488.95	547.99	565.11	567.09
Modern horses	389.65	393.07	444.81	462.82	464.69
Domestic horses	311.08	312.70	349.58	362.89	368.15
Arabian, Norwegian Fjord, Standardbred, and Thoroughbred (Twilight)	289.57	291.11	302.24	348.41	358.39
Arabian, Standardbred, and Thoroughbred (Twilight)	211.60	213.81	220.44	233.05	235.78
CGG10022 and CGG10023	329.16	341.29	388.26	405.31	412.43
Icelandic (P5782) and Icelandic (unnamed)	155.08	156.99	207.62	220.79	223.05
Standardbred and Thoroughbred (Twilight)	159.29	161.63	170.40	229.32	232.97

Supplementary Table S15. Times (kyr) to the most recent common ancestors (relaxed clock)

Ages of population splits estimated using 'r8s' in thousands of years, under the assumption of a relaxed molecular clock (NPRS), based on 100 bootstrap phylogenies generated from the CDS supermatrix (Supplementary Section S2.5). See Supplementary Figure S20 *for the list of domesticated and modern horses.*

S2.7 Gene flow between pre-domesticated and modern horses

S2.7.1 Background

The ABBA-BABA test was used to examine the presence of gene flow between the pre-domesticated horses (CGG10022 and CGG10023), and the modern horse breeds, using the domestic donkey (Willy) as outgroup. The methodology and theory behind the ABBA-BABA test was originally described by Green et al. 2010 (24) and Durand et al. 2011 (84). The implementation used here is described in Orlando et al. 2013 [see their Supplementary materials S12.1] (1). The Thoroughbred (Twilight) was excluded from the ABBA-BABA test, as the horse reference genome (EquCab2.0) was based on this individual, and remapping of reads from this individual is expected to introduce systematic bias into the test (i.e. this individual is significantly closer to the reference than any other horse).

To briefly summarize Orlando *et al.* 2013, given taxa H1, H2, and H3, and an outgroup, the ABBA-BABA test examines if the topology (((H1, H2), H3), outgroup) is correct, and whether or not there has been gene flow between H3 and H1 on one hand, or between H3 and H2 on the other hand. To accomplish this, bi-allelic loci in which the outgroup carries the (ancestral) allele A, and in which H3 carries the (derived) allele B, and where H1 and H2 carries different alleles (A or B) are tabulated. This yields two possible combinations of alleles: (((A, B), B), A) and (((B, A), B), A). Given the null hypothesis that the topology is correct, and there has been no gene flow between H3 and H1, or between H3 and H2, the two patterns result from incomplete lineage sorting, and are thus equally likely to occur. If however, the topology is incorrect, or if gene flow has occurred between H3 and either H1 or H2, one or the other pattern is expected to dominate.

This is tested using the statistic developed by Green *et al.* 2010 (24), where D = (nABBA – nBABA) / (nABBA + nBABA). If the null hypothesis is correct, this value should be close to zero, while a negative value suggests that H3 is closer to H1 than H2, and a positive value suggests that H3 is closer to H2 than H1. The standard error was estimated using "*delete-m Jackknife for unequal m*" based on 10 Mbp blocks (85). The blocks size was chosen to accommodate the large amount of linkage disequilibrium in horses. The Z-score is given as a measure of significance; absolute values greater than 3 indicate a statistically significant deviation from the null hypothesis.

S2.7.2 Results

To examine the presence of gene flow between the ancient samples and the domestic horses, each pairwise combination of the modern domestic horses were used for H1 and H2, and tested with either CGG10022 or CGG10023 as H3, and using the Domestic donkey (Willy) as the outgroup. The results are tabulated in Supplementary Table S16. None of the tests performed were found to be significant, consistent with a lack of gene flow between the pre-domesticated horses and any particular domesticated breed.

To examine the presence of gene flow between the ancient samples, and the modern breeds (including Przewalski's horse), each combination of a domestic horse, and Przewalski's horse as H1 and H2, with either CGG10022 or CGG10023 as H3, and the domestic donkey (Willy) as the outgroup was tested. The results are tabulated in Supplementary Table S17, all of which indicates a small but statistically significant (D = 0.03; Z-score > 5.4) violation of the topology (((Przewalski's horse, domestic horse), pre-domestic), outgroup).

In addition, we examined each quartet involving domestic horses as H1 and H2, the Prezwalski's horse as H3, and the domestic donkey as the outgroup, in order detect the presence gene flow between the Przewalski's horse and the domesticated horses. All quartets were non-significant, except for quartets involving the Icelandic (P5782) horse. This individual, however, was generated using a different methodology from that of the remaining samples (Supplementary Section S1.1). Therefore, this outcome likely reflects the sensitivity of D-statistic to the sequencing procedure used while generating this genome (2).

H1	H2	H3	Delta	Total	D	$D_{Jackknife}$	D_{Sd}	Z-score
ARA	FJORD	CGG10022	504	780476	0.00	0.00	0.004	0.1
ARA	ICE	CGG10022	-1976	751664	0.00	0.00	0.004	-0.6
FJORD	ICE	CGG10022	-2055	745203	0.00	0.00	0.004	-0.7
ARA	P5782	CGG10022	-2023	817857	0.00	0.00	0.004	-0.6
FJORD	P5782	CGG10022	-2306	801968	0.00	0.00	0.004	-0.7
ICE	P5782	CGG10022	569	724327	0.00	0.00	0.004	0.2
STA	P5782	CGG10022	1219	836645	0.00	0.00	0.004	0.3
ARA	STA	CGG10022	-3008	749934	0.00	0.00	0.004	-0.9
FJORD	STA	CGG10022	-3370	792782	0.00	0.00	0.004	-1.0
ICE	STA	CGG10022	-966	771734	0.00	0.00	0.004	-0.3
ARA	FJORD	CGG10023	-952	716326	0.00	0.00	0.004	-0.3
ARA	ICE	CGG10023	3145	688797	0.00	0.00	0.004	1.0
FJORD	ICE	CGG10023	4283	683197	0.01	0.01	0.004	1.7
ARA	P5782	CGG10023	883	745339	0.00	0.00	0.004	0.3
FJORD	P5782	CGG10023	2026	730792	0.00	0.00	0.004	0.8
ICE	P5782	CGG10023	-2090	657014	0.00	0.00	0.004	-0.7
STA	P5782	CGG10023	-742	761586	0.00	0.00	0.004	-0.2
ARA	STA	CGG10023	1879	686335	0.00	0.00	0.004	0.7
FJORD	STA	CGG10023	2925	726749	0.00	0.00	0.004	1.1
ICE	STA	CGG10023	-1635	705431	0.00	0.00	0.004	-0.6

Supplementary Table S16. ABBA-BABA tests between domestic and pre-domesticated horses

Delta is nABBA – nBABA; total is nABBA + nBABA; D is the statistic described in section S2.7; $D_{Jackknife}$ and D_{Sd} is the jackknife estimate of D and the associated standard-deviation; Z-scores are significant for values below -3 and values greater than 3; no result is statistically significant. Arabian is abbreviated as ARA, Icelandic (unnamed) as ICE, Icelandic (P5782) as P5782, Norwegian Fjord as FJORD, and Standardbred as STA. The domestic donkey (Willy) is used as the outgroup.

H1	H2	H3	Delta	Total	D	$D_{Jackknife}$	D_{Sd}	Z-score
PRZ	P5782	CGG10022	27429	896257	0.03	0.03	0.005	6.6
PRZ	ARA	CGG10022	28471	855803	0.03	0.03	0.005	6.7
PRZ	FJORD	CGG10022	28793	844897	0.03	0.03	0.005	6.7
PRZ	ICE	CGG10022	25886	826186	0.03	0.03	0.005	6.7
PRZ	STA	CGG10022	25794	872694	0.03	0.03	0.005	5.9
PRZ	P5782	CGG10023	22872	818208	0.03	0.03	0.005	5.6
PRZ	ARA	CGG10023	21328	786418	0.03	0.03	0.005	5.4
PRZ	FJORD	CGG10023	20671	775691	0.03	0.03	0.005	5.4
PRZ	ICE	CGG10023	24065	757525	0.03	0.03	0.005	6.5
PRZ	STA	CGG10023	23385	801365	0.03	0.03	0.005	5.7

Supplementary Table S17. ABBA-BABA tests between ancient, domestic and Przewalski's horse

Delta is nABBA – nBABA; total is nABBA + nBABA; D is the statistic described in section S2.7; $D_{Jackknife}$ and D_{Sd} is the jackknife estimate of D and the associated standard-deviation; Z-scores are significant for values below -3 and values greater than 3; all results are statistically significant. Arabian is abbreviated as ARA, Icelandic (unnamed) as ICE, Icelandic (P5782) as P5782, Norwegian Fjord as FJORD, Przewalski's horse as PRZ, and Standardbred as STA. The domestic donkey (Willy) is used as the outgroup.

H1	H2	H3	Delta	Total	D	$D_{Jackknife}$	D_{Sd}	Z-score
ARA	FJO	PRZ	9611	837791	0,01	0,01	0,005	2,1
ARA	ICE	PRZ	469	803937	0,00	0,00	0,005	0,1
FJO	ICE	PRZ	-8358	799970	-0,01	-0,01	0,005	-2,2
ARA	STD	PRZ	-284	802264	0,00	0,00	0,005	-0,1
FJO	STD	PRZ	-10182	850770	-0,01	-0,01	0,005	-2,2
ICE	STD	PRZ	-1085	826023	0,00	0,00	0,005	-0,3
ARA	P5782	PRZ	-10882	870620	-0,01	-0,01	0,005	-2,5
FJO	P5782	PRZ	-20913	857015	-0,02	-0,02	0,005	-5,0
ICE	P5782	PRZ	-11298	770846	-0,01	-0,01	0,005	-3,1
STD	P5782	PRZ	-10962	890656	-0,01	-0,01	0,005	-2,6

Supplementary Table S18. ABBA-BABA tests between two domestic and Przewalski's horse

Delta is nABBA – nBABA; total is nABBA + nBABA; D is the statistic described in section S2.7; $D_{Jackknife}$ and D_{Sd} is the jackknife estimate of D and the associated standard-deviation; Z-scores are significant for values below -3 and values greater than 3; all results are statistically significant. Arabian is abbreviated as ARA, Icelandic (unnamed) as ICE, Icelandic (P5782) as P5782, Norwegian Fjord as FJORD, Przewalski's horse as PRZ, and Standardbred as STA. The domestic donkey (Willy) is used as the outgroup.

S2.7.3 Proportion of the genome of domesticated horses deriving from admixture

We used the f,^ estimator to calculate the proportion of the genome within domesticated horses that might be derived from admixture with the ancient horse population (24, 84). This procedure is formally described by Cahill and coworkers (86) and estimates the proportion of an admixed genome by the fraction of derived allele sharing compared to that observed in a completely admixed individual. Originally, two individuals I_1 and I_2 from a first species (species$_I$) are considered, and two individuals M_1 and M_2 from another species (species$_M$). Admixture is detected between I_2 and M_2. We can calculate the proportion of the I_2 genome that result from admixture with M_2, f,$^\wedge$, as:

$$f,^{\wedge} = ABBA\text{-}BABA_{(I1,I2;M2,Outgroup)} / ABBA\text{-}BABA_{(I1,M1;M2,Outgroup)}$$

where ABBA-BABA represents the numerator of the D-statistic (A=Ancestral allele, B=Derived allele, relative to the outgroup). This fraction provides a minimal boundary for the proportion of the admixed genome to deriving from admixture, which is proportional to the time difference between the admixture and speciation event. Here, we calculated f,$^\wedge$ estimators for all combinations of two ancient horses (representing M_1 and M_2) and pairs of Przewalski's and domesticated horses in our study (representing I_1 and I_2, respectively). The admixture event detected in the following quartet (Domesticated, Przewalski; Ancient, Outgroup) has occurred following the population split between Przewalski's horses and the horse population that will later be domesticated. However, with the data presented in this study, we cannot date the admixture event, which can have occurred at any time following divergence from the Przewalski's horse. Using F-statistics (Supplementary Section S2.8.2), we estimated this time to be ~43.0-51.8 kyr ago, which represents ~29.4-44.0% of the divergence between the ancient horse population and the population that led to modern horses (including domesticated and Przewalski's horses; 117.8-146.2 kyr ago, Supplementary Table S20).

Correcting the values observed for the f,^ estimator therefore provided an upper boundary for the proportion of the genome of modern domesticated horses that results from admixture with the descent of the ancient horses **(**Supplementary Table S19**)**.

The D-statistic can be biased whenever the genomes considered as I_1, I_2 and M_1 show difference in error rates (1, 87). In such cases, errors taking place at AABA sites introduce spurious ABBA or BABA sites, that influence the D-statistic calculation, hence, the f,^ estimator. The genome of the ancient horse CGG10023 shows higher error rates than that of CGG10022 (Supplemental Section S1.4). Therefore, quartets where M_1 and M_2 are CGG10023 and CGG10022, respectively, are more subject to errors than quartets where M_1 and M_2 are CGG10022 and CGG10023. We therefore consider all calculations resulting from the later quartets more reliable.

Quartet Numerator	Quartet Denominator	All substitutions		Transversions only	
		Min. boundary (f,^)	Max. boundary (f,^ time corrected)	Min. boundary (f,^)	Max. boundary (f,^ time corrected)
(Arabian, Przewalski; CGG10022, Outgroup)	(Arabian, CGG10023; CGG10022, Outgroup)	22.9%	52.1-77.9%	20.0%	45.5-68.0%
(Fjord, Przewalski; CGG10022, Outgroup)	(Fjord, CGG10023; CGG10022, Outgroup)	23.3%	52.9-79.2%	19.9%	45.3-67.7%
(Icelandic, Przewalski; CGG10022, Outgroup)	(Icelandic, CGG10023; CGG10022, Outgroup)	21.3%	48.4-72.5%	21.4%	48.6-72.8%
(Standardbred, Przewalski; CGG10022, Outgroup)	(Standardbred, CGG10023; CGG10022, Outgroup)	19.8%	45.0-67.4%	20.0%	45.4-68.0%
(P5782, Przewalski; CGG10022, Outgroup)	(P5782, CGG10023; CGG10022, Outgroup)	21.0%	47.6-71.3%	15.7%	35.8-53.5%
(Arabian, Przewalski; CGG10023, Outgroup)	(Arabian, CGG10022; CGG10023, Outgroup)	14.8%	33.7-50.4%	14.0%	31.9-47.8%
(Fjord, Przewalski; CGG10023, Outgroup)	(Fjord, CGG10022; CGG10023, Outgroup)	14.4%	32.6-48,9%	13.7%	31.1-46.5%
(Icelandic, Przewalski; CGG10023, Outgroup)	(Icelandic, CGG10022; CGG10023, Outgroup)	17.8%	40.5-60.7%	17.3%	39.4-59.0%
(Standardbred, Przewalski; CGG10023, Outgroup)	(Standardbred, CGG10022; CGG10023, Outgroup)	16.1%	36.5-54.7%	17.3%	39.2-58.7%
(P5782, Przewalski; CGG10023, Outgroup)	(P5782, CGG10022; CGG10023, Outgroup)	15.3%	34.8-52.1%	12.9%	29.4-44.0%

Supplementary Table S19. Proportion of the genomes of domesticated horses resulting from admixture.

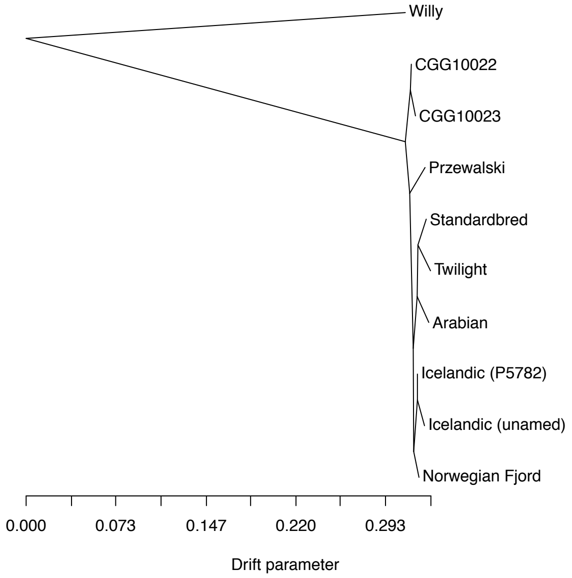
S2.8 Population tree and split times

S2.8.1 Population Tree using TreeMix

We merged the species-specific VCF files for ancient specimens, all domesticated horses, the Przewalski's horse and the donkey outgroup (including non-variants) with '*bcftools*' (15), strictly restricting to sites passing quality filters and with biallelic SNPs (Supplemental Section S2.5). Subsequently, the merged VCF file was converted into PLINK format using '*vcftools*' (88), which resulted in final SNP matrices of 4,207,193 SNPs (with donkey) or 2,686,345 SNPs (without the domestic donkey).

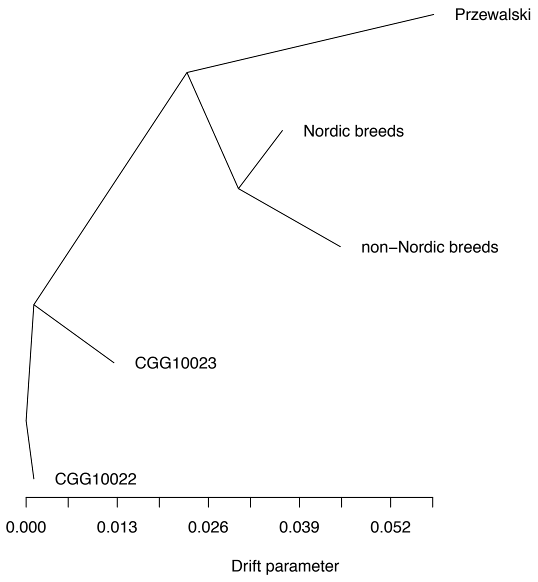
Two sets of TreeMix analyses (89) were performed in parallel, depending on whether the domestic donkey was included as outgroup or not. The PLINK SNP matrix was converted to TreeMix-format using the supplied python script ('plink2treemix.py'). In a first series of analyses, each sample was considered individually. In a second set of analyses, we grouped breeds according to their known historical affinities, namely considering the Arabian, the Standardbred and Thoroughbred (Twilight) in a first group (non-Nordic) and the Icelandic (unnamed), Icelandic (P5782) and the Norwegian Fjord in a second group (Nordic). For each analysis, we ran TreeMix considering up to one migration edge (-m 0-1), global perturbation of populations (-global) and 5,000 SNPs per block (-k 5000). Finally, the TreeMix output was plotted with the TreeMix R functions resulting in Supplementary Figure S21 to Supplementary Figure S23.

Using the individuals as distinct populations, the population tree obtained has the same topology as the phylogenetic tree of the samples (Supplementary Figure S22). The first migration edge for the analysis based on the grouped individuals was between the non-Nordic breeds and one of the ancient individuals (Supplementary Figure S23), we noticed that the gene flow direction was likely poorly inferred, as gene flow was inferred from modern domesticated horses into the ancient specimens, despite 16.5-43 kyr of time difference. However, incorrect inference of the directionality of gene flow represents one of the major types of errors for TreeMix (89), as also indicated by our admixture tests based on f_3 statistics (Supplementary Section S2.8.2).



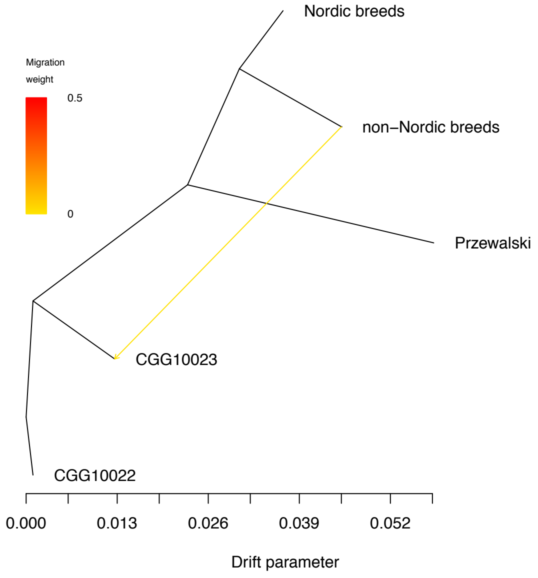
Supplementary Figure S21. **TreeMix population tree with no migration edges, considering each individual as a separate population.**

99.99959% of the variation is explained by the tree.



Supplementary Figure S22. **TreeMix population tree with no migration edges, by using the Nordic (Icelandic (unnamed), Icelandic (P5782) and the Norwegian Fjord) and non-Nordic (Arabian, the Standardbred and Twilight) grouping for the domestic breeds.**

99.99938% of the variation is explained by the tree.



Supplementary Figure S23. **TreeMix population tree with 1 migration edge, by using the Nordic (Icelandic (unnamed), Icelandic (P5782) and the Norwegian Fjord) and non-Nordic (Arabian, the Standardbred and Twilight) grouping for the domestic breeds.**

99.99990*% of the variation is explained by the graph.*

S2.8.2 Population split times

We estimated a minimal boundary for the date of divergence between horse populations using the F-statistics, originally introduced by Green and colleagues (24). Here, this statistics is calculated using 3-ways alignments, including one outgroup (the donkey) and two horses (hereafter referred to as Horse_1 and Horse_2), originating from two different populations. At sites where Horse_2 shows heterozygosity, we random-sampled one base from the sequence data underling the Horse_1 genome and counted how often the derived allele (i.e. different to that present in the Outgroup genome) was found. This frequency represents the F-statistics, which has been described to decrease with increasing population split times between Horse_1 and Horse_2.

In order to avoid alignment biases toward the horse reference, trio pileups were built using BAM files aligned to the *de novo* donkey assembly (Orlando et al. 2013). The command '*mpileup*' from the SAMtools suite (15) was used with the options -EA as described in Orlando et al. 2013 (Supplementary section 5.2.a). The following filters were then applied to the sites and discarded:

- scaffolds predicted to be from the X or Y chromosomes by Orlando et al. 2013
- bases with a quality < Phred score 30
- Depth-of-coverage for Outgroup and Horse_2 < 8
- tri-allelic sites
- indels present in any of individual in the trio
- variants, homozygous derived or heterozygous for the Outgroup to itself

Finally, scaffolds with a minimal size of 10 kb and with at least 5 kb covered sites were retained. Coalescent simulations under a simple model consisting of two populations splitting at a given time T were then performed across a uniform grid of possible value for T (every 2,500 years; Supplementary Table S20), and Approximate Bayesian Computation was used to recover a posterior distribution for T (Supplementary Figure S24), following the methodology presented by Orlando and colleagues (1) and using the '*abc*' R package (90) with loclinear regression method and a tolerance value of 2.5%. This procedure is known to be sensitive to the demographic model used for simulations of the Horse_2 population, but is robust to that considered for the Horse_1 population (1, 24). We therefore used a simplified demographic model deriving from PSMC inference (Supplemental Section S2.10) for time periods older than 10,000 years. As PSMC inference is known to be unreliable for younger times, we relied on the recent demographic expansion described by Lippold and coworkers (35) and Achilli and coworkers (91) based on the variation present in mitochondrial genomes. We averaged the population size estimated in both studies. Details about the demographic model are presented in Supplementary Figure S24. We performed a total number of 100 simulations per population time split using *fastsimcoal2* (92) and independent genomic blocks of at least 5 kb. These blocks corresponded to all *de novo* genomic contigs available for the donkey outgroup and where sequence information was available for both Horse_1 and Horse_2 genomes. We used a generation time of 8 years and a global mutation rate of $7.242\text{x}10^{-9}$ mutation per site per generation for all mutation types and $2.108\text{x}10^{-9}$ mutation per site per generation for transversions (1).

Of note, coalescent simulations were performed without gene flow post-population split. Yet, following gene flow, derived mutations specific to any of the two horse populations will be shared with the other population, resulting in an increase of the F-statistics and longer time periods would be needed before the value for the F-statistics observed between Horse_1 and Horse_2 could be reached. D-statistics calculation and prior work (1) show that assumption of isolation is valid when considering populations of domesticated and Przewalski's horses. However, admixture tests (SI section S2.7) have revealed the presence of significant gene flow between the population of our ancient horses (and their descent) and the population of domesticated horses. Therefore, the methodology used only provides a minimal boundary for the age of population split between ancient and modern horses.

In order to determine how our estimates for the time of this population split are affected by increasing amounts of gene flow, we performed the same simulations as those described above but assuming three possible migration rates (Supplementary Table S21). Migrations were assumed to go from the population including the ancient horses into the population leading to modern domesticated horses. The range selected spanned three orders of magnitudes (10^{-4}, 10^{-5} and 10^{-6} migrants per generation) and comprised migration rate estimates recovered from *∂a∂i* (SI section S2.9). In our simulations, such migration rates correspond to 1 migrant per generation, or 1 migrant per 10 or 100 generations. Migrations were allowed between 40 kya and 4 kya, following our findings that the gene flow could have occurred at any time after the split between Przewalski's horses and domesticated horses, and until the early stages of domestication. We then performed ABC (tolerance = 0.025) to recover posterior distributions of population time splits that we compared to those obtained in absence of gene flow (Supplementary Figure S25). We observed only a marginal effect on the population split time between ancient and modern horses from increasing levels of gene flow, except when migration rates were superior to 1 migrant per generation to 1 migrant per generation. These analyses were performed on the two following trios (Outgroup,(CGG10023, Icelandic)) and (Outgroup,(CGG10023,Twilight)).

In order to test the migration edge inferred between the CGG10023 and the non-Nordic breeds by TreeMix (Supplementary Section S2.8.1), we calculated f_3-statistics (93) using the 'threepop' program (-k 5000) provided in the TreeMix package. Here we use the same notation as Patterson et al. (93), that is, if f_3(C;A,B) is negative then it suggests that population C has contribution from both ancestral populations of A and B. The resulting f_3-statistics are in Supplementary Table S22 to Supplementary Table S24. Given that the addition of the migration edge explains a minute proportion of the residual variance (0.00052%; compare Supplementary Figure S22 and Supplementary Figure S23) and the f_3(CGG10023; CGG10022, non-Nordic horses) statistics is positive (f_3 = 0.0114; Supplementary Table Supplementary Table S22), we conclude that this inferred migration edge is probably of spurious nature.

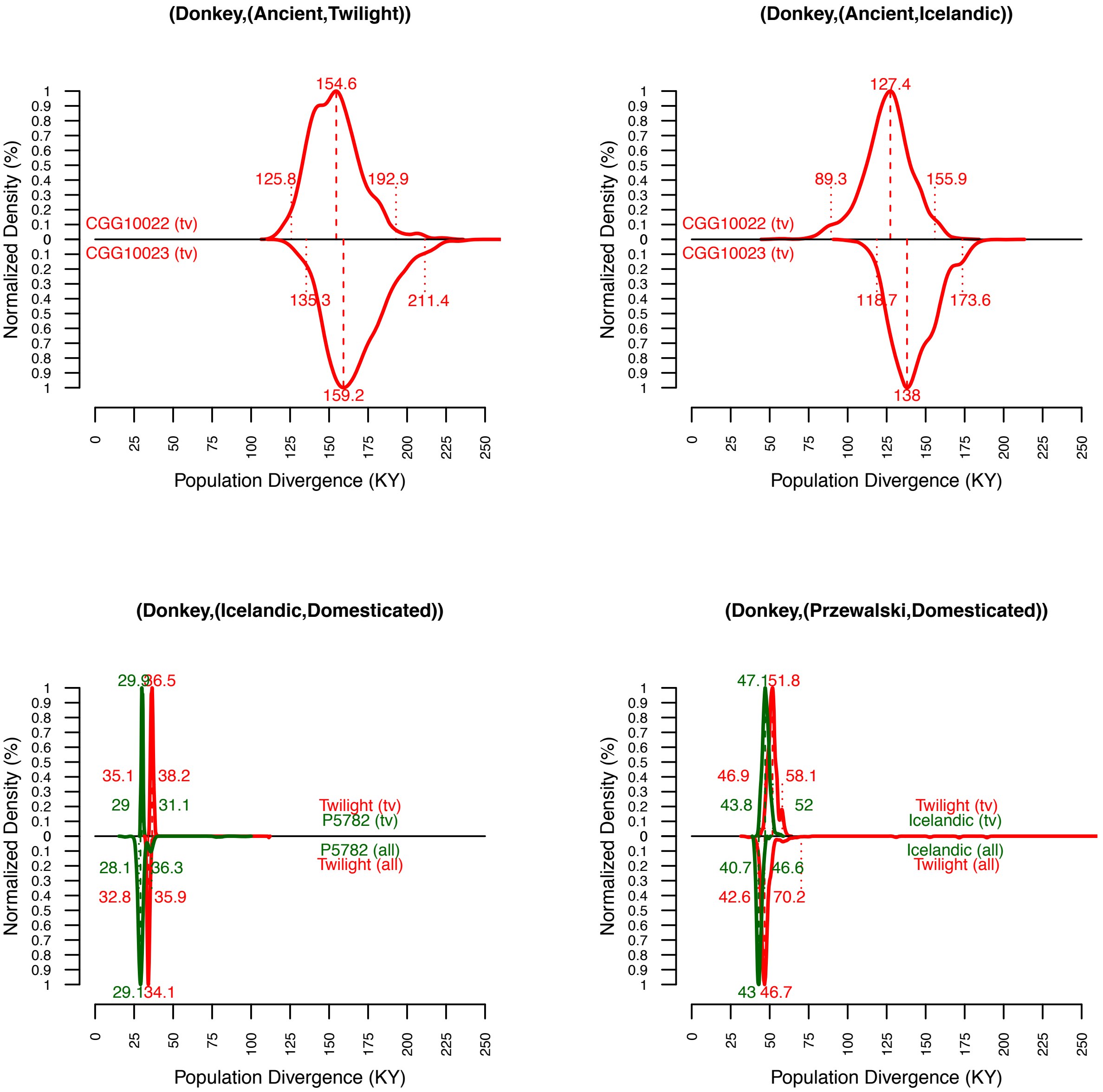
We further tested the migration edge inferred by TreeMix by attempting to detect the presence of gene flow between CGG10023 and the modern horses using D-statistics (Supplementary Table S25), and found no support for admixture in all but one test, namely between CGG10023 and the Arabian horse. This is in contrast with the migration edge inferred by TreeMix (from non-Nordic breeds into CGG10023) and hence we do not consider this as supporting the migration edge inferred by TreeMix. We further tested for the possibility of gene flow from the modern breeds into the ancient populations using f_3 (Supplementary Table S26) but, as all tests yielded positive f_3 values, we did not find support for such gene flow in any of the tests.

						Posterior Time (T, KY)		
3-way (Outgroup,(Horse$_1$,Horse$_2$))	Mutation	#Blocks	#Length (bp)	T_{min} (KY)	T_{max} (KY)	2.5%	mode	97.5%
(Outgroup, (CGG10022, Twilight))*	tv	32,083	1,596,551,988	47.5	450	125.8	154.6	192.9
	all	32,083	1,596,551,988	47.5	450	80.7	117.9	145.4
(Outgroup, (CGG10023, Twilight))*	tv	31,608	1,553,167,427	20	450	135.3	159.2	211.4
	all	31,608	1,553,167,427	20	450	103.1	126.3	153.1
(Outgroup, (CGG10022, Icelandic))*	tv	24,663	751,879,578	47.5	450	89.3	127.4	155.9
	all	24,663	751,879,578	47.5	450	57.2	76.3	115.4
(Outgroup, (CGG10023, Icelandic))*	tv	24,444	736,932,893	20	450	118.7	138.0	173.6
	all	24,444	736,932,893	20	450	68.9	107.1	128.7
(Outgroup, (Twilight, Icelandic))	tv	24,808	758,019,473	2.5	300	35.1	36.5	38.2
	all	24,808	758,019,473	2.5	300	32.8	34.1	35.9
(Outgroup, (P5782, Icelandic))	tv	24,815	758,350,065	2.5	300	29.0	29.9	31.1
	all	24,815	758,350,065	2.5	300	28.1	29.1	36.3

(Outgroup, (Przewalski, Icelandic))	tv	24,729	753,877,917	2.5	300	43.8	47.1	52.0
	all	24,729	753,877,917	2.5	300	40.7	43.0	46.6
(Outgroup, (Przewalski, Twilight))	tv	32,036	1,596,534,893	2.5	300	46.9	51.8	58.1
	all	32,036	1,596,534,893	2.5	300	42.6	46.7	70.2

Supplementary Table S20. Population split times with no migration

*T_{min} and T_{max} represent the time grid used for coalescent simulations, which changed given the difference in age between our ancient and modern samples. The mode, 2.5% and 97.5% quantiles are indicated for each posterior distribution (in thousands of years, KY). (tv): only tranversions were considered. (all): all mutation types were considered. *: lower boundary.*



Supplementary Figure S24. Posterior distributions of population split times.

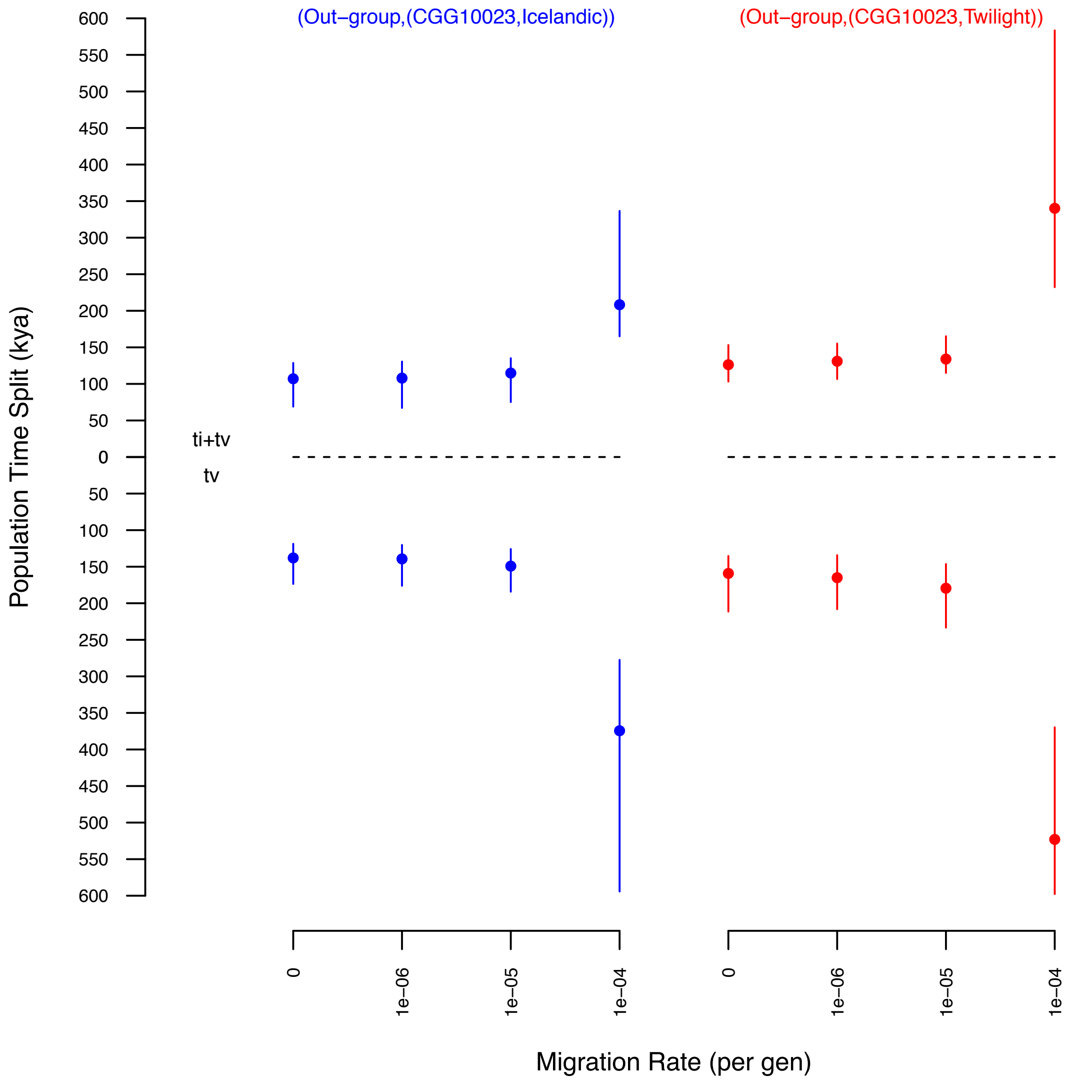
The 3-way alignment considered is provided above each graph. The mode, 2.5% and 97.5%

quantiles are indicated for each posterior distribution (in thousands of years, KY). Top Left panel: Population split time between the ancient and modern horse population, using Twilight as Horse$_2$. Top Right panel: Population split time between the ancient and modern horse population, using Icelandic as Horse2. Bottom Left panel: Population split time between two modern domesticated horses, using either Twilight or Icelandic as Horse$_2$. Bottom Right panel: Population split time between the Przewalski's horse population and modern domesticated horses, using either Twilight or Icelandic as Horse$_2$. (tv): only tranversions were considered. (all): all mutation types were considered.

					Posterior Time (T, KY)		
3-way (Outgroup,(Horse$_1$,Horse$_2$))	Mutation	MigRate	T_{min} (KY)	T_{max} (KY)	2.5%	mode	97.5%
(Outgroup, (CGG10023, Twilight))	tv	10^{-4}	20	600	369.7	523.0	597.7
	all	10^{-4}	20	600	232.3	340.1	583.5
	tv	10^{-5}	20	450	146.4	179.5	233.4
	all	10^{-5}	20	450	115.0	134.0	165.2
	tv	10^{-6}	20	450	134.3	165.0	208.0
	all	10^{-6}	20	450	106.6	131.0	155.1
(Outgroup, (CGG10023, Icelandic))	tv	10^{-4}	20	600	277.5	374.4	594.3
	all	10^{-4}	20	600	165.1	208.3	336.5
	tv	10^{-5}	20	450	125.8	149.2	184.2
	all	10^{-5}	20	450	75.2	114.9	135.17
	tv	10^{-6}	20	450	120.3	139.4	176.1
	all	10^{-6}	20	450	67.1	107.9	130.6

Supplementary Table S21. Population split times with migration

T_{min} and T_{max} represent the time grid used for coalescent simulations. The mode, 2.5% and 97.5% quantiles are indicated for each posterior distribution (in thousands of years, KY). (tv): only tranversions were considered. (all): all mutation types were considered. MigRate = Migration rate (proportion of individuals migrating from the ancient population into the modern population). The number of genomic blocks simulated as well as their total length is indicated in Supplementary Table S20*.*



Supplementary Figure S25. Sensitivity of Population split times to migration rates

We used the F-statistics, serial coalescent simulations and Approximate Bayesian Computation to estimate the population split time between ancient and modern domesticated horses. Simulations were performed assuming no migration, or increasing migration rates (proportion of migrant per generation = 10^{-6}, 10^{-5} and 10^{-4}) from the ancient population into the population of modern domesticated horses. Migration rates were considered constant between 40,000 and 4,000 years ago but zero at any other time. ti+tv = all mutations were considered. tv = transversions only.

(C=CGG10023; A=CGG10022, B)	**f3**	**std**	**Z**
CGG10023; CGG10022, Nordic	0.0115	0.000483	23.8
CGG10023; CGG10022, non-Nordic	0.0114	0.000470	24.1

CGG10023; CGG10022, Icelandic	0.0113	0.000496	22.7
CGG10023; CGG10022, P5782	0.0115	0.000494	23.2
CGG10023; CGG10022, Standardbred	0.0113	0.000510	22.1
CGG10023; CGG10022, Arabian	0.0115	0.000474	24.3
CGG10023; CGG10022, Fjord	0.0118	0.000506	23.3
CGG10023; CGG10022, Przewalski	0.0115	0.000472	24.4
CGG10023; CGG10022, Twilight	0.0113	0.000473	23.9

Supplementary Table S22. f-3 statistic for CGG10023; CGG10022, (non-)Nordic horses

(C=CGG10023; A, B)	**f3**	**std**	**Z**
CGG10023; Nordic, non-Nordic	0.0407	0.000685	59.3
CGG10023; Nordic, Przewalski	0.0241	0.000637	37.9
CGG10023; non-Nordic, Przewalski	0.0332	0.000638	52.1
CGG10023; Icelandic, P5782	0.0504	0.000924	54.6
CGG10023; Icelandic, Standardbred	0.0407	0.000744	54.7
CGG10023; Icelandic, Arabian	0.0411	0.000781	52.6
CGG10023; Icelandic, Fjord	0.0417	0.000710	58.8
CGG10023; Icelandic, Przewalski	0.0331	0.000687	48.2
CGG10023; Icelandic, Twilight	0.0406	0.000753	53.9
CGG10023; P5782, Standardbred	0.0407	0.000784	51.9
CGG10023; P5782, Arabian	0.0404	0.000735	55.0
CGG10023; P5782, Fjord	0.0420	0.000757	55.4
CGG10023; P5782, Przewalski	0.0331	0.000630	52.5
CGG10023; P5782, Twilight	0.0401	0.000752	53.3
CGG10023; Standardbred, Arabian	0.0480	0.000888	54.0
CGG10023; Standardbred, Fjord	0.0410	0.000782	52.4
CGG10023; Standardbred, Przewalski	0.0331	0.000656	50.5
CGG10023; Standardbred, Twilight	0.0512	0.001028	49.8
CGG10023; Arabian, Fjord	0.0405	0.000744	54.4
CGG10023; Arabian, Przewalski	0.0335	0.000679	49.3
CGG10023; Arabian, Twilight	0.0512	0.001011	50.6
CGG10023; Fjord, Przewalski	0.0341	0.000686	49.8
CGG10023; Fjord, Twilight	0.0410	0.000704	58.3
CGG10023; Przewalski, Twilight	0.0332	0.000664	50.0

Supplementary Table S23. f-3 statistic for CGG10022; modern horse, modern horse

(C=CGG10022; A, B)	**f3**	**std**	**Z**
CGG10022; Nordic, non-Nordic	0.0315	0.000683	46.1
CGG10022; Nordic, Przewalski	0.0241	0.000637	37.9
CGG10022; non-Nordic, Przewalski	0.0241	0.000646	37.3
CGG10022; Icelandic, P5782	0.0414	0.000942	43.9
CGG10022; Icelandic, Standardbred	0.0319	0.000737	43.2
CGG10022; Icelandic, Arabian	0.0320	0.000809	39.6
CGG10022; Icelandic, Fjord	0.0324	0.000736	44.0

CGG10022; Icelandic, Przewalski	0.0241	0.000682	35.3
CGG10022; Icelandic, Twilight	0.0317	0.000745	42.6
CGG10022; P5782, Standardbred	0.0317	0.000768	41.2
CGG10022; P5782, Arabian	0.0311	0.000729	42.7
CGG10022; P5782, Fjord	0.0324	0.000720	45.0
CGG10022; P5782, Przewalski	0.0238	0.000635	37.4
CGG10022; P5782, Twilight	0.0309	0.000723	42.8
CGG10022; Standardbred, Arabian	0.0389	0.000920	42.3
CGG10022; Standardbred, Fjord	0.0317	0.000766	41.3
CGG10022; Standardbred, Przewalski	0.0241	0.000654	36.8
CGG10022; Standardbred, Twilight	0.0423	0.001026	41.2
CGG10022; Arabian, Fjord	0.0309	0.000741	41.7
CGG10022; Arabian, Przewalski	0.0242	0.000694	34.8
CGG10022; Arabian, Twilight	0.0426	0.001028	40.9
CGG10022; Fjord, Przewalski	0.0245	0.000679	36.2
CGG10022; Fjord, Twilight	0.0316	0.000712	44.4
CGG10022; Przewalski, Twilight	0.0241	0.000673	35.8

Supplementary Table S24. f-3 statistic for CGG10023; modern horse, modern horse

(H1, H2; H3, Outgroup)	**D**	**Std**	**Z**	**Jackknife**
(CGG10022, CGG10023; Arabian, Outgroup)	-0.0137	0.00455	-3.0	-0.0137
(CGG10022, CGG10023; Fjord, Outgroup)	-0.0137	0.00473	-2.9	-0.0137
(CGG10022, CGG10023; Icelandic, Outgroup)	-0.0095	0.00497	-1.9	-0.0095
(CGG10022, CGG10023; P5782, Outgroup)	-0.0100	0.00463	-2.2	-0.0100
(CGG10022, CGG10023; Przewalski, Outgroup)	-0.0069	0.00455	-1.5	-0.0069
(CGG10022, CGG10023; Standardbred, Outgroup)	-0.0078	0.00481	-1.6	-0.0079

Supplementary Table S25. ABBA-BABA tests between modern and two ancient horses

Delta is nABBA – nBABA; total is nABBA + nBABA; D is the statistic described in section S2.7; Jackknife and Std is the jackknife estimate of D and the associated standard-deviation; Z-scores are significant for values below -3 and values greater than 3. The domestic donkey (Willy) is used as the outgroup.

(C; A, B)	**f3**	**std**	**Z**
non-Nordic; Nordic, CGG10022	0.0146	0.000451	32.4
non-Nordic; Nordic, CGG10023	0.0145	0.000453	31.9
non-Nordic;Przewalski,CGG10022	0.0220	0.000517	42.5
non-Nordic;Przewalski,CGG10023	0.0219	0.000540	40.5
Nordic; non-Nordic, Przewalski	0.0062	0.000313	19.8
Standardbred; Icelandic, CGG10022	0.0289	0.001174	24.7
Standardbred; Icelandic, CGG10023	0.0290	0.001197	24.2
Arabian; Icelandic, CGG10022	0.0345	0.001097	31.4
Arabian; Icelandic, CGG10023	0.0347	0.001140	30.5
Twilight; Icelandic, CGG10022	0.0393	0.001172	33.5
Twilight; Icelandic, CGG10023	0.0394	0.001154	34.1
Standardbred; P5782, CGG10022	0.0291	0.001156	25.2
Standardbred; P5782, CGG10023	0.0289	0.001171	24.7
Arabian; P5782, CGG10022	0.0354	0.001114	31.8

Arabian; P5782, CGG10023	0.0354	0.001130	31.4
Twilight; P5782, CGG10022	0.0401	0.001182	33.9
Twilight; P5782, CGG10023	0.0399	0.001167	34.2
Standardbred; Fjord, CGG10022	0.0292	0.001150	25.4
Standardbred; Fjord, CGG10023	0.0286	0.001166	24.6
Arabian; Fjord, CGG10022	0.0356	0.001109	32.1
Arabian; Fjord, CGG10023	0.0353	0.001123	31.5
Twilight; Fjord, CGG10022	0.0394	0.001152	34.2
Twilight; Fjord, CGG10023	0.0390	0.001120	34.8
Standardbred; Przewalski, CGG10022	0.0368	0.001171	31.4
Standardbred; Przewalski, CGG10023	0.0365	0.001203	30.3
Arabian; Przewalski, CGG10022	0.0423	0.001126	37.6
Arabian; Przewalski, CGG10023	0.0424	0.001188	35.7
Twilight; Przewalski, CGG10022	0.0470	0.001159	40.5
Twilight; Przewalski, CGG10023	0.0468	0.001119	41.8
Icelandic; Standardbred, Przewalski	0.0253	0.000842	30.0
Icelandic; Arabian, Przewalski	0.0253	0.000895	28.2
Icelandic; Twilight, Przewalski	0.0254	0.000917	27.7
P5782; Arabian, Przewalski	0.0093	0.000605	15.4
P5782; Standardbred, Przewalski	0.0086	0.000533	16.2
P5782; Twilight, Przewalski	0.0093	0.000589	15.9
Fjord; Arabian, Przewalski	0.0130	0.000854	15.2
Fjord; Standardbred, Przewalski	0.0122	0.000837	14.5
Fjord; Twilight, Przewalski	0.0122	0.000823	14.8

Supplementary Table S26. f-3 statistic testing for admixture for (non-)Nordic

S2.9 Joint demographic inference using *∂a∂i*

We refined the joint demographic history of horses using ∂a∂i (version 1.6.3), a program which estimates demographic parameters based on the diffusion approximation to the site frequency spectrum (SFS) (93). We used the 3-dimensions SFS (3D-SFS) using the Domestic horse (DOM), CGG10022 (ANC) and Przewalski's horse (PRZ) samples. We did not include CGG10023 due to its lower sequencing coverage and we did not take into account SNPs leading to transitions, as the latter is more likely to be due to DNA damage in ancient samples. We folded the spectrum to avoid biases when assessing the ancestral allelic states.

We adopted the demographic model used in simulations using fastsimcoal2 (see Section S2.8). We allowed only instantaneous population size changes, except for the most recent expansion of domesticated horses where we modeled an exponential growth (Supplementary Figure S26).

We ran the program 20 times with varying starting points to ensure convergence, and retained the fitting with the highest likelihood. Gene flow from ANC to DOM/PRZ was modeled as a continuous migration event from the time of the split to the ANC collecting time (43 kya). After the split between DOM and PRZ, the same rate of gene flow was imposed from ANC solely to DOM, as supported by previous analyses (see Supplementary Section S2.7).

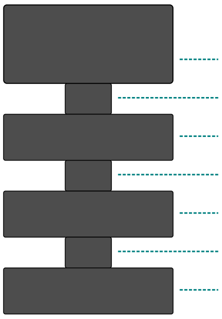
We calibrated our estimated model parameters into years and effective population sizes using estimates previously obtained from fastsimcoal2 analyses. We let 4 parameters free to be estimated: the split time between ANC and DOM/PRZ, the migration rate from ANC to DOM/PRZ, the split time between DOM and PRZ, and the current effective population size of ANC and PRZ.

Supplementary Table S27 shows the most likely values of estimated parameters. The resulting population split time at around ~169k year BP support the results obtained from the F-Statistics (Supplementary Section S2.8), and closely matches estimates obtained when assuming a migration rate of 10^{-5}. Observed and modeled spectra are presented in Supplementary Figure S27.

We assessed the statistical significance of the model improvement observed following the addition of migration event using a Likelihood Ratio Test (LRT) with a block re-sampling approach. For this sole purpose, we generated a SFS by randomly sampling one unique single site for each non-overlapping window of 25 kb across the entire genome. This procedure ensured that a sufficiently large number of sites was considered and that the observations were approximately independent.

A model was fitted assuming no migration between ANC and DOM/PRZ and assuming migration from ANC to DOM/PRZ, following the same procedure as above where the fit was computed several times and the best likelihood was recorded. P-value was computed by doubling the difference in log-likelihood between models and assuming a χ^2 distribution with 1 degree of freedom. To assess the robustness in the resulting p-value, 10 different sampling procedures were used and the highest (and therefore most conservative) value was recorded.

We find statistical support for the model including migration from ANC to DOM/PRZ, with the highest p-value among all 10 replicates being equal to 3.23 × 10^{-8}.



Years BP	DOM	PRZ	ANC
1100k	-	110k	-
550k	-	20k	-
230k	-	200k	-
110k	90k	90k	20k
70k	180k	180k	150k
10k	3k	2k	2k
0	300k	2k	2k

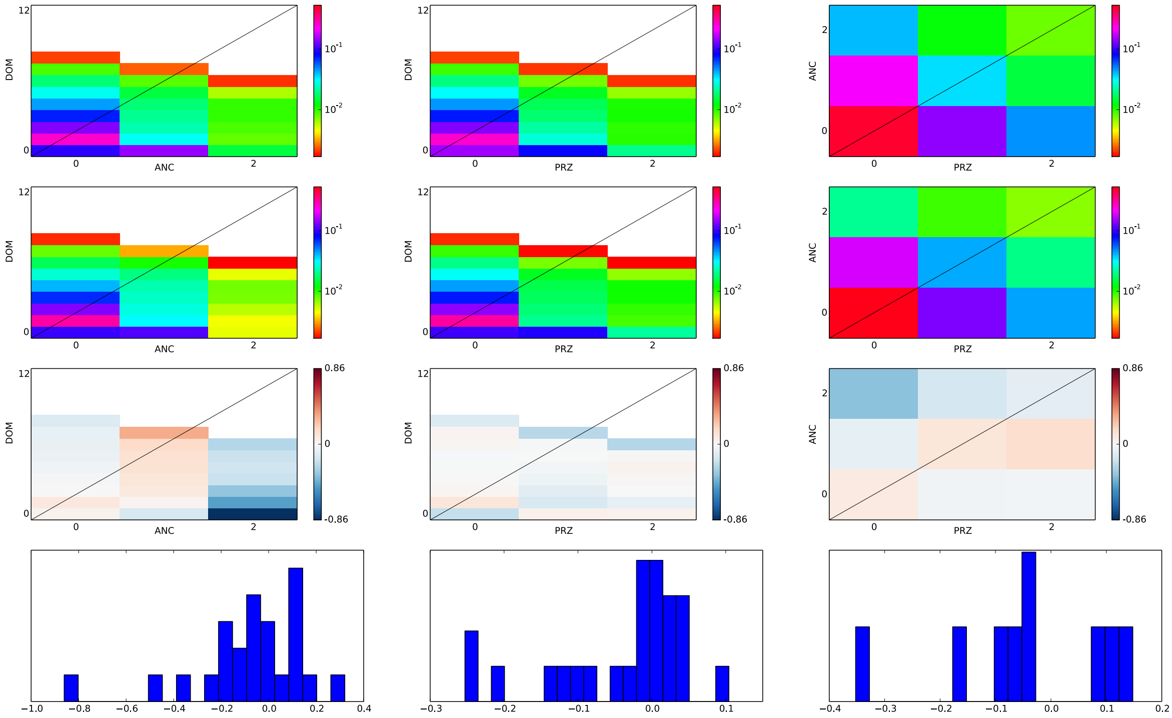
Supplementary Figure S26. Population models for the ∂a∂i analysis

Demographic shifts have been modelled as instantaneous for tractability reasons; for the first three shifts, the population sizes are identical for the three populations, as these have not yet split. Gene flow was allowed between the ANC and DOM/PRZ populations from the initial population split between ANC and DOM/PRZ, and between ANC and DOM after the DOM/PRZ split, but not between ANC and PRZ in agreement with the population model supported by admixture tests. The expansion in the DOM population from 10k BP to now was modelled as an exponential expansion.

Parameter	**Estimated value**

Split time between ANC and DOM/PRZ	169,351 years BP
Migration rate from ANC to DOM/PRZ (fraction made up of new migrants each generation)	5.94×10^{-5}
Split time between DOM and PRZ	46,200 years BP
Current effective population size for ANC and PRZ	1,973

Supplementary Table S27. The ∂a∂i maximum-likelihood estimates for the horses demographic inferences



Supplementary Figure S27. Observed and modeled spectra from ∂a∂i analyses

Pairwise Site Frequency Spectra (SFS) of horses from our observed data (upper panel) and from our model estimated using ∂a∂i (second panel from the top). Residuals (third and fourth panel from the top) show the distance between the data and the model. The first column from left represents the comparison between ANC and DOM, the second between DOM and PRZ, the third between ANC and PRZ. We show 2D-SFS instead of 3D-SFS for ease of visual inspection of the fitting.

S2.10 Demographic reconstruction based on the nuclear genome information

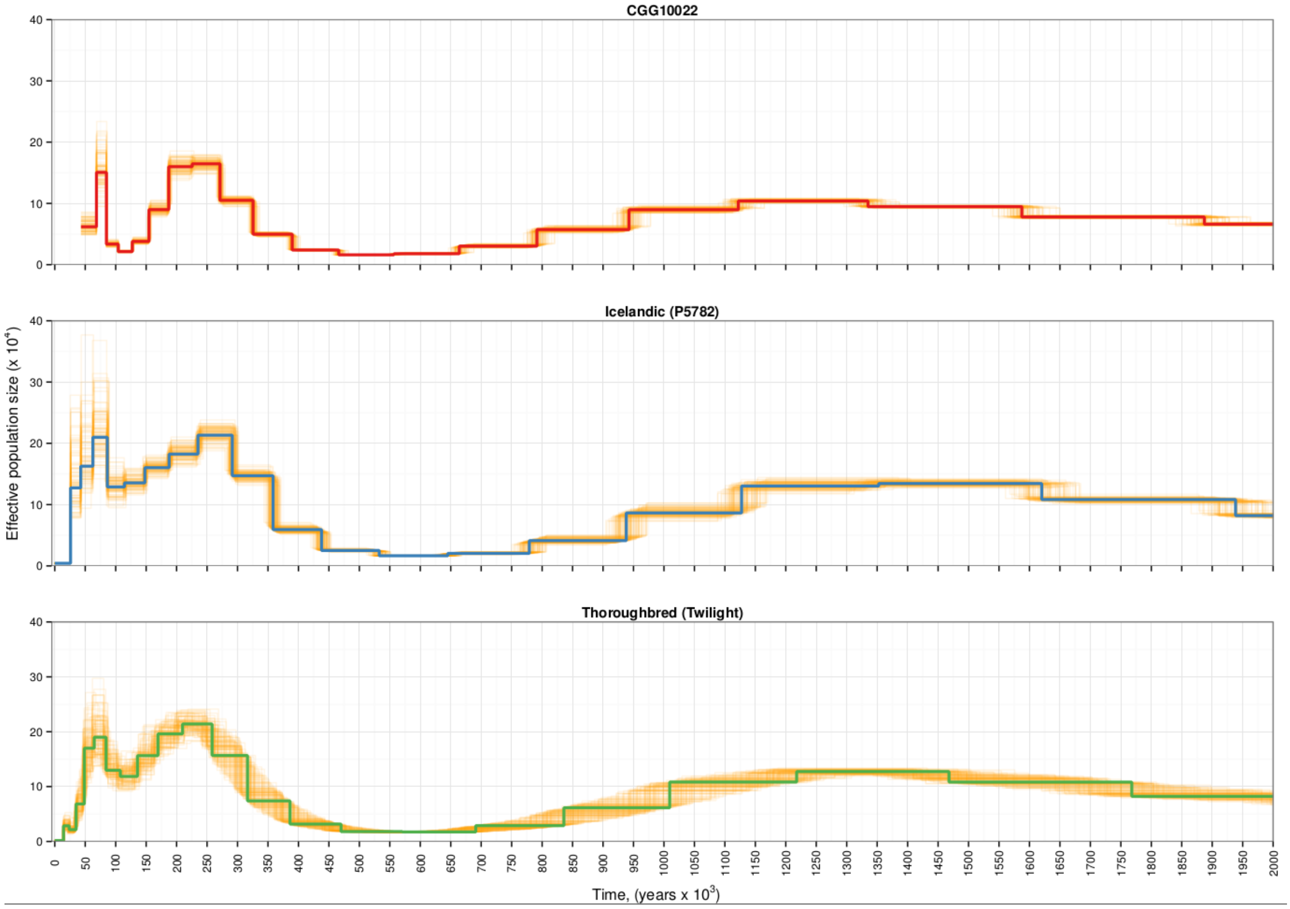
The depth-of-coverage obtained for CGG10022 (24.27X, Supplementary Table S4) is compatible with the inference of past demographic population sizes using Pairwise Sequentially Markovian Coalescent (PSMC), which requires at least 20X coverage (94). For comparison, the same procedure was applied to the previously published modern horse genomes for which the depth-of-coverage was sufficient, namely the Thoroughbred (Twilight) (1) and the Icelandic (P5782) (2).

The demographic inferences were performed as described earlier [Orlando *et al.* 2013, see their Supplementary Information section S9.2](1) with slight modifications, namely per-sample maximum coverage threshold. Briefly, a consensus sequence was generated for the autosomes (excluding chrUn) using the BAM files for alignments against the horse reference genome EquCab2.0 (3). SNPs were called using SAMtools (15) v0.1.18 and filtered using the '*vcfutils.pl vcf2fq*' command with the following parameters:

- Minimum depth-of-coverage: 8
- Maximum depth-of-coverage: Supplementary Table S13
- Indels filtered in a windows size of 5 bp
- Minimum RMS mapping quality: 10

The coverage distribution was determined using the *'depths*' tool included with the PALEOMIX pipeline (10), and excludes sites with depth 0. In addition, bases with Phred quality scores inferior to 35 were filtered.

The PSMC inference was run using input parameters recommended by the developers (Number of iterations = 25; maximum $2N_0$ coalescent time = 15; initial θ/ρ = 5). 100 bootstrap pseudo-replications were performed by splitting chromosomal sequences into shorter fragments of 500 kb and randomly selected among these (with replacement) in order to evaluate the spread of the PSMC reconstructions. For scaling the demographic inferences, we took advantage of the recent recalibration time of all *Equus* species, ranging between 4 to 4.5 Myr (Orlando *et al.* 2013). For a calibration point at 4.5 Myr, the resulting mutation rate used for scaling all PSMC was 7.242×10^{-9} per site per generation. The generation time was set to 8 years (Supplementary Figure S28).



Supplementary Figure S28. Demographic fluctuations over the last 2 millions years

*The 100 bootstrap pseudo-replications are depicted in yellow for each of the three specimens. The demographic inference and the associated bootstrap pseudo-replications for the pre-domesticated horse (CGG10022) are plotted starting from the sample age, namely 43 kyr BP (*Supplementary Table S1*).*

S3 Genetic load and inbreeding

S3.1 Genetic load in the genomes of modern and pre-domesticated samples

S3.1.1 Comparison between modern and pre-domesticated samples

We restricted the analyses described here to the high-coverage genome sequence of specimen CGG10022 and disregarded the sequence data from specimen CGG10023 that showed lower depth-of-coverage and higher error rates (Supplementary Section S1.4 and Supplementary Figure S3). Similarly, we did not measure the genetic load for Twilight, as this specimen corresponds to the individual that was originally sequenced to generate the horse reference sequence EquCab2.0 and is consequently expected to show a deficit in derived alleles, with respect to the reference (3).

We used Genomic Evolutionary Rate Profiling (GERP) scores to quantify the level of evolutionary constraint at each polymorphic site. GERP scores computed from the alignment of 35 mammals to the human genome reference hg19 were downloaded from the UCSC platform (95). GERP scores are defined as the number of substitutions expected under neutrality minus the number of substitutions observed at that position (96). Positive scores, larger than 2, represent a substitution deficit, which is expected for sites under selective constraints; scores smaller than 2, including negative values, indicate that a site is probably evolving neutrally (97). New mutations at constrained sites are expected to be deleterious. We converted the EquCab2.0 genomic coordinates of the horse polymorphic sites into the hg19 coordinates with the liftover tool (95). GERP scores at each site were then extracted using the hg19 coordinates. We validated GERP score calculations by checking that the class of sites associated with GERP scores greater than or equal to 2 (as opposed to the class of sites with GERP scores between -2 and 2) was enriched for exonic and non-synonymous sites (Supplementary Figure S29). Functional classifications into exons and non-synonymous sites were obtained with ANNOVAR applied to the Ensembl horse transcripts (98).

We restricted our analyses to genomic regions corresponding to coding DNA sequences (CDS) from Ensembl v72. The 10 bp upstream and downstream of each exon were also considered to enable calling SNP following the same quality filters as described in Supplemental Section S2.5. For each polymorphic site observed in a given horse sample, we defined the ancestral state using the donkey sequence data. We then computed a measure of genetic load as the product of the GERP score at each site and the number of derived alleles carried by this individual at this site (Supplementary Figure S30), averaged across sites for each individual. We considered only sites with GERP scores ≥ -2, since only those sites are considered as under selection.

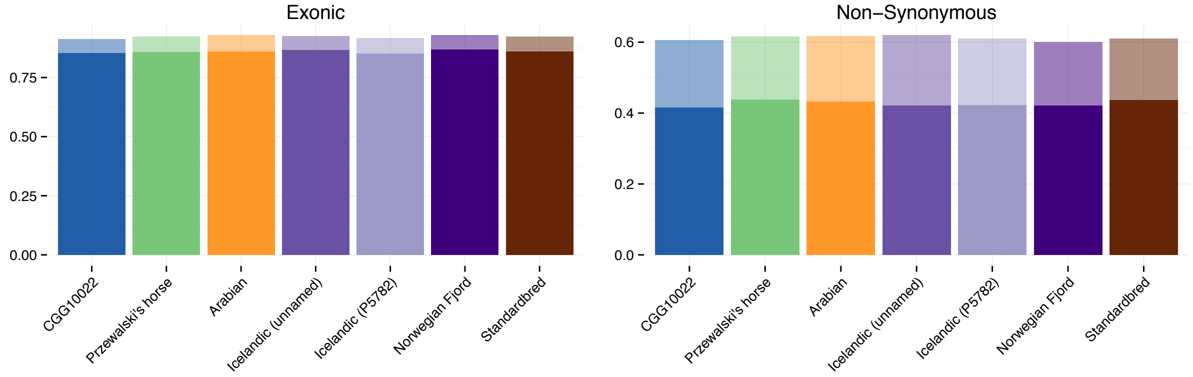
We plotted the distributions of genetic loads (Supplementary Figure S31), and performed QQ plots (Supplementary Figure S32) and Kolmogorov-Smirnov tests to compare the distribution of genetic load measures among individuals. The statistical significance of each of those tests is presented in Supplementary Table S28. The full set of analyses was repeated disregarding sites with transitions to limit the bias introduced by nucleotide misincorporations at positions affected by *post-mortem* DNA damage (Supplementary Table S29 and Supplementary Figure S33).

That the high-quality ancient horse genome and the genomes from domesticated horses show different genetic load measures is in line with the 'Cost of Domestication' hypothesis (99-101), which supposes that the repeated bottleneck events associated with domestication have changed the patterns of molecular evolution by limiting the strength of purifying selection. As a result, the genomes of domesticated animals, such as dogs (99), and plants, such as rice (101, 102) and tomatoes (103), but not fungi (104), show an excess of (slightly) deleterious mutations. The difference observed in genetic load values could also result from the higher levels of inbreeding measured in domesticated horses (Figure 2; Supplemental Section S3.2), as homozygous sites contribute to a greater extent than heterozygous sites in the calculation of the genetic load. This is because GERP scores are counted twice at homozygous sites, while at heterozygous sites, GERP scores at counted only once.

In order to control for the various levels of inbreeding found across individuals, we stratified our analyses across homozygous and heterozygous sites and tested for each category whether the high-quality ancient genome showed significantly fewer mutations at sites under selection than the genomes of domesticated horses and the Przewalski's horse (Supplementary Figure S34). To accomplish this, we first stratified heterozygous and homozygous sites for each individual. We then considered sites with GERP scores in the range between -2 and 2, representing sites under no or very weak selective constraints, and sites with GERP scores greater than or equal to 4, representing sites under strong selective constraint. We next counted the number of sites containing mutations in each of the two classes of sites and calculated the ratio between numbers of mutations for the domestic horses in aggregate, and for the Przewalski's horse by itself. Ratios were calculated separately for sites containing homozygous or heterozygous derived mutations.

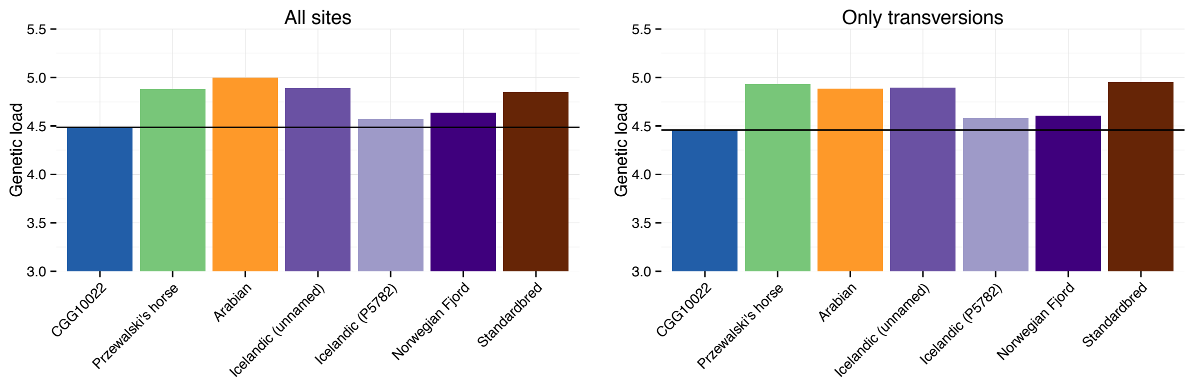
We next calculated the odds-ratio between these ratios and the ratios observed for CGG10022 and performed one-sided ("greater") Fisher exact-tests using R. Following correction for multiple tests using Holm-correction, we found that sites under strong selective constraint in the genomes of the domestic horses were enriched for homozygous mutations relative to CGG10022 (p_{hom} = 0.033; p_{het} = 0.072). We furthermore found that the genome of the Przewalski's horse was enriched for both homozygous and heterozygous mutations at sites under strongly selective constraint (p_{hom} = 0.033; p_{het} = 0.041).

Overall, our results support the presence of an excess of mutations at sites under selection amongst domestic horses and therefore the 'Cost of domestication' hypothesis. Interestingly, a similar excess was found in the genome of the Przewalski's horse, which has never been successfully domesticated, most likely as a result of the recent massive bottleneck experienced following the foundation of the captive stock from only 13 individuals (105).

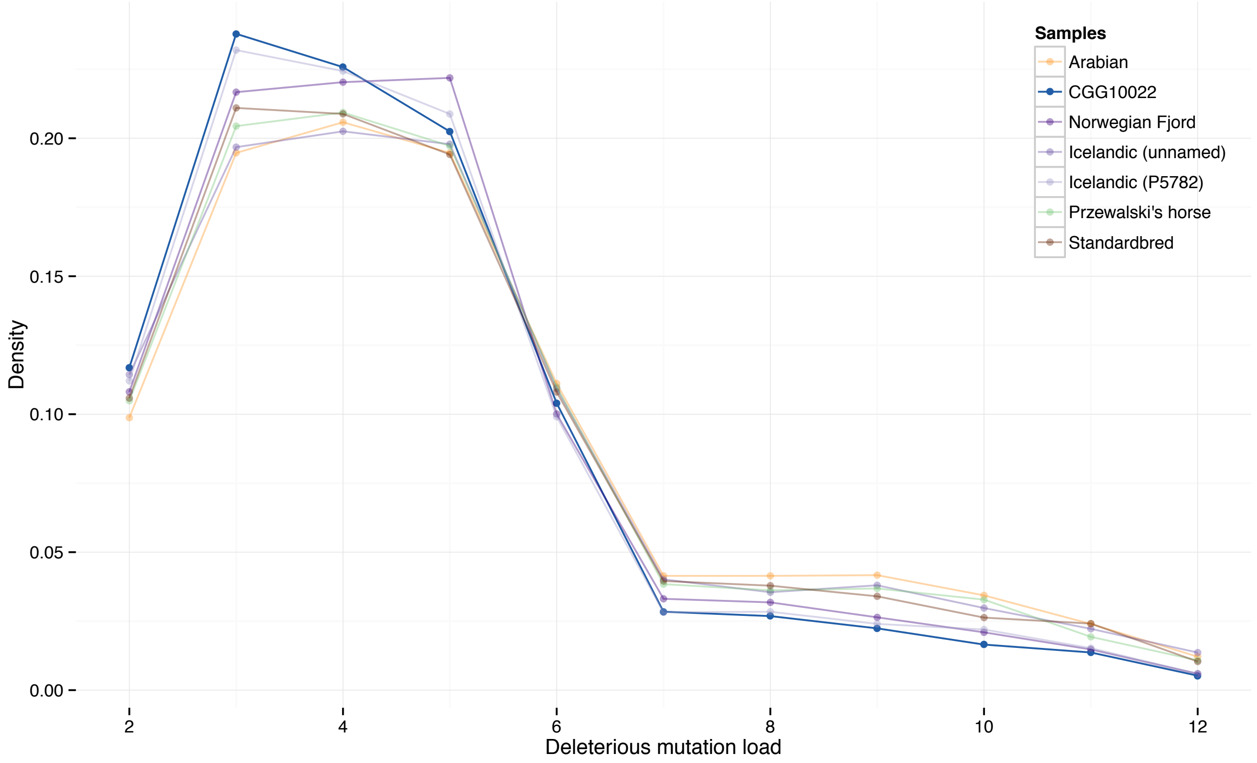


Supplementary Figure S29 Mutations stratified by GERP scores and annotation

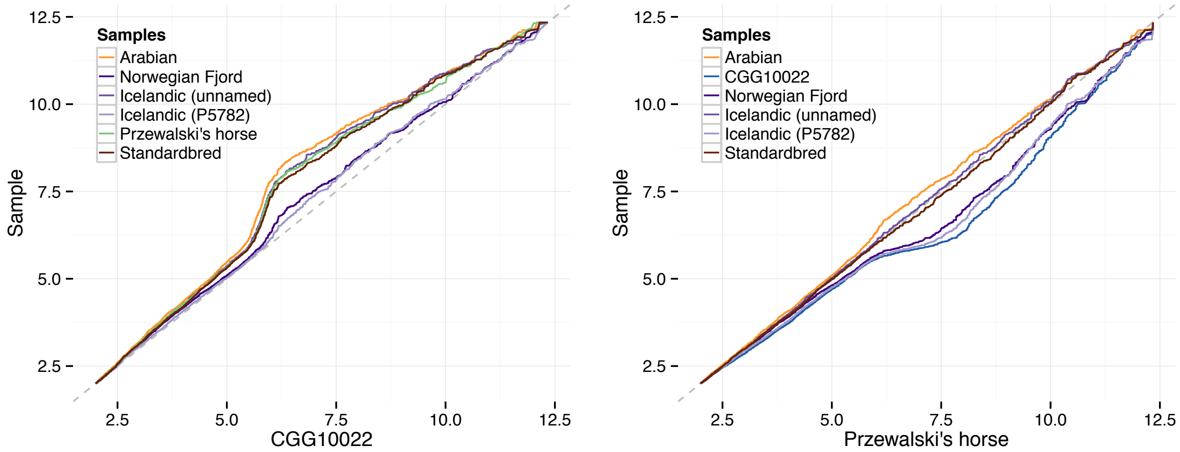
Plot showing the fraction of sites with a given annotation (exonic or non-synonymous), stratified by GERP score. The dark bars show the fraction of sites for GERP scores in the range [-2; 2[, superimposed on the lighter bars showing the fraction of sites for GERP scores in the range [2;[.



Supplementary Figure S30 Genetic load calculated for sites with / without transitions



Supplementary Figure S31 Distribution of genetic loads

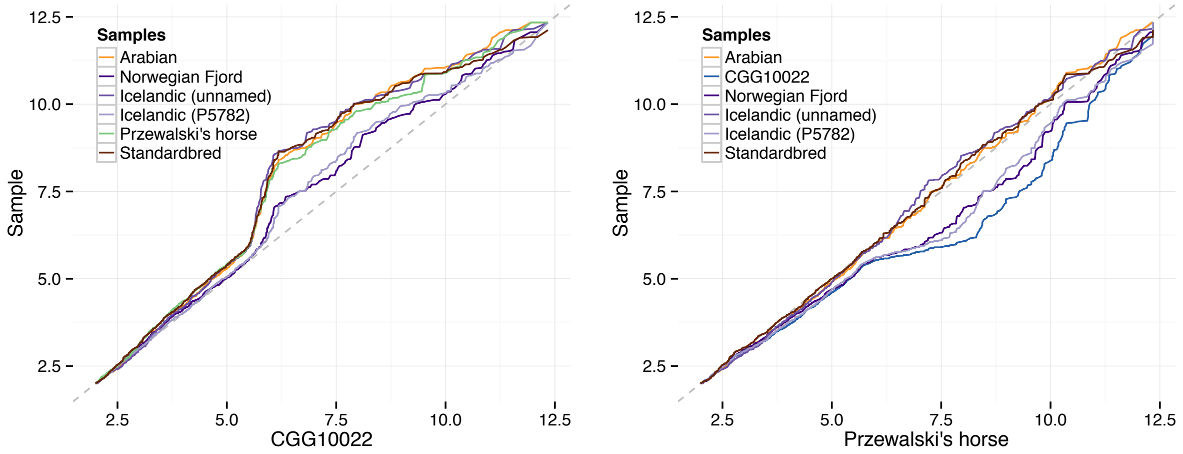


Supplementary Figure S32 QQ-plots of genetic load vs CGG10022 / Przewalski's horse (all sites)

Vs	Sample	*p*-value	*q*-value
CGG10022	Arabian	0.00×10^{0}	**0.00×10^{0}**
CGG10022	Norwegian Fjord	5.28×10^{-4}	**2.64×10^{-3}**
CGG10022	Icelandic (unnamed)	1.24×10^{-9}	**9.94×10^{-9}**
CGG10022	Icelandic (P5782)	2.25×10^{-1}	6.74×10^{-1}
CGG10022	Przewalski's horse	2.86×10^{-11}	**3.15×10^{-10}**
CGG10022	Standardbred	1.36×10^{-10}	**1.22×10^{-9}**
Przewalski's horse	Arabian	8.26×10^{-2}	3.30×10^{-1}
Przewalski's horse	CGG10022	2.86×10^{-11}	**3.15×10^{-10}**
Przewalski's horse	Norwegian Fjord	2.36×10^{-5}	**1.41×10^{-4}**
Przewalski's horse	Icelandic (unnamed)	9.30×10^{-1}	9.30×10^{-1}
Przewalski's horse	Icelandic (P5782)	4.95×10^{-9}	**3.47×10^{-8}**
Przewalski's horse	Standardbred	3.90×10^{-1}	7.81×10^{-1}

Supplementary Table S28 Kolmogorov-Smirnov test of genetic loads (all sites)

Significant tests (after correction for multiple tests) are shown in bold.

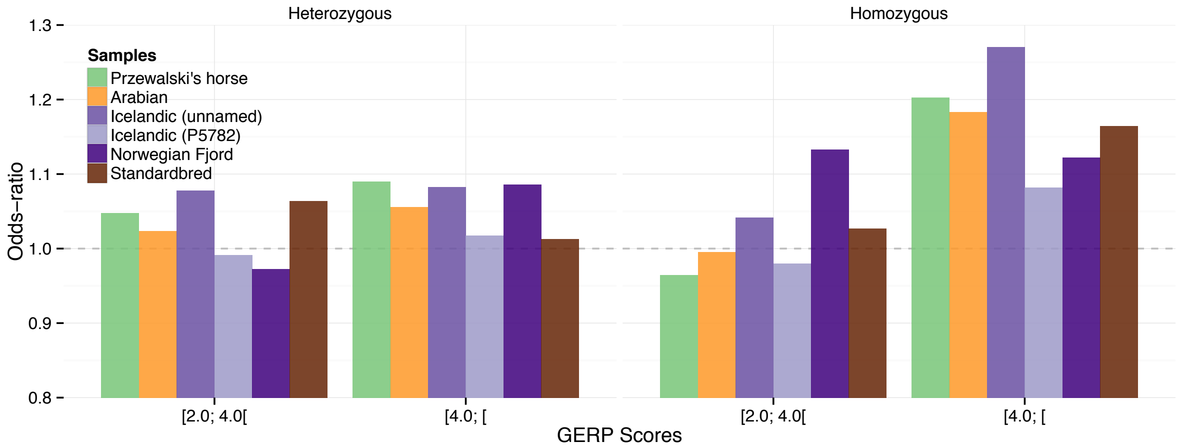


Supplementary Figure S33 QQ-plots of genetic load scores vs CGG10022 / Przewalski's horse (tv only)

Vs	Sample	p-value	q-value
CGG10022	Arabian	7.09×10^{-5}	**6.38×10^{-4}**
CGG10022	Norwegian Fjord	2.85×10^{-1}	1.00×10^{0}
CGG10022	Icelandic (unnamed)	1.65×10^{-3}	**9.87×10^{-3}**
CGG10022	Icelandic (P5782)	5.07×10^{-1}	$1.00\times10^{+0}$
CGG10022	Przewalski's horse	9.97×10^{-6}	**1.10×10^{-4}**
CGG10022	Standardbred	5.76×10^{-6}	**6.91×10^{-5}**
Przewalski's horse	Arabian	5.46×10^{-1}	1.00×10^{0}
Przewalski's horse	CGG10022	9.97×10^{-6}	**1.10×10^{-4}**
Przewalski's horse	Norwegian Fjord	8.10×10^{-4}	**5.67×10^{-3}**
Przewalski's horse	Icelandic (unnamed)	3.41×10^{-1}	1.00×10^{0}
Przewalski's horse	Icelandic (P5782)	1.65×10^{-4}	**1.32×10^{-3}**
Przewalski's horse	Standardbred	8.70×10^{-1}	1.00×10^{0}

Supplementary Table S29 Kolmogorov-Smirnov test of genetic loads (transversions only)

Significant tests (after correction for multiple tests) are shown in bold.

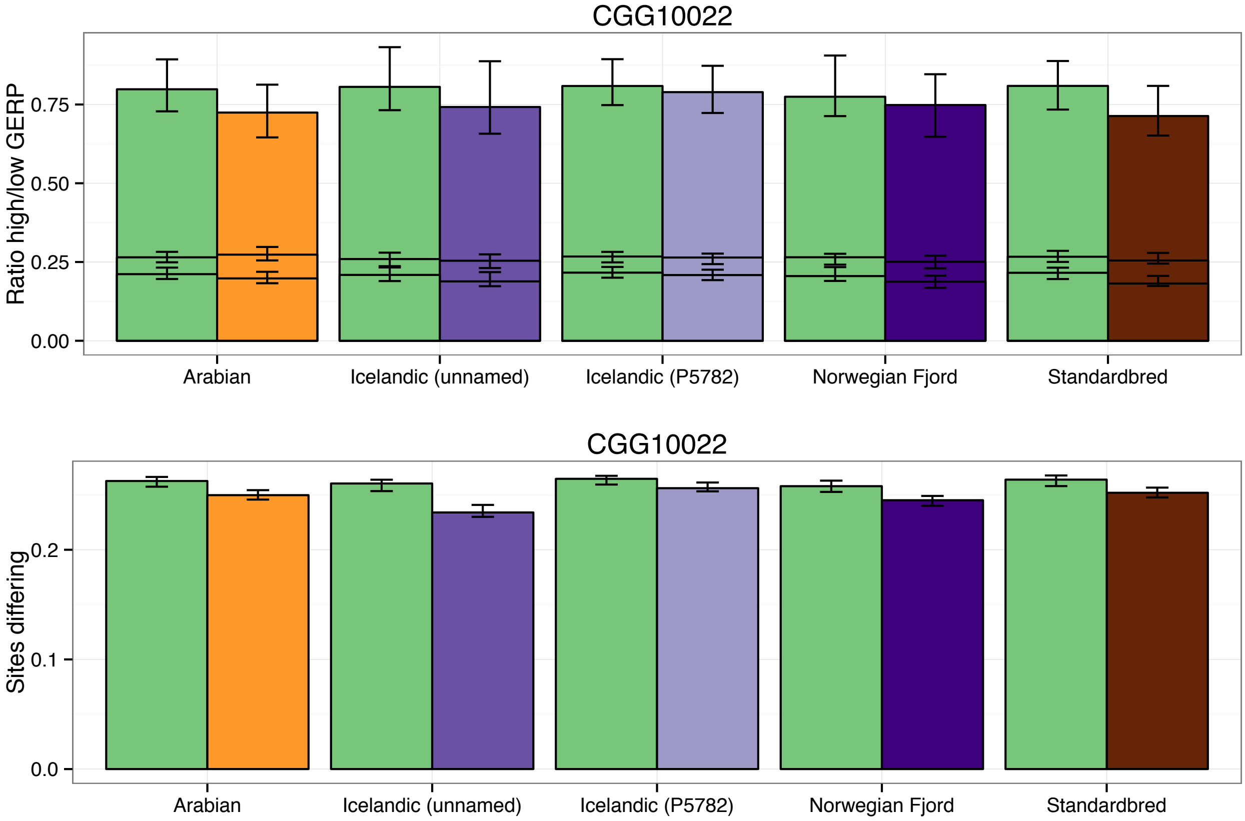


Supplementary Figure S34 Odds-ratios vs CGG10022 for GERP scores across homo/heterozygous sites

S3.1.2 Comparison of the Przewalski's horse with domestic breeds

To determine if the deleterious mutation load observed for the Przewalski's horse differed from that observed for the domesticated horses, sites in coding regions containing SNPs in at least one modern breed were stratified according to the GERP scores, and the number of differences calculated for trios of horses. Each trio included the high-quality pre-domesticated sample (CGG10022), the Przewalski's horse, and one domestic horse. Differences were calculated by pairwise comparisons between the pre-domesticated horse and the two modern horses (Przewalski's and domestic), counting any site that did not match the pre-domestic horse as 1 difference. For a given trio, only sites for which all 3 samples were called were included in the comparison. The Thoroughbred (Twilight) was excluded from the comparison.

The fractions of sites differing from the pre-domesticated horse were calculated for each modern horse in a trio, stratified by the two classes of GERP scores described in Supplementary Section S3.1.1 (namely scores [-2;2[and scores [4;[). More specifically, these fractions were calculated as the number of sites showing differences (as defined above) to the pre-domestic horse divided by the total number of sites in the class, for a given trio. The odds-ratios were calculated as the ratio of these two fractions. 100 bootstrap pseudo-replicates were performed by sampling with replacement among the set of sites which were included in at least one comparison, and estimating the ratio of fractions as described above. The overlapping intervals are consistent with the hypothesis that rates are similar/same between the modern horses and the Przewalski's horse (Supplementary Figure S35).



Supplementary Figure S35. GERP scores for CGG10022

The top plot shows the ratios of sites stratified according to the top/bottom 5% of GERP scores. The two smaller error bars indicate the rates observed for each of these classes respectively. The bottom plot shows the overall rate of differences across all sites. Error bars were estimated using 100 bootstrap pseudo-replicates. Each pair of bars represent a comparison with the Przewalski's horse (green).

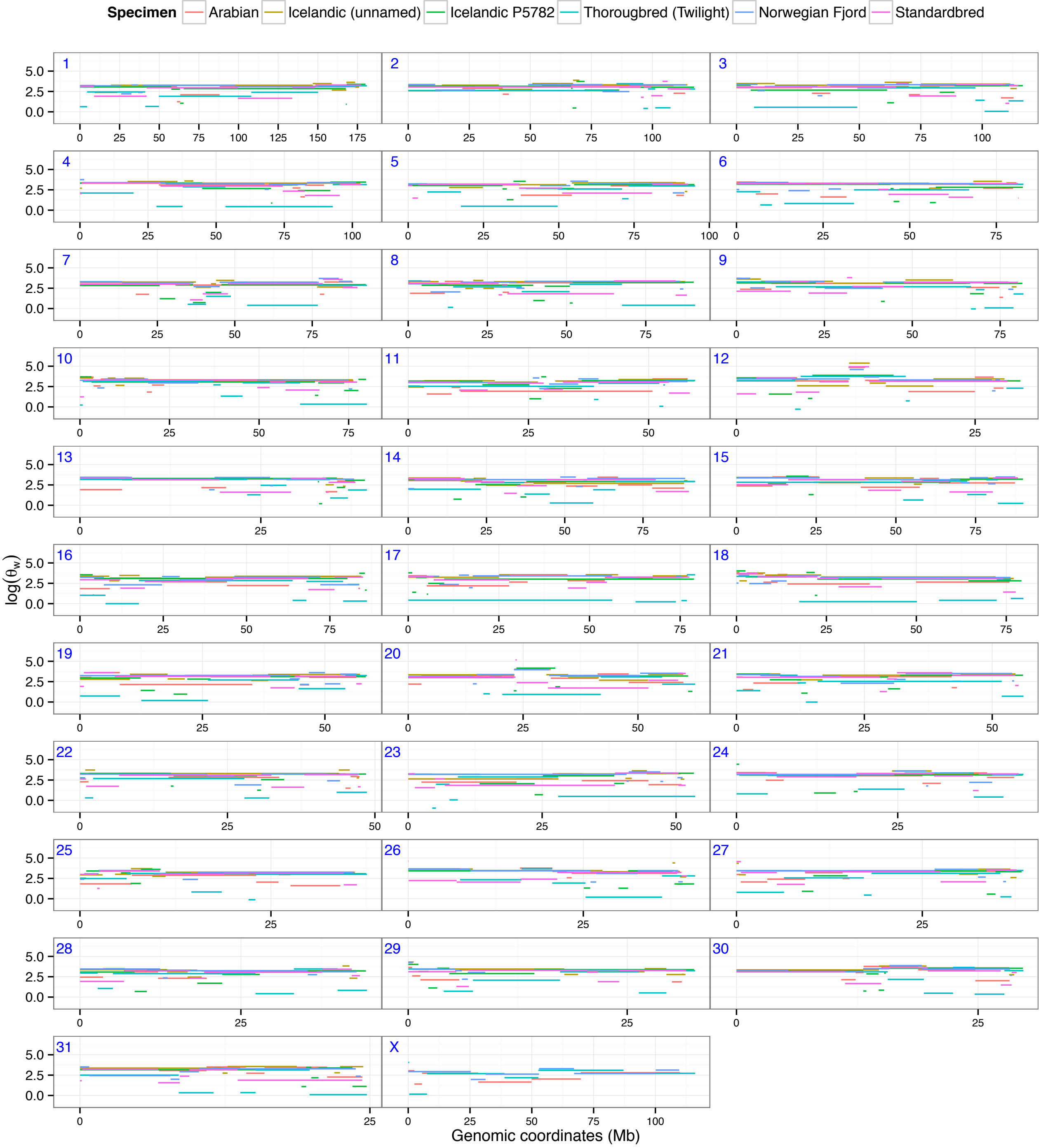
S3.2 Estimation of inbreeding in domestic and wild horses

Inbreeding in the domestic and wild horses was estimated using a methodology similar to that detailed in Prüfer *et al*. 2013 [see their Supplementary Information section S10] (106), in which the proportion of mostly homozygous genomic segments is calculated. These regions are termed homozygous-by-descent (HBD) in the following section, in line with Prüfer *et al.* (106).

To detect HBD regions, heterozygosity was estimated across the samples by calculating the $\hat{\theta}_w$ for sliding windows of 50 kb with a 10 kb step size (Supplementary Section S1.6), excluding regions where less than 90% of bases (45 kb) were covered, and excluding transitions to account for the presence of *post-mortem* DNA damage in CGG10022 and CGG10023 (plots with transitions available upon request). Segments with local changes in hat(theta) values were estimated using the R package 'changepoints' (http://cran.r-project.org/web/packages/changepoint/index.html) using the binary segmentation algorithm. Segment coordinates and corresponding heterozygosity estimates (mean log($\hat{\theta}_w$)) were extracted (Supplementary Figure S36 and Supplementary Figure S37), and segments were plotted according to the segment length and heterozygosity estimate for each sample (Supplementary Figure S38). We excluded chromosome X from male individuals.

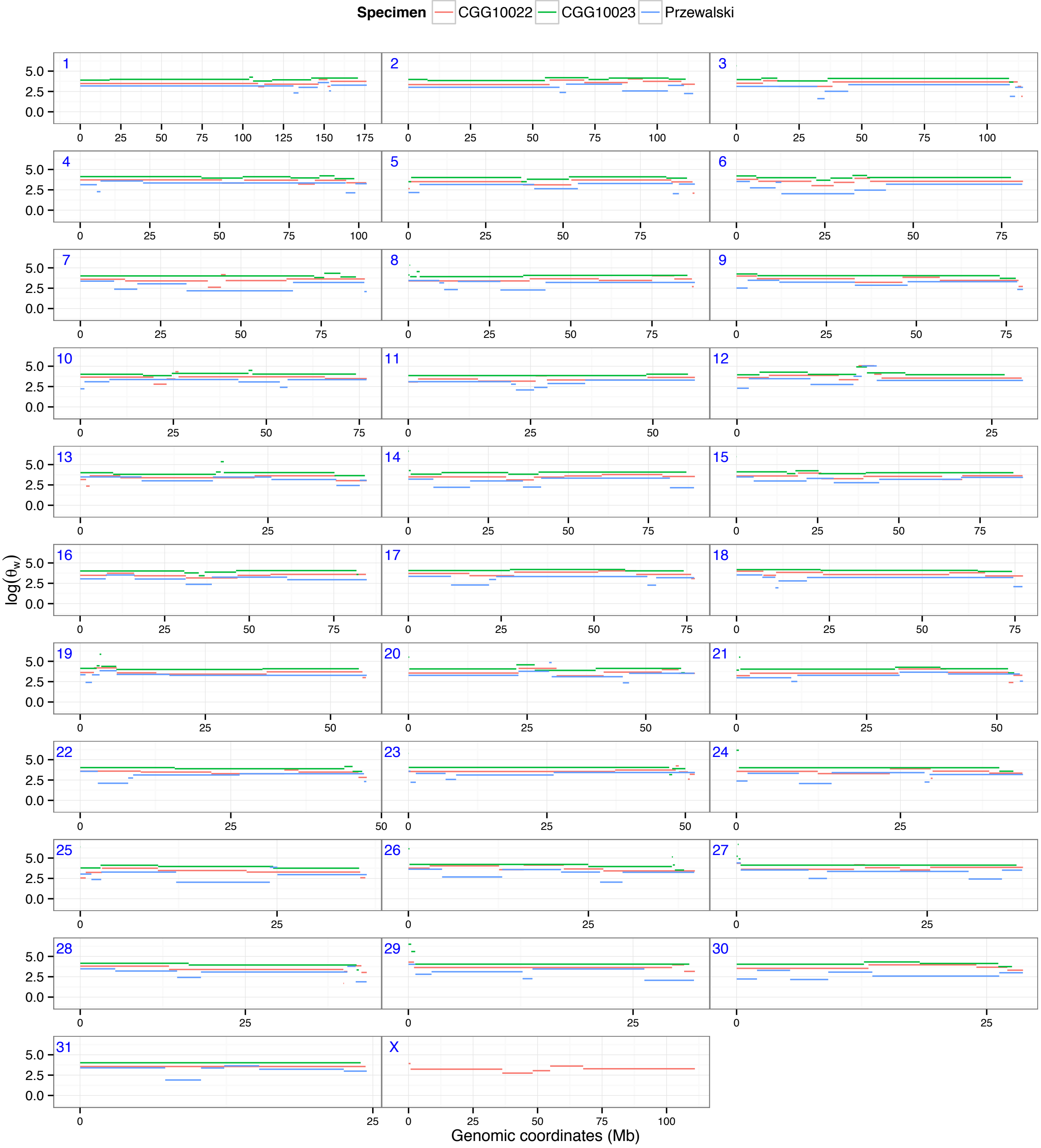
A bimodal distribution indicates inbreeding in the corresponding individual. The R package 'turnpoints' (http://www.sciviews.org/pastecs/) was used to determine the coordinate of the lowest point (pit) between the two modes for bimodal distributions, indicated by a vertical red line on the plots (Supplementary Figure S38), providing a threshold used to group genomic segments into the categories 'high' and 'low' heterozygosity. The proportion of HBD tracks was subsequently calculated as the total size of 'low' heterozygosity regions divided by the total size of all regions (value shown adjacent to the line indicating the turnpoint).

The results suggest that the Arabian, Standardbred, and Thoroughbred (Twilight) are the most inbred individuals, although all modern horses appear to be inbred to some extent (4.1 - 29.8% regions HBD). This is expected for the Thoroughbred (Twilight) as this individual was selected for the *Equus caballus* reference genome because of its high level of inbreeding (3). Lower levels of inbreeding were observed for the Icelandic horses, the Norwegian Fjord, and the Przewalski's horse (4.1 - 13.8% regions HDB). The latter is encouraging in light of efforts to preserve the Przewalski's horse as representing the last remaining truly wild horses (1). Low to non-existent levels of inbreeding were detected for the two pre-domesticated specimens, in line with the high amount of genetic variability known to predate domestication (107), while the current population of Przewalski's horses derive from a small population of captured individuals, of which only 12 individuals contributed to the current population (108).



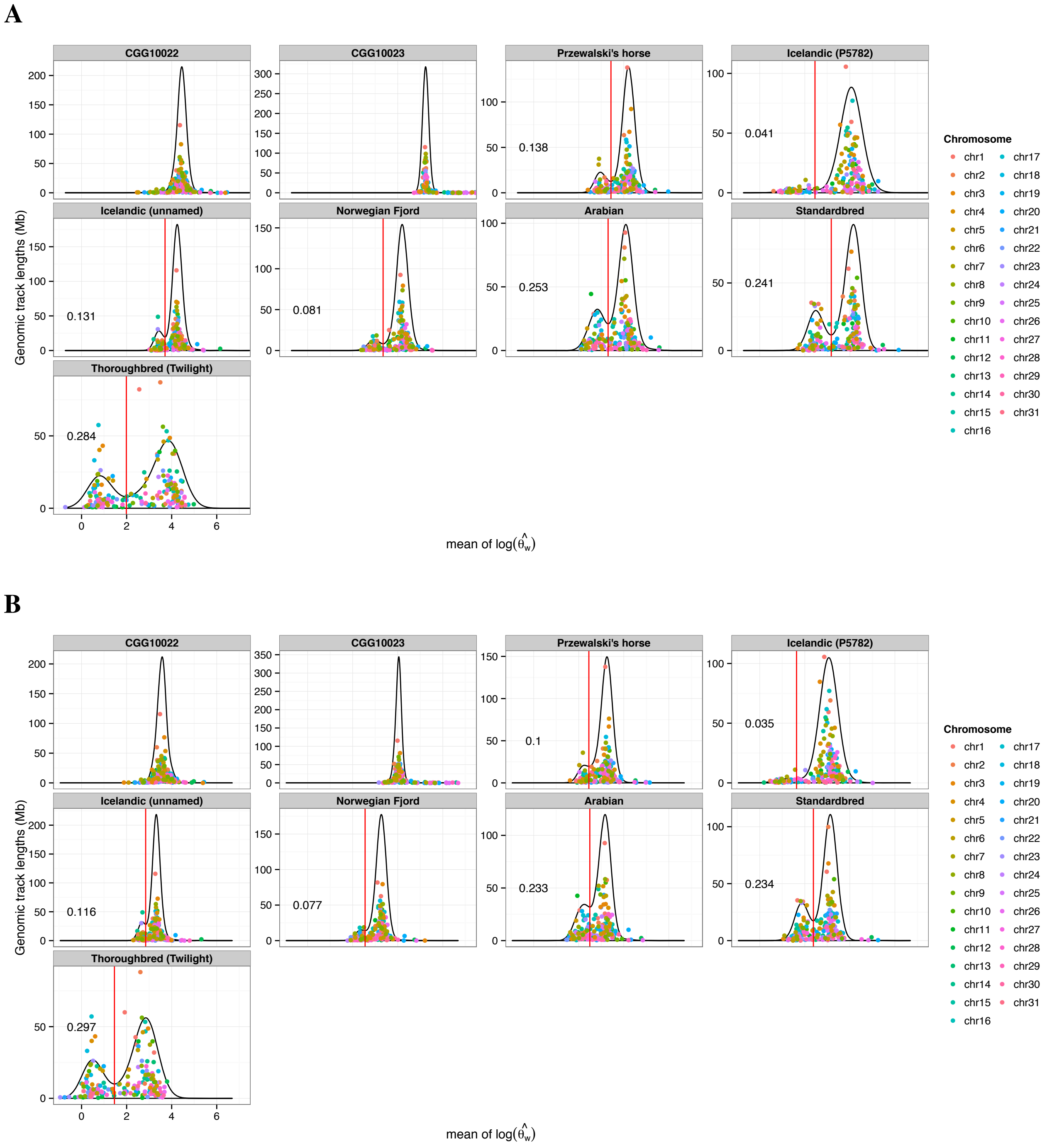
Supplementary Figure S36. Heterozygosity of genomic regions for domestic horses

Blue numbers in the upper-left corner of plots indicate the chromosome. Heterozygosity was estimated across the samples by calculating the $\hat{\theta}_w$ for sliding windows of 50 kb with a 10 kb step size, excluding regions where less than 90% of bases (45 kb) were covered. Transitions were excluded to allow comparison with pre-domesticated samples.



Supplementary Figure S37. Heterozygosity of genomic regions for wild horses

Blue numbers in the upper-left corner of plots indicate the chromosome. Heterozygosity was estimated across the samples by calculating the $\hat{\theta}_w$ for sliding windows of 50 kb with a 10 kb step size, excluding regions where less than 90% of bases (45 kb) were covered. Transitions were excluded to allow comparison with pre-domesticated samples.



Supplementary Figure S38. Inbreeding coverage estimates for domestic and wild horses

(A) Inbreeding estimates calculated using both transitions and transversions; (B) Inbreeding estimtates calculated using only transversions. The average level of inbreeding coverage is indicated for bi-modal distributions. Unimodal distributions suggested low levels of inbreeding, if any. In such case, the average level of inbreeding is left undetermined.

S4 Selection scans

S4.1 Screening for positive selection using PAML

Transcripts for which at least 80% of sites were called (across all samples) were selected from the 20,384 transcripts genotyped for the phylogenetic inference (Supplementary Section S2.5), yielding 9,663 candidates. PAML (109) 'codeml' was subsequently used to estimate the number of synonymous and non-synonymous substitutions at each branch, using the topology inferred as described in the section "Phylogenetic Inference", while allowing a different ω for each branch (model = 1), with no clock, F3X4 codon frequencies, no dN/dS variation among sites, κ fixed at 2.2 as determined from the substitution type counts between the domestic donkey genome and modern horse genomes (Orlando et al. 2013, see their Supplemental section S10.1.c). Genes for which there were at least 2 non-synonymous substitutions in the domestic clade, and for which the ω of the domestic clade was both > 1 and greater than the background were selected for subsequent analysis, yielding a total of 454 candidate genes.

PAML 'codeml' was run on the 454 selected genes, allowing 2 possible values for ω; one value for the domestic clade and one value for the rest of the tree. In addition, a null model in which a single ω was assigned to the entire tree was used. To account for problems with the numerical optimization in PAML, each test was run twice, with different starting values for ω (1.5 and 3.0); if the log-likelihood of a test for a given gene was lower than that of the null test, PAML codeml was re-run for that gene. Subsequently, the best model was selected for both the null model and the test based on the log-likelihood of each pair of tests, and the likelihood-ratio test LRT was calculated for each gene.

Due to the highly conserved nature of protein coding genes, the LRT scores do not follow a Chi^2 distribution, and could therefore not be used directly to determine *p*-values of tests. To detect statistically significant tests, *q*-values were estimated using a false discovery rate (FDR) modulated sequential MC algorithm (110); for each simulation, the parameters derived from the null model for the gene were used to simulate sequences of the same length as the gene using PAML 'evolver'. PAML 'codeml' was run as described above for the genes themselves. For each simulation for each gene, the resulting LRT was compared with the LRT observed for a given gene, with a cut-off value (h) of 10 LRT values at least as significant as that observed for the real sequences. Consequently, 36 genes were selected with a FDR of 5% (Supplementary Table S30).

Notably, among these genes are two olfactory receptors (*OR4E2* and *OR2A25*), and one gene involved in the regulation of axon growth in olfactory sensory neurons (*RAP1GAP2*). Olfactory receptors were similarly observed for genomic scans of the domesticated pig (*Sus scrofa*) (111), suggesting a potential commonality in the effect of domestication of these species. In addition to olfactory genes, several genes (*ADAMTS1*, *PLEKHM1*, *RTRD1*, and *THSD7A*) have been observed to relate to bone structure, potentially resulting from the use of the horse as a draft animal. The gene *SYNJ2* is furthermore known to be a longevity gene candidate in humans, with SNPs in this gene associated with statistically significant changes in levels of agreeableness (112, 113).

Ensembl Gene ID	Gene Name	q-value	ω(null)	ω(bg)	ω(domestic)
ENSECAG00000000264	*DCT*	0.048	0.149	0.049	88.628
ENSECAG00000000292	*THSD7A*	0.031	0.138	0.061	∞
ENSECAG00000001279	*AGTR1*	0.031	0.272	0.000	∞
ENSECAG00000002752	*OR4E2*	0.031	0.440	0.145	∞
ENSECAG00000002769		0.031	0.262	0.000	∞
ENSECAG00000006310	*OR2A25*	0.048	0.204	0.000	218.467
ENSECAG00000007936	*SARDH*	0.031	0.061	0.020	∞
ENSECAG00000008303	*LAMB1*	0.035	0.217	0.158	104.501
ENSECAG00000008427	*SLC43A1*	0.031	0.238	0.000	∞
ENSECAG00000010028		0.048	0.076	0.059	∞
ENSECAG00000010092	*POLR1A*	0.031	0.058	0.035	168.09
ENSECAG00000010370	*CDK5RAP1*	0.031	0.257	0.000	18.475
ENSECAG00000010758	*KIF24*	0.050	0.633	0.361	∞
ENSECAG00000011312	*RTDR1*	0.031	0.163	0.000	∞
ENSECAG00000011659	*KIAA1549*	0.031	0.335	0.237	330.347
ENSECAG00000012056	*MLXIP*	0.035	0.222	0.110	443.218
ENSECAG00000012646	*SYNJ2*	0.035	0.190	0.100	1.894
ENSECAG00000013824	*POC5*	0.031	0.209	0.031	∞
ENSECAG00000014500	*ART5*	0.031	0.189	0.000	428.945
ENSECAG00000014501	*TMEM54*	0.048	0.702	0.000	∞
ENSECAG00000015852	*GOT1L1*	0.048	0.201	0.000	∞
ENSECAG00000016339	*ADAMTS1*	0.031	0.085	0.000	8.039
ENSECAG00000016408	*CABP4*	0.050	0.163	0.000	0.813
ENSECAG00000017284	*ANKDD1A*	0.048	0.108	0.023	1.123
ENSECAG00000018068	*MROH2B*	0.031	0.208	0.113	∞
ENSECAG00000018566	*TMEM63A*	0.048	0.074	0.049	∞
ENSECAG00000019738	*MTMR3*	0.050	0.346	0.172	∞
ENSECAG00000020278	*RAP1GAP2*	0.035	0.158	0.063	127.614
ENSECAG00000021040	*CDC42BPB*	0.031	0.353	0.253	∞
ENSECAG00000022722	*ESPL1*	0.048	0.162	0.129	218.485
ENSECAG00000022735	*DOPEY2*	0.031	0.176	0.136	375.465
ENSECAG00000023888	*ALDH1L2*	0.031	0.091	0.000	∞
ENSECAG00000024194	*POLA2*	0.031	0.080	0.000	∞
ENSECAG00000024397	*UBR4*	0.041	0.110	0.081	1.313
ENSECAG00000024405	*CATSPER1*	0.048	0.180	0.085	1.406
ENSECAG00000025023	*PLEKHM1*	0.031	0.110	0.044	∞

Supplementary Table S30. Genes undergoing positive selection in the domestic clade

The infinity symbol signifies that maximum value allowed by 'codeml' during the model optimization stage (999.0); ω(null) signifies the global ω of the null model, ω(bg) and ω(domestic) signifies the omegas of the background (wild horses and the domestic donkey (Willy)) and the domestic horses respectively.

S4.2 Screening for selective sweeps using $\widehat{\theta}_w$ and Tajima's D statistic

S4.2.1 Overall methodology

Regions in which the domestic horses show low genetic diversity and deviation from neutrality as compared to either the pre-domesticated horses or Przewalski's horses are proposed as candidates for evolving under selection since the onset of horse domestication. To identify such regions, the genomes were analyzed by calculating the Watterson estimator ($\hat{\theta}_w$) (18) and Tajima's D statistic (114) using a sliding window approach. To compensate for missing data and the variation in the depth-of-coverage across the different genomes (ranging from 7.44× to 32.66×; Supplementary Table S1) an empirical Bayes method based on genotype likelihoods was used to calculate the posterior probabilities for the sample frequency spectrum using a maximum likelihood estimate of the site frequency spectrum as the prior (115). The implementation is available in the software *angsd* (http://www.popgen.dk/angsd). In addition to $\hat{\theta}_w$, the nucleotide diversity (π) was estimated and used for the Tajima's D statistic. The prior was estimated on chromosome 22 for each set, and was used to screen the genomes using a window size of 50 kb and a step size of 10 kb. Only the genomic windows in which at least 90% of bases were covered were considered in order to avoid coverage-related bias.

For each genomic window, the $\hat{\theta}_w$ log-ratio was calculated as $ln(\hat{\theta}_{w_{PD}}) - ln(\hat{\theta}_{w_D})$, where $\hat{\theta}_{w_{PD}}$ is the Watterson estimator for the pre-domesticated horse genomes, and $\hat{\theta}_{w_D}$ is the Watterson estimator for the domestic horse genomes, using the natural logarithm. Selective sweeps results in decreased variation in regions under positive natural selection (116), causing a local increase in the ratio calculated above. For modern horse breeds the Tajima's D statistic was calculated using the same sliding windows approach in order to detect deviation from neutrality; more specifically, selective sweeps decrease the amount of polymorphism around the site under selection, resulting in a negative Tajima's D value. Therefore, candidate genomic regions are selected when a local increase in the $\hat{\theta}_w$ log-ratio is associated with a decrease of the Tajima's D statistic for modern horses within the same region.

For ancient specimens, *post-mortem* DNA damage artificially increases the nucleotide diversity and consequently the $\hat{\theta}_w$ log-ratio. Supplementary Figure S39 shows the correlation between the $\hat{\theta}_w$ of pre-domesticated horses, calculated with or without transition, which indicates that the presence of *post-mortem* DNA damage results in a *global* - but importantly, not a local - shift in the $\hat{\theta}_w$. As a consequence, the relationship between the $\hat{\theta}_w$ of different genomic regions remains constant, and the presence of *post-mortem* DNA is not expected to lead to an increase in the false discovery rate.

Three scans using the $\hat{\theta}_w$ log-ratio were performed: In analysis I (Supplementary Section S4.2.2), we calculated the $\hat{\theta}_w$ log-ratios for the domestic horses and the pre-domesticated horses, but excluded the Icelandic (P5782); in analysis II (Supplementary Section S4.2.4), we calculated the $\hat{\theta}_w$ log-ratios for *all* of the domestic horses and the pre-domesticated horses; and in analysis III (Supplementary Section S4.2.5) we calculated the $\hat{\theta}_w$ log-ratio the domesticated horses, and Przewalski's horse, but excluded the Icelandic (P5782). Analyses were performed with and without the Icelandic (P5782) on the ground that it was sequenced using a different technology than the remaining samples, and could therefore potentially introduce a systematic bias in the analyses.

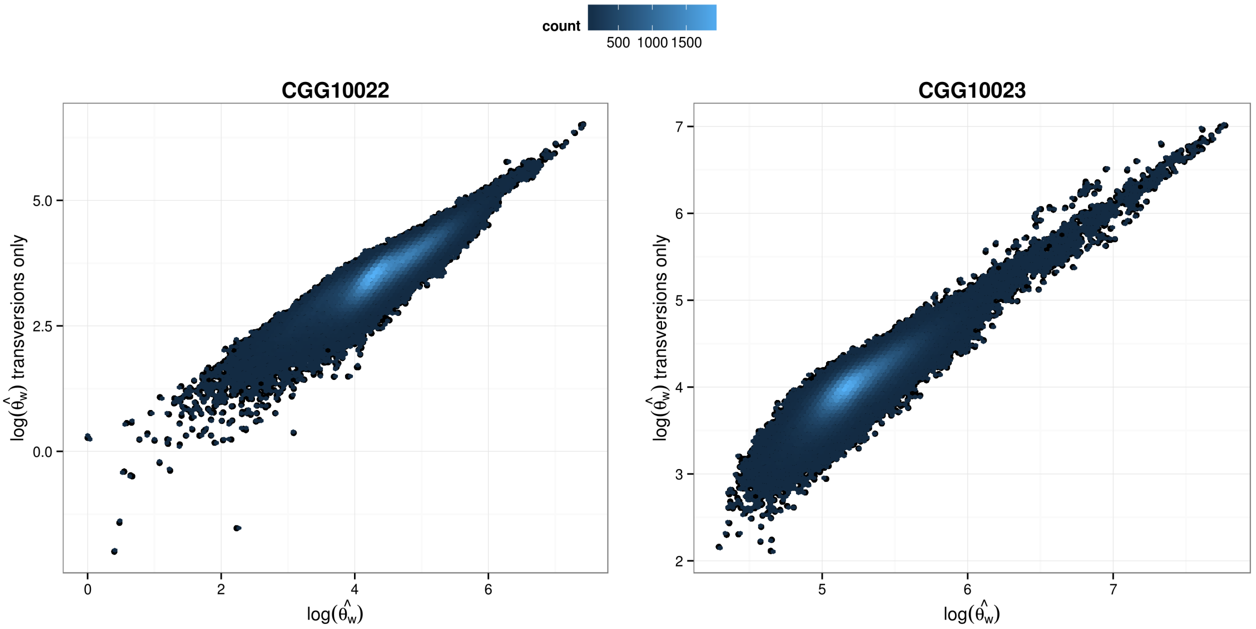
S4.2.2 Analysis I: Determining outliers in $\widehat{\theta}_w$ log-ratio and Tajima's D values

The procedure to extract outliers was defined as follows: A cubic spline was fitted to the $\hat{\theta}_w$ log-ratio and Tajima's D values from all genomic windows, using the R function *smooth.spline* (26).

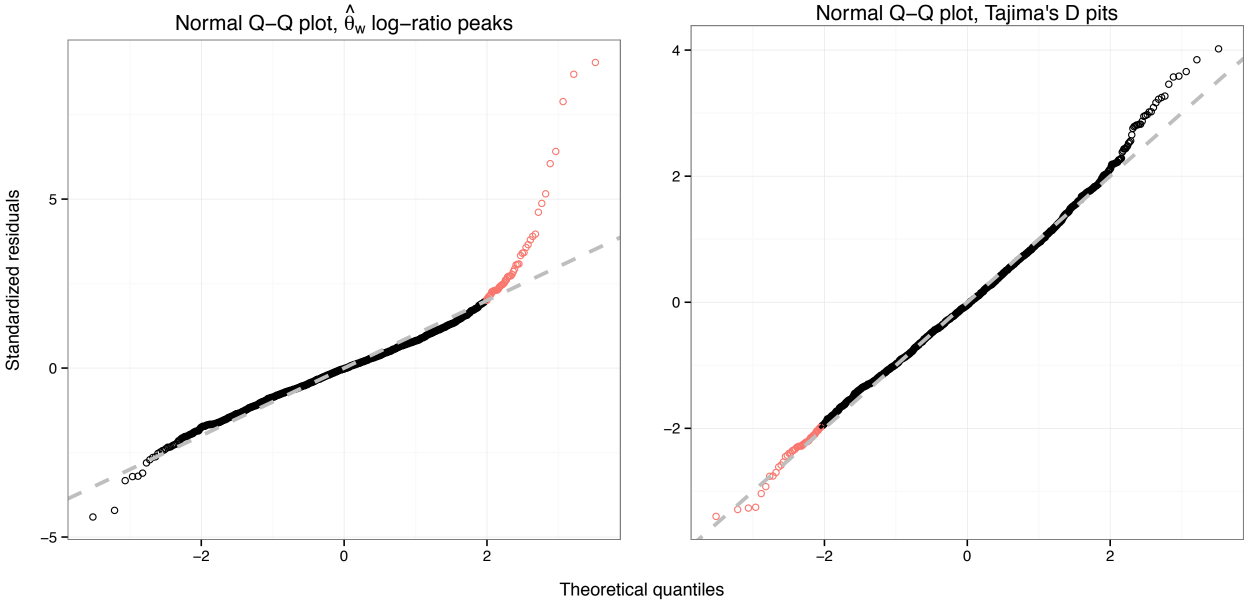
Peaks and pits (local increase and decrease) were determined using the *turnpoints* function in the R package ‘pastecs’ (http://cran.r-project.org/web/packages/pastecs/). The obtained pits and peaks were then used to define the limits of genomic regions showing a local increase and decrease for the $\widehat{\theta}_w$ log-ratio, and Tajimas’s D values respectively. In the regions defined by the pits and peaks, mean values for the summary statistics ($\widehat{\theta}_w$ log-ratio and Tajima's D) were computed. A normal approximation was applied to the summary statistics averages for the peaks/pits defined regions (Supplementary Figure S40). Outliers were defined as values found greater than the 95% upper predictive quantile for the $\widehat{\theta}_w$ log-ratio (red points; Supplementary Figure S40, left) and less than the 5% lower predictive quantile for the Tajima's D values (red points; Supplementary Figure S40, right panel). The resulting outliers are depicted as triangles on Supplementary Figure S41.

57 candidate regions with a significantly higher $\widehat{\theta}_w$ log-ratio were detected and are reported in Supplementary Table S31. Among those regions, 16 also intersect with regions showing significant low Tajima's D. The intersection of these regions is reported in Supplementary Table S32. The quantile-quantile plot of the summary statistics for the Tajima's D demonstrate that the outliers do not deviate from the normal distribution quantile (Supplementary Figure S40). Hence, using modern domestic horses does not bring the sensitivity needed to detect deviation from neutrality. Thus intercepting high $\widehat{\theta}_w$ log-ratio and low Tajima's D is over-conservative for identifying the candidate regions for selection.

Of note, those 16 candidate regions also include two genes identified in previous genome selection scans: *MC1R* (chr3: 35,115,000 - 37,265,000) which plays a role in coat color; and *KITLG (*chr28: 14,375,000 - 15,135,000), a ligand that was previously identified as a potential gene under selection while scanning modern domestic horse genomes (1). *KITLG* encodes a ligand which is a pleiotropic factor involved in fertility, neural cell development and hematopoiesis, and which is associated with the roan coat-color in cattle (*Bos taurus;* gene symbol *MGF)* (117).

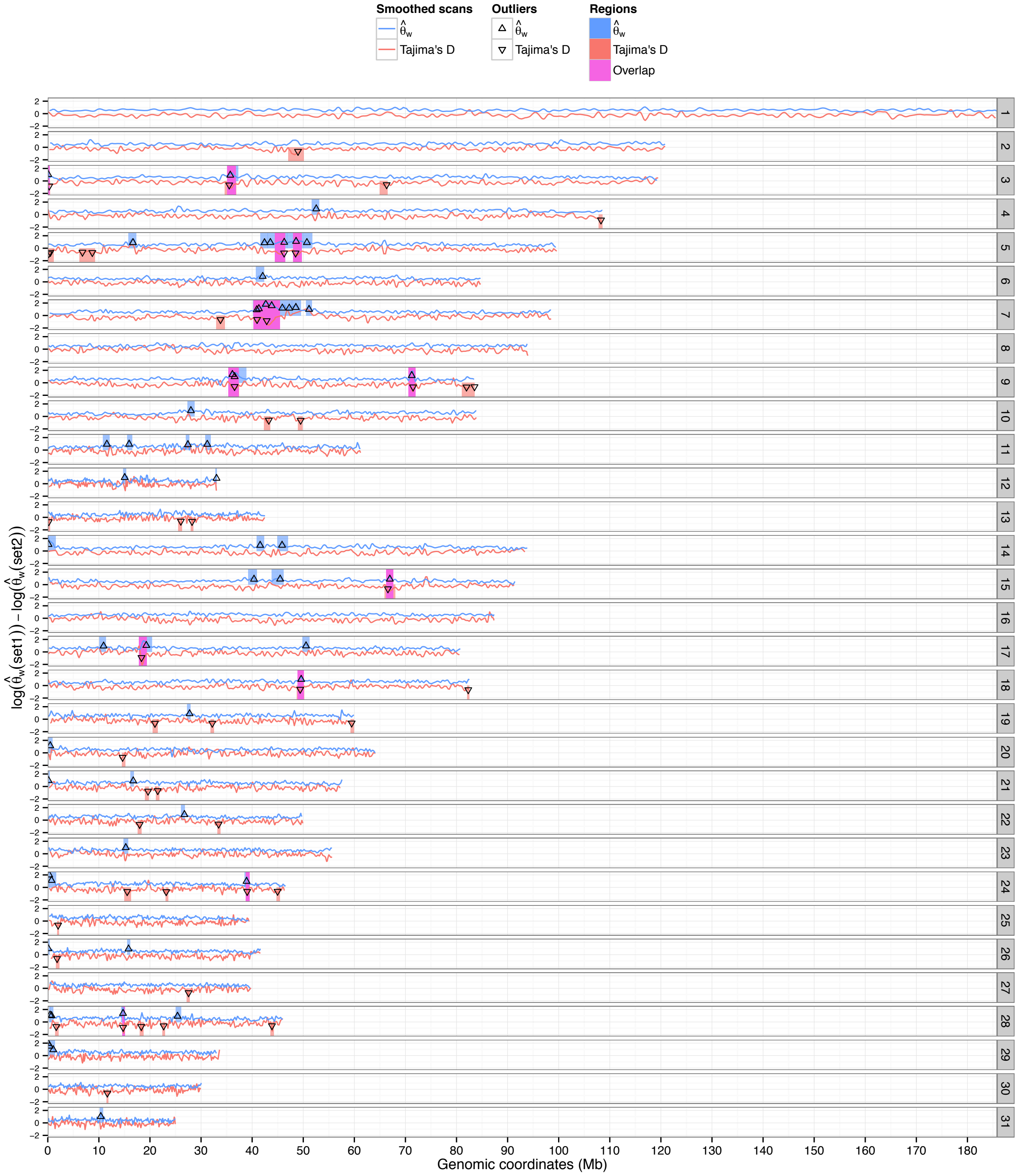


Supplementary Figure S39. Correlation between $\widehat{\theta}_w$* log values *with or without transitions for pre-domesticated horses



Supplementary Figure S40. Q-Q plots for the normal approximation to the local average of the summary statistics (*$\widehat{\theta}_w$ **log-ratio and Tajima's D*) in the comparison between pre-domestic horses and domesticated horses (excluding Icelandic (P5782))

Comparison between the pre-domesticated horses, CGG10022, CGG10023, and the domestic horse breeds, Arabian, Icelandic, Norwegian Fjord, Standardbred, and Thoroughbred (Twilight). Outliers are defined using the upper or lower 5% predictive intervals for $\widehat{\boldsymbol{\theta}}_w$ *log-ratio and Tajima's D respectively and depicted in red.*



Supplementary Figure S41. Screening for selective sweeps using the $\widehat{\theta}_w$ log-ratio between pre-domesticated horses (set1) and domestic horses (set) and Tajima's D values among domestic

Autosomes are depicted. The cubic spline fitted to the $\widehat{\theta}_w$ log-ratio and Tajima's D values for all 50 kb sliding windows with a 10 kb step is reported by blue and red lines respectively. The outlier regions are depicted with rectangle and mean of the summary statistics within the outlier regions are presented with an arrow.

Chromosome	Start	End	Peak	y_{mean}
3	**35,000**	**375,000**	**75,000**	**1.047**
3	**35,115,000**	**37,265,000**	**35,715,000**	**0.922**
4	51,685,000	53,135,000	52,545,000	0.917
5	15,795,000	17,365,000	16,675,000	0.897
5	41,615,000	43,045,000	42,415,000	0.885
5	43,045,000	44,415,000	43,565,000	0.922
5	**44,415,000**	**46,845,000**	**46,275,000**	**0.948**
5	**46,845,000**	**50,555,000**	**48,595,000**	**1.100**
5	50,555,000	51,815,000	50,715,000	0.905
6	40,755,000	42,385,000	42,065,000	0.875
7	**40,165,000**	**41,065,000**	**40,905,000**	**0.920**
7	**41,065,000**	**41,905,000**	**41,315,000**	**1.086**
7	**41,905,000**	**43,165,000**	**42,665,000**	**1.811**
7	**43,165,000**	**45,445,000**	**43,825,000**	**1.576**
7	45,445,000	46,875,000	45,905,000	1.187
7	46,875,000	47,795,000	47,265,000	1.176
7	47,795,000	49,615,000	48,555,000	1.290
7	50,495,000	51,675,000	51,135,000	0.985
9	**35,315,000**	**36,465,000**	**36,135,000**	**1.332**
9	**36,465,000**	**38,875,000**	**36,565,000**	**0.991**
9	**70,565,000**	**71,965,000**	**71,235,000**	**1.160**
10	27,375,000	28,735,000	28,015,000	0.928
11	10,795,000	12,085,000	11,565,000	0.939
11	15,565,000	16,565,000	15,915,000	0.953
11	27,005,000	27,715,000	27,425,000	0.874
11	30,855,000	31,935,000	31,215,000	0.922
12	14,705,000	15,325,000	15,055,000	0.988
12	32,805,000	33,065,000	33,025,000	0.869
14	35,000	1,575,000	145,000	1.044
14	40,875,000	42,345,000	41,565,000	0.915
14	44,895,000	46,985,000	45,855,000	0.923
15	39,175,000	40,895,000	40,365,000	0.870
15	43,845,000	46,185,000	45,475,000	0.894
15	**66,255,000**	**67,615,000**	**66,945,000**	**0.885**
17	10,085,000	11,355,000	10,925,000	0.944
17	**17,885,000**	**20,435,000**	**19,285,000**	**1.042**
17	49,805,000	51,175,000	50,545,000	0.971
18	**48,885,000**	**50,135,000**	**49,595,000**	**0.997**
19	27,275,000	27,945,000	27,755,000	0.874
20	185,000	965,000	485,000	1.137
21	155,000	335,000	155,000	0.936
21	16,185,000	16,875,000	16,725,000	0.900
22	26,075,000	26,805,000	26,745,000	0.892
23	14,785,000	15,705,000	15,265,000	0.988
24	195,000	355,000	275,000	1.995
24	355,000	1,635,000	735,000	1.124
24	**38,525,000**	**39,485,000**	**38,895,000**	**0.970**
26	65,000	365,000	105,000	0.958

26	15,495,000	16,105,000	15,805,000	0.918
28	235,000	525,000	445,000	1.097
28	525,000	1,095,000	735,000	1.011
28	**14,375,000**	**15,135,000**	**14,725,000**	**1.377**
28	25,065,000	26,125,000	25,395,000	0.925
29	85,000	275,000	105,000	1.940
29	275,000	695,000	345,000	1.519
29	695,000	1,355,000	1,065,000	0.949
31	10,145,000	10,805,000	10,375,000	1.023

Supplementary Table S31. Candidate genomic regions for selective sweeps, in the comparison between two pre-domesticated horses and five domestic horse breeds

Start and end refers to the external coordinates for the regions defined by two valleys of the $\hat{\theta}_w$ log-ratios cubic spline. The columns "Peak" and y_{mean} corresponds to the genome coordinate of the peak of the $\hat{\theta}_w$ log-ratio and the local mean within the region respectively. Regions that also show a significantly decreased Tajima's D value are reported in bold.

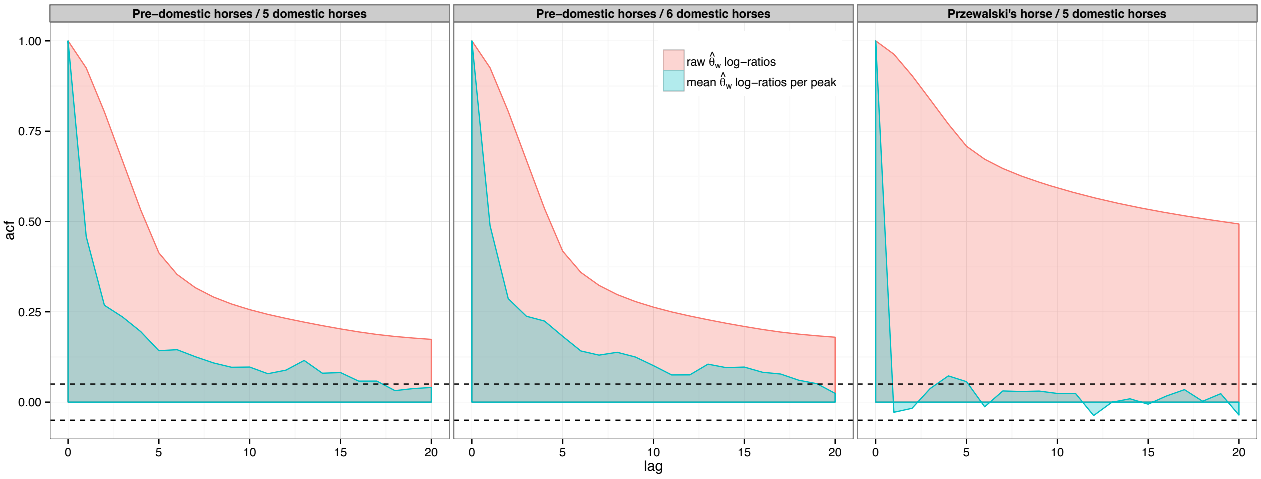
Chromosome	Start	End	$y_{mean(\hat{\theta}_w)}$	$y_{mean(D)}$
3	35,000	375,000	1.047	-0.806
3	35,115,000	36,855,000	0.922	-0.650
5	44,455,000	46,445,000	0.948	-0.770
5	48,035,000	49,765,000	1.100	-0.771
7	40,195,000	41,065,000	0.920	-0.641
7	41,065,000	41,905,000	1.086	-0.641
7	41,905,000	41,935,000	1.811	-0.641
7	41,935,000	43,165,000	1.811	-0.830
7	43,165,000	45,435,000	1.576	-0.830
9	35,315,000	36,465,000	1.332	-0.606
9	36,465,000	37,415,000	0.991	-0.606
9	70,585,000	71,965,000	1.160	-0.669
15	66,255,000	67,615,000	0.885	-0.673
17	17,885,000	19,405,000	1.042	-0.907
18	48,885,000	50,135,000	0.997	-0.601
24	38,765,000	39,485,000	0.970	-0.658
28	14,615,000	15,105,000	1.377	-0.879

Supplementary Table S32. Intersection of genomic regions for selective sweeps defined by the overlap of between $\hat{\theta}_w$ peaks and Tajima's D valleys, in the comparison between two pre-domesticated horses and five domestic horse breeds

*Candidate genomic regions were selected based on the presence of an increased $\hat{\theta}_w$ log-ratios, overlapping a region with a decreased Tajima's D (*Supplementary Table S31*); these were truncated to just the region of overlap between the two measures.*

S4.2.3 Quality controls

The three analyses were performed using a window size of 50 kb with a step of 10 kb. Because of this, and due to moderate LD in horses (twice the r^2 background at 100 – 150 kb (3)), the genomic regions analyzed are not independent. The autocorrelation of the summary statistics was reduced, without resorting to using large non-overlapping regions with the accompanying poor resolution, by using the mean of the summary statistics within regions defined by the local extremities. The lag correlation of the summary statistics with different offsets was calculated with the '*acf*' function in R *(26)* and the results are shown in Supplementary Figure S42 for the pre-domesticated horses and the domestic horses, excluding Icelandic (P5782), for the same dataset including the Icelandic (P5782), and when using the Przewalski instead of the pre-domesticated horses.



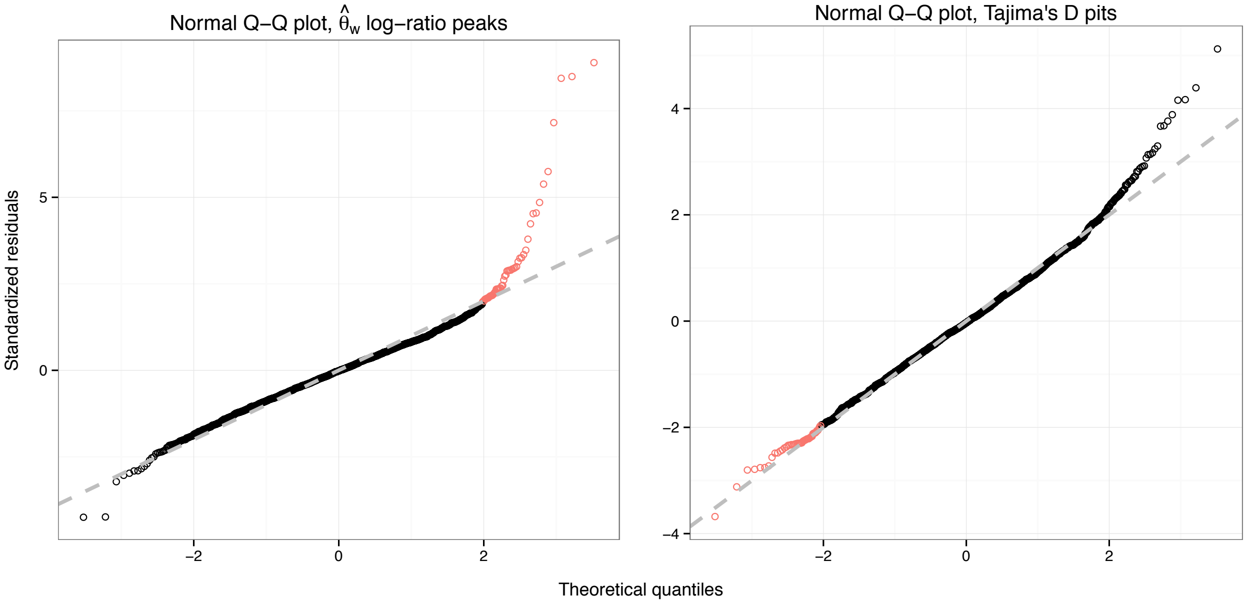
Supplementary Figure S42. Autocorrelation plots of all $\hat{\theta}_w$ log-ratios (left) and mean $\hat{\theta}_w$ log-ratios retained per peak (right)

Left: *Auto-correlation for the pre-domesticated horses and the domestic horses, excluding the Icelandic (P5782);* ***Center:*** *for the pre-domesticated horses and all 6 domestic horses;* ***Right:*** *for Przewalski's horse and the domestic horses, excluding Icelandic (P5782). Performed using the 'acf' function in R (26).; lag represents the distance in 10kbp windows between the each window, and the window with which it is compared; the stippled lines demarcate the region indicating +/- 0.05 Pearson-correlation coefficient.*

S4.2.4 Analysis II: Adding the Icelandic (P5782) genome to the set of domestic horse breeds

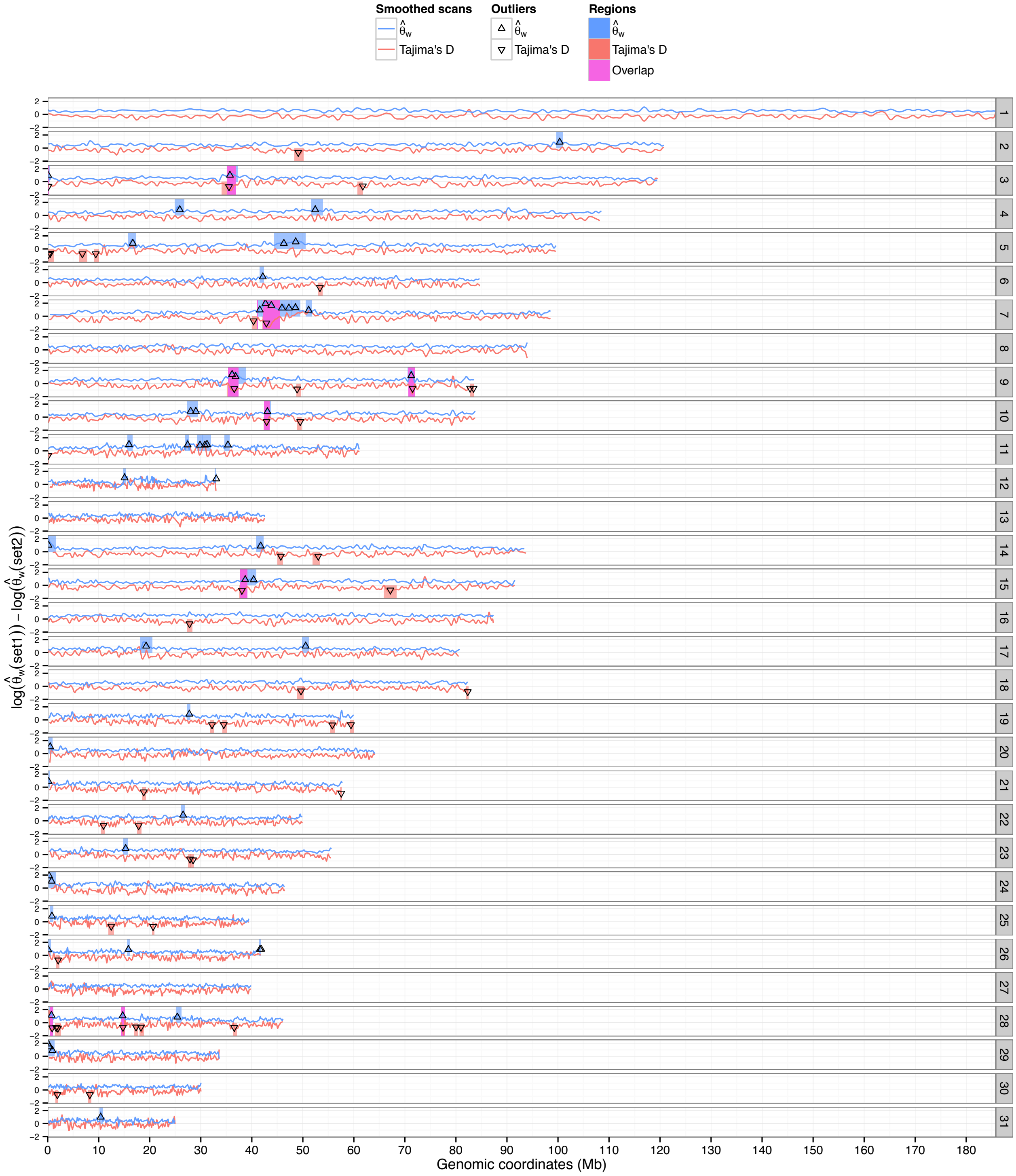
In the second analysis, we investigated the same parameters, the $\hat{\theta}_w$ log-ratio and Tajima's D, but adding an additional domestic horse: Icelandic (P5782) (2). Following the same procedure as described in section S4.2.2, we selected outliers for both the $\hat{\theta}_w$ log-ratios (Supplementary Figure S43, left panel) and Tajima's D values (Supplementary Figure S43, right panel) using a 5% threshold for predictive intervals. The resulting selection scans are depicted on Supplementary Figure S44.

We obtained 55 candidate regions using the 5% threshold for detecting $\hat{\theta}_w$ log-ratio outliers to the upper predictive interval. 46 of these regions overlapped the 57 regions detected when excluding the Icelandic (P5782) in analysis I, accounting for 74.1% of the genomic regions in the first analysis. Out of the 55 candidate regions, 12 were confirmed by unusual low Tajima's D values, i.e., outliers to 5% of the lower predictive interval (Supplementary Table S33); the regions intersecting the two measures are reported in Supplementary Table S34. The 12 truncated regions overlap 9 of the 16 truncated regions detected in the first analysis (see above), accounting for 47.0% of the genomic regions in the previous set. Notably, these 9 overlapping regions include both of the genes described previously (*MC1R* and *KITLG*).



Supplementary Figure S43. Q-Q plots for the normal approximation to the local average of the summary statistics (*$\hat{\theta}_w$ **log-ratio and Tajima's D**) in the comparison between pre-domesticated horses and the domestic horses***

Comparison between the pre-domesticated horses, CGG10022, CGG10023, and the domestic horse breeds, Arabian, Icelandic, Norwegian Fjord, Standardbred, Thoroughbred (Twilight), and the Icelandic (P5782). Outliers are defined using the upper or lower 5% predictive intervals for $\hat{\theta}_w$ *log-ratio and Tajima's D respectively and depicted in red.*



Supplementary Figure S44. Screening for selective sweeps using the $\widehat{\theta}_w$ log-ratio between pre-domesticated (set1) and domestic horses (set2) including Icelandic (P5782), 5% significant threshold

The smooth spline that fits the $\widehat{\theta}_w$ log-ratio and Tajima's D values for all 50 kb sliding windows with a 10 kb step is reported using blue and red lines respectively. The outlier regions are depicted with rectangles and the mean of the summary statistics within the outlier regions are presented with an arrow.

Chromosome	Start	End	Peak	y_{mean}
2	99,725,000	101,015,000	100,335,000	0.907
3	**35,000**	**375,000**	**75,000**	**0.992**
3	**35,115,000**	**37,265,000**	**35,715,000**	**0.994**
4	24,855,000	26,725,000	25,885,000	0.864
4	51,565,000	53,905,000	52,425,000	0.852
5	15,795,000	17,365,000	16,675,000	0.872
5	44,325,000	46,805,000	46,275,000	0.861
5	46,805,000	50,495,000	48,595,000	1.086
6	41,525,000	42,395,000	42,155,000	0.877
7	**41,105,000**	**41,945,000**	**41,555,000**	**0.972**
7	**41,945,000**	**43,135,000**	**42,665,000**	**1.952**
7	**43,135,000**	**45,455,000**	**43,825,000**	**1.675**
7	45,455,000	46,885,000	45,905,000	1.258
7	46,885,000	47,805,000	47,265,000	1.255
7	47,805,000	49,485,000	48,555,000	1.307
7	50,545,000	51,675,000	51,135,000	0.921
9	**35,315,000**	**36,475,000**	**36,135,000**	**1.391**
9	**36,475,000**	**38,895,000**	**36,815,000**	**1.049**
9	**70,555,000**	**71,965,000**	**71,235,000**	**1.208**
10	27,355,000	28,695,000	28,015,000	0.911
10	28,695,000	29,415,000	29,035,000	0.882
10	**42,405,000**	**43,725,000**	**43,055,000**	**0.848**
11	15,605,000	16,565,000	15,925,000	0.967
11	26,995,000	27,715,000	27,425,000	0.907
11	29,325,000	30,315,000	29,845,000	0.869
11	30,315,000	30,825,000	30,805,000	0.876
11	30,825,000	31,955,000	31,215,000	0.990
11	34,635,000	35,655,000	35,315,000	0.896
12	14,715,000	15,335,000	15,055,000	1.005
12	32,765,000	33,065,000	33,025,000	0.856
14	35,000	1,575,000	145,000	0.999
14	40,835,000	42,325,000	41,735,000	0.875
15	**37,735,000**	**39,145,000**	**38,705,000**	**0.866**
15	39,145,000	40,885,000	40,355,000	0.851
17	18,105,000	20,475,000	19,285,000	1.001
17	49,815,000	51,145,000	50,545,000	1.010
19	27,275,000	27,975,000	27,755,000	0.885
20	185,000	965,000	485,000	1.033
21	155,000	355,000	155,000	0.891
22	26,065,000	26,835,000	26,535,000	0.878
23	14,805,000	15,735,000	15,265,000	0.912
24	195,000	385,000	275,000	1.880
24	385,000	1,605,000	755,000	1.066
25	495,000	1,055,000	815,000	0.866
26	55,000	605,000	105,000	0.860
26	15,515,000	16,125,000	15,805,000	0.906
26	41,425,000	41,785,000	41,595,000	0.923
26	41,785,000	41,835,000	41,835,000	0.948
28	**235,000**	**1,095,000**	**735,000**	**1.137**
28	**14,385,000**	**15,155,000**	**14,685,000**	**1.051**
28	25,085,000	26,175,000	25,395,000	0.863
29	85,000	275,000	105,000	1.888

29	275,000	705,000	345,000	1.450
29	705,000	1,335,000	925,000	0.907
31	10,145,000	10,805,000	10,375,000	0.993

Supplementary Table S33. Genomic regions candidate for selective sweeps, in the comparison between 2 pre-domesticated horses and 6 domestic horse breeds

Start and end refers to the external coordinates for the regions defined by two valleys of the $\hat{\theta}_w$ log-ratios cubic spline. The columns "Peak" and y_{mean} corresponds to the genome coordinate of the peak of the $\hat{\theta}_w$ log-ratio and the local mean within the region, respectively. Regions that also show a significantly decreased Tajima's D value are reported in bold.

Chromosome	**Start**	**End**	$\mathbf{y_{mean(\widehat{\theta}_w)}}$	$\mathbf{y_{mean(D)}}$
3	35,000	345,000	0.992	-0.712
3	35,115,000	36,905,000	0.994	-0.781
7	41,105,000	41,185,000	0.972	-0.715
7	42,125,000	43,135,000	1.952	-1.045
7	43,135,000	45,455,000	1.675	-1.045
7	45,455,000	45,455,000	1.258	-1.045
9	35,315,000	36,475,000	1.391	-0.773
9	36,475,000	37,415,000	1.049	-0.773
9	70,655,000	71,965,000	1.208	-0.740
10	42,405,000	43,505,000	0.848	-0.688
15	37,735,000	39,125,000	0.866	-0.756
28	525,000	1,015,000	1.137	-0.799
28	14,405,000	15,115,000	1.051	-0.721

Supplementary Table S34. Truncated genomic regions for selective sweeps delimited by $\hat{\theta}_w$ valleys and Tajima's D peaks, in the comparison between 2 pre-domesticated horses and 6 domestic horse breeds

*Candidate genomic regions were selected based on the presence of an increased the $\hat{\theta}_w$ log-ratios, overlapping a region with a decreased Tajima's D (*Supplementary Table S33*); these were truncated to just the region of overlap between the two measures.*

S4.2.5 Analysis III: Comparing the Przewalski's horse to domestic horse breeds

The Przewalski's horse sequenced in Orlando *et al.* 2013 (1) did not show significant levels of admixture with any of the four investigated domestic horse breeds: Arabian, Icelandic, Norwegian Fjord, and Standardbred. The study also estimated that the Przewalski's horse population and the population that led to modern domestic lineages diverged 38-72 kyr BP, i.e. much earlier than the earliest known evidence of horse domestication (5.5 kyr BP) (5). We therefore compared the $\hat{\theta}_w$ between the Przewalski horse and the domestic horses (excluding the Icelandic (P5782)) in order to detect selection that differentiated both populations, possibly in relation – but not exclusively – with domestication.

As a single individual represents the Przewalski's horse in this study, the $\hat{\theta}_w$ equals the heterozygozity. As consequence, homozygous regions in the genome of this individual will bias the $\hat{\theta}_w$ log-ratio, preventing the direct detection of outliers using the methodology described in section S4.2.2. To resolve this issue, we used a running median on 151 windows (1.55 Mbp). This number of windows was determined empirically to offer a reasonable tradeoff between over-smoothing and insufficient smoothing. The difference between the polynomial smooth-lines with and without the running median was subtracted from the $\hat{\theta}_w$ log-ratios. Following correction, genomic regions were determined as described in section S4.2.2. The corrected regions and log-ratios are henceforth referred to as "adjusted". Resulting selection scans are displayed on Supplementary Figure S46, with the uncorrected $\hat{\theta}_w$ log-ratios shown in green, and adjusted $\hat{\theta}_w$ log-ratios shown in blue. The effect of running median correction could be assessed on the normal qqplots of $\hat{\theta}_w$ log-ratios (Supplementary Figure S45).

This resulted in 34 candidate regions, overlapping 6 of the regions detected using the pre-domesticated horses (Analysis I), and covering 8.1% of the genomic regions detected for that test; similarly five of the regions detected for the present analysis overlapped the regions detected in analysis II, accounting for 7.5% of the genomic regions.

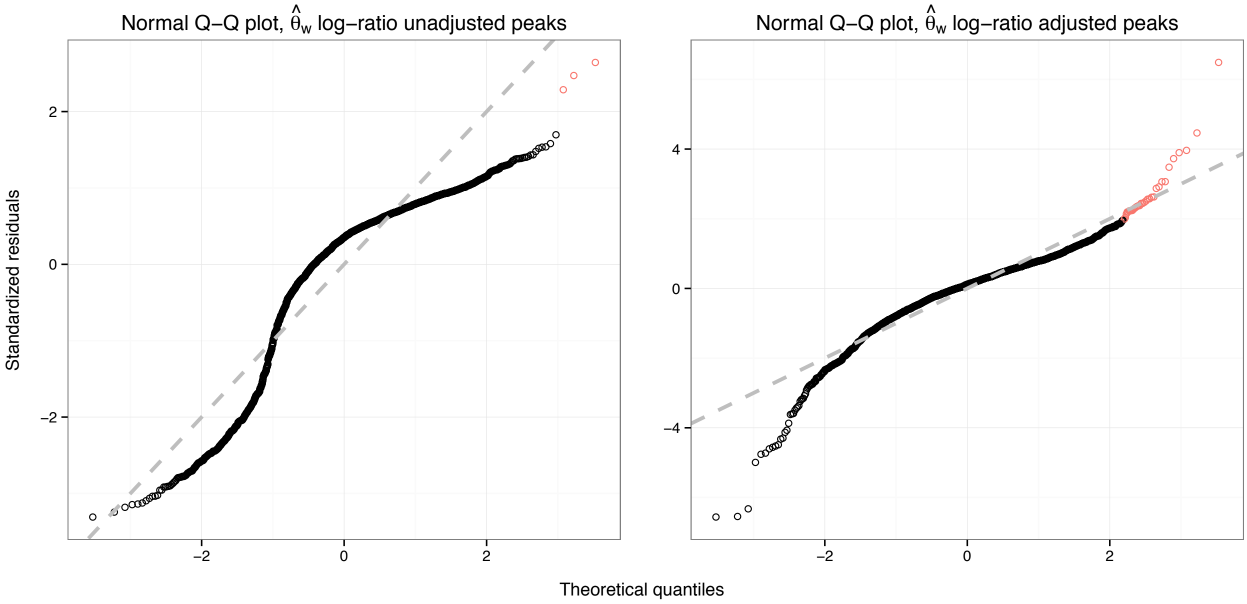
Of out of the 34 candidate regions detected using $\hat{\theta}_w$ log-ratio outliers to the upper predictive interval, five were characterized by low Tajima's D values, i.e., outliers to 5% of the lower predictive interval (Supplementary Table S35). The regions intersecting the two measures are reported in Supplementary Table S36. These five regions intersected with three of the 16 regions detected using the five domestic horses and the pre-domesticated horses (Analysis I), accounting for 14.9% of the genomic regions detected for that analysis (I). Notably, these three overlapping regions include one of the genes described previously (*KITLG*), but not the other (*MC1R*).

The sensitivity of this third analysis was lower due to the fact that $\hat{\theta}_w$ values for the Przewalski's horse were obtained using only one individual, and due to the fact that smoothing of $\hat{\theta}_w$ log-ratios was performed globally, and hence also smoothed out genuine outliers, which reduces our ability to detect peaks of medium intensities. As adding an additional domestic horse decreased our ability to detect regions undergoing selective sweeps (Supplementary Section S4.2.4), we did not perform the comparison of the Przewalski's horse with the domestic horses including the Icelandic (P5782).

S4.2.6 Detecting outlier genes in smaller regions

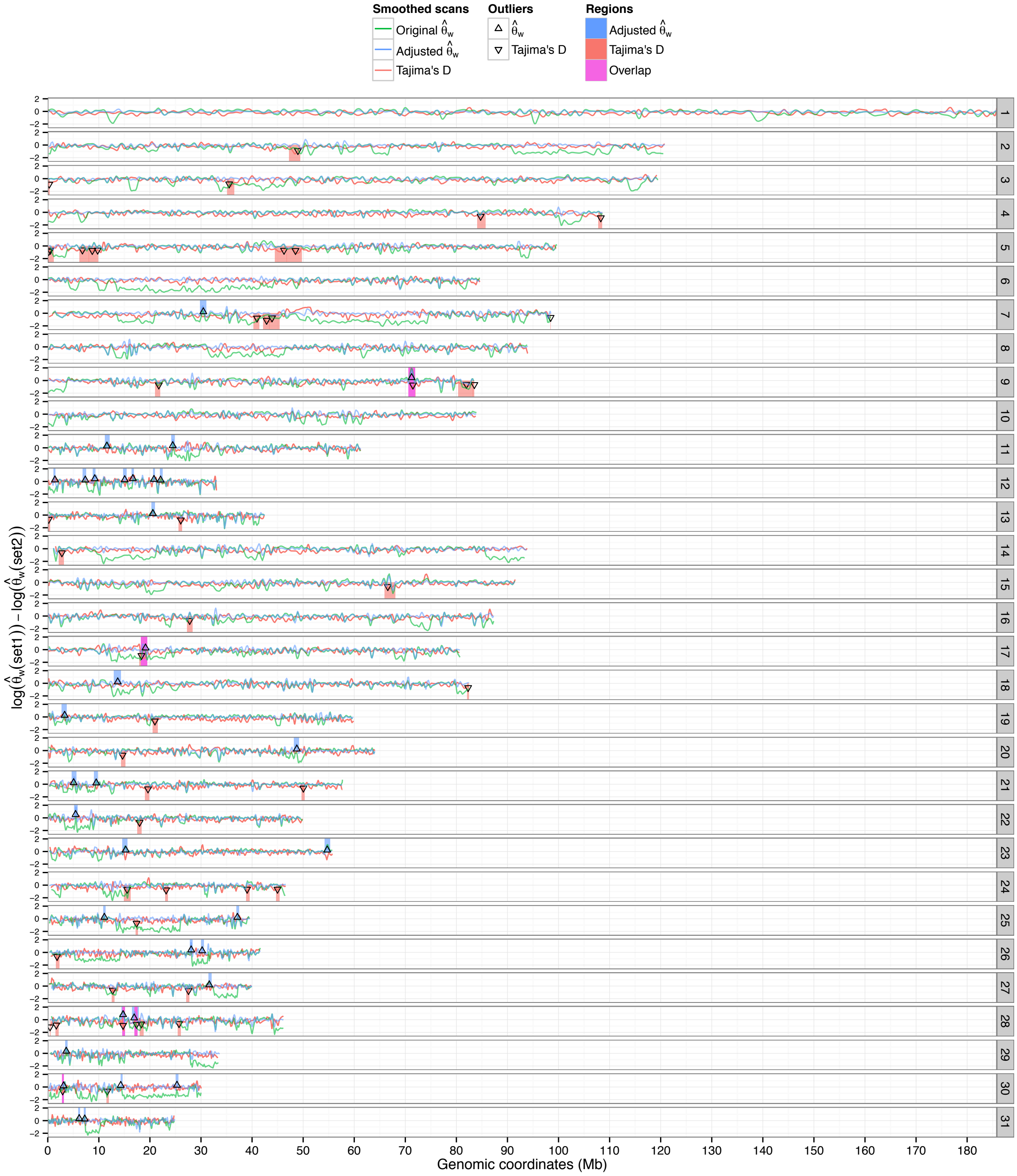
Smoothing the raw $\hat{\theta}_w$ log-ratios for all sliding windows to delimit regions prevent us from detecting peaks in smaller regions, and any gene found in such regions. In order to facilitate testing of all genes, we used an alternative method, in which each gene annotated in Ensembl v72 (80) was assigned the $\hat{\theta}_w$ log-ratio of the closest 50 kb sliding window. This was achieved by finding the window for which the distance between the center-point of the window, and the center point of the gene was minimized. The genes were ranked by the $\hat{\theta}_w$ log-ratio of the closet window and the top 1% or 5% were used as outliers. The procedure was applied to the three analyses performed earlier.

Notably, both *KITLG* and *MC1R* are ranked in the top 1% of results in the comparisons, with the exception of *MC1R* in the comparison between the domestic horses and the Przewalski's horse (analysis III, Supplementary Section S4.2.5).



Supplementary Figure S45.* Q-Q *plots for the normal approximation to the local average of the summary statistics (*$\hat{\theta}_w$ log-ratio and Tajima's D*) in the comparison between Przewalski's horse and domestic horses (excluding Icelandic (P5782))

Comparison between the Przewalski's horse, and the domestic horse breeds, Arabian, Icelandic, Norwegian Fjord, Standardbred, and Thoroughbred (Twilight). Outliers are defined using the upper or lower 5% predictive intervals for $\hat{\boldsymbol{\theta}}_w$ *log-ratio and Tajima's D respectively and depicted in red.*



Supplementary Figure S46. Screening for selective sweeps using $\widehat{\theta}_w$ log-ratio between the Przewalski's horse and domestic horses, 5% significant threshold

All autosomes are depicted. The smooth polynominal model that fits the $\widehat{\theta}_w$ log-ratio and Tajima's D values for all 50 kb sliding windows with a 10 kb step is reported using blue and red lines respectively. $\widehat{\theta}_w$ log-ratios were corrected for low-heterozygosity regions specific to the Przewalski's horse genome using a running median approach (Supplementary Section S4.2.5).

The outlier regions are depicted with rectangle and mean of the summary statistics within the outlier regions are presented with an arrow.

Chromosome	Start	End	Peak	y_{mean}
7	29,785,000	31,045,000	30,455,000	0.232
9	**70,625,000**	**71,935,000**	**71,235,000**	**0.465**
11	11,215,000	12,105,000	11,565,000	0.283
11	24,205,000	24,865,000	24,465,000	0.344
12	1,125,000	1,455,000	1,405,000	0.250
12	6,855,000	7,505,000	7,345,000	0.226
12	8,835,000	9,315,000	9,205,000	0.433
12	14,705,000	15,405,000	15,065,000	0.276
12	16,405,000	16,845,000	16,665,000	0.456
12	20,625,000	20,975,000	20,795,000	0.318
12	21,975,000	22,495,000	22,085,000	0.227
13	20,285,000	20,945,000	20,585,000	0.244
17	**18,215,000**	**19,435,000**	**19,135,000**	**0.269**
18	12,895,000	14,315,000	13,695,000	0.219
19	2,745,000	3,715,000	3,305,000	0.264
20	48,165,000	49,135,000	48,765,000	0.240
21	4,735,000	5,615,000	5,065,000	0.247
21	9,075,000	9,845,000	9,485,000	0.204
22	5,175,000	5,815,000	5,445,000	0.532
23	14,645,000	15,615,000	15,255,000	0.232
23	54,175,000	55,255,000	54,705,000	0.230
25	10,875,000	11,305,000	11,075,000	0.236
25	36,965,000	37,495,000	37,125,000	0.231
26	27,845,000	28,325,000	28,075,000	0.400
26	30,045,000	30,525,000	30,225,000	0.259
27	31,445,000	32,045,000	31,615,000	0.197
28	**14,395,000**	**15,095,000**	**14,745,000**	**0.806**
28	**16,535,000**	**17,805,000**	**16,915,000**	**0.284**
29	3,405,000	3,815,000	3,615,000	0.343
30	**2,815,000**	**3,135,000**	**3,105,000**	**0.199**
30	14,285,000	14,675,000	14,285,000	0.251
30	25,045,000	25,595,000	25,295,000	0.277
31	6,005,000	6,335,000	6,155,000	0.323
31	7,115,000	7,455,000	7,235,000	0.259

Supplementary Table S35. Genomic regions candidate for selective sweeps, in the comparison between the Przewalski's horse and 5 domestic horse breeds

Start and end refers to the external coordinates for the regions defined by two valleys of the $\hat{\theta}_w$ log-ratios cubic spline. The columns "Peak" and y_{mean} corresponds to the genome coordinate of the peak of the $\hat{\theta}_w$ log-ratio and the local mean within the region respectively. Regions that also show a significantly decreased Tajima's D value are reported in bold.

Chromosome	Start	End	$y_{mean(\hat{\theta}_w)}$	$y_{mean(D)}$
9	70,625,000	71,935,000	0.465	-0.719
17	18,215,000	19,435,000	0.269	-0.935

28	14,595,000	15,075,000	0.806	-0.904
28	16,985,000	17,555,000	0.284	-0.737
30	2,815,000	3,135,000	0.199	-0.616

Supplementary Table S36. Truncated genomic regions for selective sweeps delimited by $\hat{\theta}_w$ valleys and Tajima's D peaks, in the comparison between the Przewalski's horse and 5 domestic horse breeds

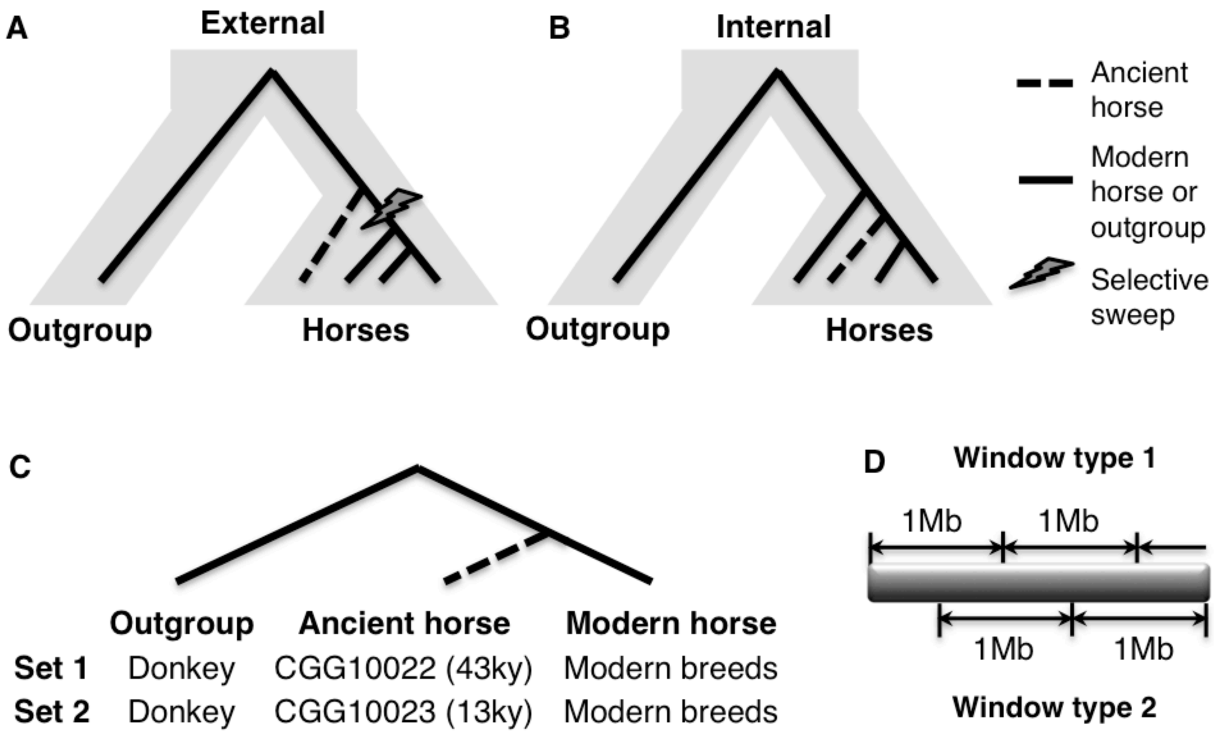
*Candidate genomic regions were selected based on the presence of an increased the $\hat{\theta}_w$ log-ratios, overlapping a region with a decreased Tajima's D (*Supplementary Table S35*); these were truncated to just the region of overlap between the two measures.*

S4.3 Selective sweep scan with coalescent and SNP BeadChip genotypes

S4.3.1 Background

To search for signatures of selection in the horse genomes, putatively affected by early domestication, we modified an approach developed by Green et al. (2010) and used to identify signatures of selective sweeps in the early stage of human evolution, using Neanderthal and modern human genomes [see their Supplementary Section SOM13] (24). This approach examines the timing of coalescence of domestic and pre-domesticated alleles at each locus. If strong artificial selection targeted a specific region in the early stages of domestication, alleles of domestic horses should coalesce within the domestic clade more recently than the coalescence of ancient alleles, locating the pre-domesticated horses basal to the domesticated horses, and resulting in what we term the 'external' topology (illustrated in Supplementary Figure S47A). Importantly, the ancient horse genome will contain substantially reduced numbers of derived alleles (polarized relative to the outgroup) in regions under selection, and the lengths of these regions will be positively correlated with strengths of selection. Alternatively, if genomic regions are evolving neutrally, ancient and domestic alleles may be sorted incompletely, potentially resulting in what we term the 'internal' topology (Supplementary Figure S47B). For neutral loci, derived allele frequencies in domestic horses should be positively correlated with those in ancient horses, while genomic regions that exhibit continuous distributions of 'external' states are likely to be affected by domestication. To find these external regions, we fit a parametric normal distribution to the genome-wide differences between observed and expected derived allele frequencies in ancient horses, and seek outliers of this distribution.

We made use of the publicly available genotypes of more than 400 horses of 32 modern breeds at 54,602 SNP loci, genotyped with the Illumina EquineSNP50 Genotyping BeadChip (38, 39), from the NAGRP Community Data Repository for Livestock Animal Genomics. We also used whole genome shotgun sequences of the samples described in Supplementary Section S1.1, excluding the Przewalski's horse, and using the domestic donkey (Willy) as the outgroup.

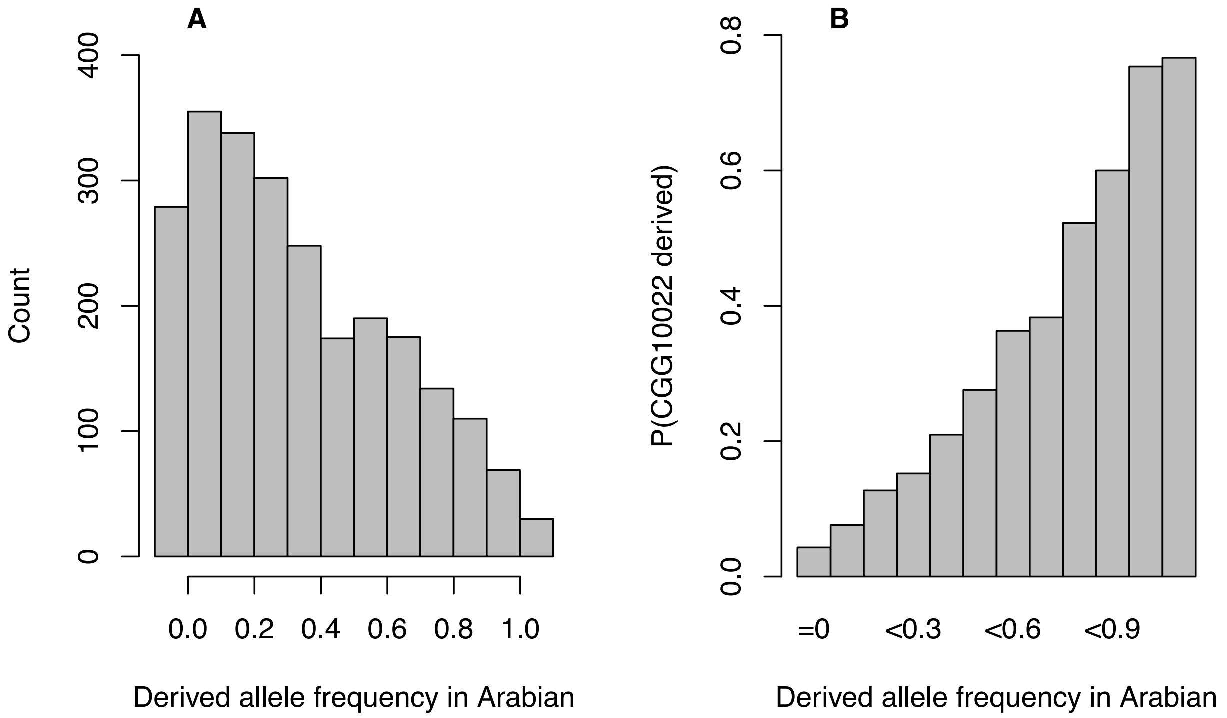


Supplementary Figure S47. Study design of the selective sweep scans with SNP Beadchip array data

*(**A-B**) Coalescence of ancient and modern horses based on (**A**) alleles affected by selective sweeps after divergence of the pre-domestication breed and (**B**) neutral alleles. Dashed lines represent ancient horses and the grey area represents the two lineages of outgroup species and horses. (**C**) Two sets of selective sweep scans that are composed of different combinations of ancient horse individuals (CGG10022 and CGG10023) and all 32 modern breeds. (**D**) Illustration of the two types of 1 Mb windows used in selection scans for a stretch of chromosome; type 2 windows are produced by shifting the coordinates of type 1 windows by 500kb.*

S4.3.2 Data processing and model training

We 'haploidized' whole genome shotgun sequencing data of pre-domesticated horses (CGG10022 and CGG10023) and the outgroup with the program '*pu2fa*' developed by Green *et al.* (24). This program picks one random allele at each site that passes a set of quality filter and consequentially transforms a diploid genome into a haploid genome. We extracted bases at the 54,602 positions included in the SNP array from the resulting haploid genomes. We then selected SNPs that were biallelic and autosomal, and polarized them by designating the allele in outgroups as ancestral. To minimize the interference of ancient DNA damage on the final results, which manifests mainly as C→T/G→A transitions (118), we divided SNPs into different transitions/transversion categories and modeled them individually. We estimated the conditional probability of derived alleles in ancient horses $P(ancient\ derived|modern\ derived)$ empirically by estimating the proportion of loci in each frequency bin of domestic horses that also possess derived alleles in ancient horses for each category of SNPs (Supplementary Figure S48).



Supplementary Figure S48. Model training

***A:** Derived allele frequency spectrum of the modern breed. We use an example of the Arabian breed at SNP loci with A→G transitions. **B:** Probabilities of derived alleles of the ancient horse conditioning on the derived allele frequencies in modern breeds. We use the sample combination of the Domestic donkey (Willy), CGG10022 and Arabian horse at the A→G SNP transition type as an example.*

S4.3.3 Selection scan setup

We divided the horse genome into non-overlapping windows of 1 Mb (this set of windows constitutes what we call hereafter type 1 windows; Supplementary Figure S47D), which include, on average, 23 SNPs. We estimated differences (D_i) between observed frequency of derived alleles (o_i) and the expected frequency of derived alleles in the ancient horse $E(P(ancient\ derived))_i$ under no selection with the following equations:

For *n* numbers of SNP loci in window *i* of 1 Mb:

$D_i = o_i - E\big(P(ancient\ derived)\big)_i$, in which

$o_i = \sum_{k=1}^{n} Bernoulli(ancient\ derived)$,

$E\big(P(ancient\ derived)\big)_i = \sum_{k=1}^{n} P_k\ (ancient\ derived)$,

and P_k*(ancient derived)* at each site was estimated from the respective empirical distribution of the transition/transversion type as illustrated in Supplementary Figure S48b.

Finally, we z-transformed all D_i, quantified the importance of every region by their z-scores, and designated the top 1% of regions (with z-scores less than -2.325) as candidate regions. We performed selective sweep scans on two sets of data; each scan includes combination of one ancient horse and one modern breed (e.g. a scan of the Domestic Donkey (Willy), CGG10022, Akhal Teke) (Supplementary Figure S47C). We repeated the two sets of analyses by moving each window by 500kb to accommodate any signature of selective sweeps that intercepts with two neighboring windows (this set of windows constitutes type 2 windows; Supplementary Figure S47D). Regions above the 99th and 95th percentiles of all comparisons, or regions that are repetitively recovered from scans with both ancient horses are candidates for artificial selection.

S4.3.4 Gene annotation and enrichment

We annotated candidate regions in the top 1% of every analysis with Ensembl 72, and recorded the Ensembl gene ID, genomic coordinates, associated gene names and biostatus of genes that overlap with these candidate regions. Then we performed enrichment analyses of these genes with DAVID, using horses as the background and the Benjamini-Hochberg correction to correct multiple testing. We also performed the same type of enrichment analyses on the human orthologs of these annotated genes.

S4.3.5 Results

Among all 301 candidate regions above the 99th percentiles of all sets of analyses with both window types, four regions are in the top 1% of more than 30 scans of modern breeds in each set of sample combinations (Supplementary Table S37), while 38 regions are ranked as the top 5% in all 32 scans of each set (Supplementary Table S38). SNPs in the top 1% candidate regions rarely possess the derived alleles in the ancient horses, but show moderate to high frequencies of derived alleles in modern breeds comparing with their neighboring regions (Supplementary Figure S49).

Candidate regions from selection scans with two pre-domesticated horses (182 regions for CGG10022 and 181 regions for CGG10023) overlap at 62 regions, a number of overlapping regions that is significantly smaller than random chance expectations (Monte Carlo simulation *P*-value < 0.01). Among the 62 overlapping regions, 29 regions were recovered from analyses with type 1 windows and the other 33 from analyses of the type 2 windows (Supplementary Table S39).

Gene ontology analyses of annotated genes in all candidate regions against the horse background revealed the enrichment of neuroactive ligand-receptor interaction (*P*-value=0.015), lysosome (*P*-value=0.023) and focal adhesion pathways (*P*-value=0.042), but not significant after correction for multiple testing. Consistent with results in Orlando *et al.* 2013 (1), limited enrichment of specific functional categories implies that no specific function or phenotype was particularly favored in the domestication process of all horse breeds.

However, our results suggested several genes that may be subject to strong selection. For example, two genes known to be affected by horse domestication have showed up in our candidate regions: *KIT* (Supplementary Figure S50), and *MC1R* (Supplementary Figure S51), the latter of which was also detected using the $\hat{\theta}_w$ log-ratio scans (Supplementary Section S4.2). The candidate region containing the *KIT* gene, chr3: 77,000,000-78,000,000, is above 99 percentiles in Clydesdale, and is in the top 5% in the analyses with Saddlebred. The region containing the *MC1R* gene, chr3: 35,500,000-36,500,000, is the top 1% candidate for Belgian and Morgan breeds, and is above the 95th percentile for analyses with another 18 modern breeds (Akhal Teke, Arabian, Caspian, Finnhorse, French Trotter, Icelandic, Manglarga Paulista, Miniature, Mongolian, New Forest Pony, Paint, Peruvian Paso, PR Paso Fino, Quarter Horse, Saddlebred, Shire, Swiss Warmblood and Tuva). Discovery of these genes supports the power of our approach in detecting genes targeted by artificial selection.

Chr	Start	End	Set1	Set2	Genes

22	9,000,000	10,000,000	32	0	No annotated genes
9	74,000,000	75,000,000	31	0	*ST3GAL1, ZFAT, ENSECAG00000005903*
1	128,500,000	129,500,000	0	32	*HERC1, FBXL22, USP3, CA12, APH1B, RAB8B, RPS27L, TPM1, TLN2*
3	52,500,000	53,500000	0	30	*WDFY3, CDS1, NKX6-1, OOEP*

Supplementary Table S37. Regions in the top 1% of more than 30 breeds among all sets of comparisons of two window types

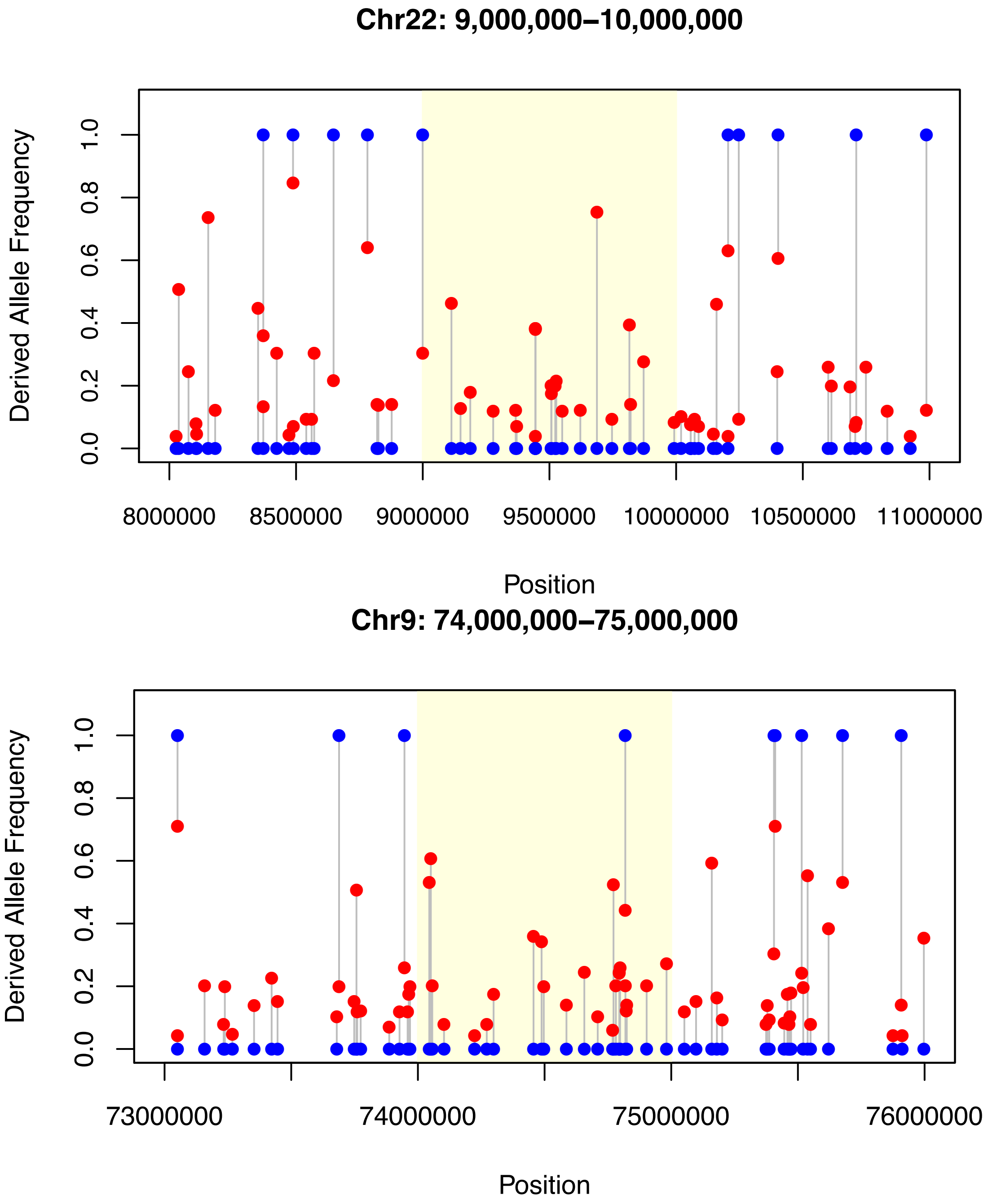
As defined in Supplementary Figure S47*C, the selection scan of set 1 makes use of donkey, CGG10022 and 32 modern breeds, while set 2 examines donkey, CGG10023 and 32 modern breeds. The two alternative window types of 1Mb are illustrated in* Supplementary Figure S47*D.*

Chr	**Start**	**End**	**Set1**	**Set2**	**Genes**
1	95,500,000	96,500,000	-	32	*AGBL1, ENSECAG00000004989, ENSECAG00000004887*
1	127,000,000	128,000,000	14	32	*PDCD7, KIAA0101, IGDCC3, TRIP4, ENSECAG00000024668, ENSECAG00000024130, PLEKHQ1, OAZ2, ANKDD1A, CLPX, RBPMS2, CILP, SPG21, MTFMT, SLC51B, PARP16, RASL12, CSNK1G1, PIF1, KBTBD13*
1	128,500,000	129,500,000	-	32	*HERC1, TLN2, CA12, USP3, FBXL22, TPM1, RPS27L, RAB8B, APH1B*
2	60,000,000	61,000,000	32	22	*GLRA3, ENSECAG00000023845, DEFB131, ADAM29, ENSECAG00000002168, DEFB136*
2	75,500,000	76,500,000	-	32	*C4orf46, FAM198B, TMEM144, RXFP1, ETFDH, PPID, FNIP2, C4orf45, ENSECAG00000003376*
3	8,500,000	9,500,000	-	32	*CPNE2, AMFR,ENSECAG00000015275, MT3, NLRC5, SLC12A3, MT4, RSPRY1, GNAO1, HERPUD1, BBS2, FAM192A, OGFOD1, NUP93, ENSECAG00000001555, NUDT21, ENSECAG00000000368, ENSECAG00000000363, ENSECAG00000000358*
3	52,500,000	53,500,000	-	32	*ENSECAG00000023821, NKX6-1, WDFY3, CDS1, OOEP*
4	80,000,000	81,000,000	-	32	*DRG1, POT1, GPR37*
5	11,500,000	12,500,000	32	28	*PAPPA2, FAM5B, ENSECAG00000024280, ASTN1*
7	10,000,000	11,000,000	-	32	*CNTN5*
7	10,500,000	11,500,000	-	32	*PGR, ARHGAP42, CNTN5*
7	70,500,000	71,500,000	32	-	*ENSECAG00000022725, LAMTOR1, LRTOMT, ANAPC15, ENSECAG00000020411, ENSECAG00000019909, PDE2A, ENSECAG00000018619, INPPL1, PHOX2A, ARAP1, STARD10, ENSECAG00000015031, ATG16L2, CLPB, FCHSD2, P2RY2, ARHGEF17*
7	79,500,000	80,500,000	-	32	*GALNT18, EIF4G2, CTR9, MRVI1, LYVE1, RNF141, ZBED5, AMPD3*
8	60,500,000	61,500,000	-	32	*No annotated genes*
9	74,000,000	75,000,000	32	-	*ZFAT, ST3GAL1, ENSECAG00000005903*
10	39,000,000	40,000,000	-	32	*ZNF292, CG, HTR1E, ENSECAG00000003407*
10	55,000,000	56,000,000	32	27	*PREP, ENSECAG00000021367, LIN28B, HACE1, ENSECAG00000008961*
11	11,500,000	12,500,000	32	-	*FAM20A, ARSG, ABCA9, ENSECAG00000022227, ABCA8, GNA13, AMZ2, SLC16A6, ABCA5, ABCA6, WIPI1, ENSECAG00000002254, PRKAR1A, ABCA10*
11	33,000,000	34,000,000	32	-	*CLTC, DHX40, ENSECAG00000018590, YPEL2, ENSECAG00000017636, GDPD1, SMG8, PRR11, ENSECAG00000015007, RPS6KB1, TUBD1, TRIM37, PPM1E, VMP1, PTRH2, ENSECAG00000003590*
11	34,500,000	35,500,000	20	32	*TBX2, TBX4, C17orf64, APPBP2, PPM1D, BCAS3,*

					ENSECAG00000003698
11	35,000,000	36,000,000	10	32	*MRM1, DHRS11, GGNBP2, MYO19, ZNHIT3, CA4, USP32, C17orf64, APPBP2, PPM1D, PIGW, ENSECAG00000003729*
13	8,000,000	9,000,000	32	-	*TFR2, ZNF3, MEPCE, MOSPD3, LRCH4, SAP25, ENSECAG00000022553, ZCWPW1, CNPY4, PCOLCE, ENSECAG00000021110, PVRIG, SLC12A9, GIGYF1, GNB2, ENSECAG00000016429, AGFG2, AP4M1, ENSECAG00000015672, FBXO24, MCM7, ACTL6B, MUC12, NYAP1, ENSECAG00000012736, SRRT, MUC3A, EPO, EPHB4, POP7, ENSECAG00000010102, COPS6, TAF6, TSC22D4, GPC2, C7orf61, ACHE, UFSP1, LAMTOR4, MBLAC1, PPP1R35, GAL3ST4, ENSECAG00000002581, TRIP6, ENSECAG00000001032*
14	22,000,000	23,000,000	32	-	*SGCD, HAVCR1, HAVCR2, FAM71B, ITK, ENSECAG00000004332, MED7*
14	68,500,000	69,500,000	-	32	*FAM174A, ST8SIA4, ENSECAG00000002396, ENSECAG00000002335*
15	25,500,000	26,500,000	32	-	*ENSECAG00000002005*
15	38,500,000	39,500,000	32	-	*FAM161A, CCT4, COMMD1, B3GNT2, TMEM17, EHBP1, ENSECAG00000002836*
15	41,500,000	42,500,000	32	-	*ENSECAG00000011245*
16	34,500,000	35,500,000	32	32	*ITIH1, ENSECAG00000024754, ENSECAG00000023559, PRKCD, RFT1, SEMA3G, MUSTN1, BAP1, TNNC1, PHF7, ENSECAG00000018596, NEK4, NT5DC2, GNL3, ALAS1, TLR9, WDR82, GLYCTK, PPM1M, STAB1, NISCH, SFMBT1, ENSECAG00000006260, TKT, ENSECAG00000004579, ITIH3, ITIH4, PBRM1, DNAH1, GLT8D1*
16	86,500,000	87,500,000	32	31	*ENSECAG00000024575, GMPS, SLC33A1, C3orf33, PLCH1, MME, ENSECAG00000001897*
17	6,500,000	7,500,000	32	-	*ENSECAG00000021095, LNX2, MTIF3, GTF3A, RASL11A, ENSECAG00000010661, USP12, POLR1D, ENSECAG00000002829, GPR12, WASF3*
17	55,500,000	56,500,000	32	-	*ENSECAG00000002806, ENSECAG00000002767, ENSECAG00000002737, SLITRK1*
18	66,000,000	67,000,000	32	-	*INPP1, ASNSD1, HIBCH, SLC40A1, MSTN, PMS1, ORMDL1, OSGEPL1, NAB1, TMEM194B, MFSD6, ANKAR, C2orf88*
19	50,500,000	51,500,000	32	-	*ALCAM, CBLB*
20	61,000,000	62,000,000	32	-	*LMBRD1, COL9A1, BAI3, COL19A1*
21	16,500,000	17,500,000	-	32	*DHX29, SKIV2L2, DDX4, GZMA, GPX8, PPAP2A, IL31RA, CDC20B, MCIN, CCNO, SLC38A9, ESM1, GZMK*
22	9,000,000	10,000,000	32	-	*No annotated genes*
24	21,500,000	22,500,000	32	12	*ANGEL1, VASH1, TGFB3, ESRRB, KIAA1737, IRF2BPL, GPATCH2L, C14orf166B, IFT43*
30	25,000,000	26,000,000	-	32	*DENND1B, NEK7, CRB1, LHX9*

Supplementary Table S38. Regions found in the top 5% of all 32 breeds of all sets of comparisons with two different window arrangements

Overlapped regions are not merged. Chr: chromosome number.



Supplementary Figure S49. Distribution of the expected frequency of derived alleles and the observed number of derived alleles in the ancient horse

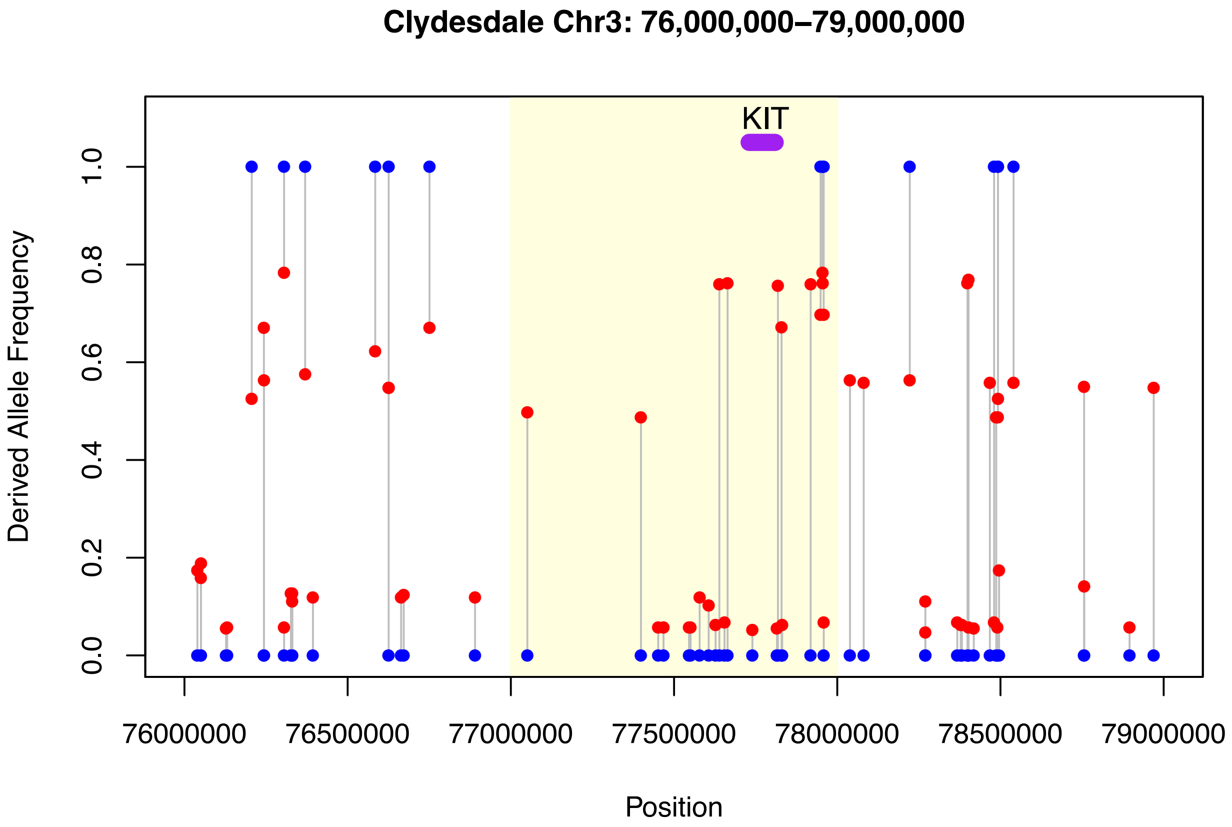
Chromosome 22: 8,000,000-11,000,000 and chromosome 9: 73,000,000-76,000,000 in the sample combination of the domestic donkey (Willy), CGG10022 and the modern breed Arabian as examples. The yellow shaded area is the top 1% candidate region (see Supplementary Table S38 *for detailed information of these two regions). Red dots represent the expected frequency of derived alleles in CGG10022, and blue dots denote the observed frequency of derived alleles in CGG10022 at each site.*

Chr	Region Start	Region End	Set 1		Set 2	
			1%	5%	1%	5%
1	0	1,000,000	8	26	1	13
1	67,000,000	68,000,000	6	20	7	19
1	67,500,000	68,500,000	22	30	16	25
1	112,000,000	113,000,000	1	6	2	9
1	112,500,000	113,500,000	1	3	1	1
1	127,000,000	128,000,000	1	14	23	32
2	27,500,000	28,500,000	7	24	22	31
2	40,500,000	41,500,000	2	6	2	5
2	42,500,000	43,500,000	1	1	1	2
2	60,000,000	61,000,000	6	32	1	22
2	100,000,000	101,000,000	6	14	2	8
2	100,500,000	101,500,000	3	9	5	12
2	104,000,000	105,000,000	5	17	3	16
3	33,500,000	34,500,000	1	9	4	27
3	35,500,000	36,500,000	1	20	2	20
3	39,000,000	40,000,000	1	15	7	30
3	77,000,000	78,000,000	1	2	1	2
3	77,500,000	78,500,000	1	5	1	2
3	100,000,000	101,000,000	1	12	1	4
3	100,500,000	101,500,000	1	14	1	14
3	118,500,000	119,500,000	5	10	5	10
4	35,000,000	36,000,000	1	8	2	10
4	54,500,000	55,500,000	6	24	1	10
4	55,000,000	56,000,000	11	23	2	12
4	65,500,000	66,500,000	3	29	1	5
5	0	1,000,000	4	20	5	22
5	11,500,000	12,500,000	22	32	3	28
5	12,000,000	13,000,000	2	13	9	28
5	50,500,000	51,500,000	1	5	5	24
6	17,500,000	18,500,000	2	19	12	30
7	39,500,000	40,500,000	1	7	10	18
7	41,000,000	42,000,000	2	4	3	6
7	41,500,000	42,500,000	19	30	3	11
7	75,000,000	76,000,000	1	2	1	2
8	85,500,000	86,500,000	5	17	5	21
9	43,500,000	44,500,000	7	18	16	31
9	44,000,000	45,000,000	9	27	10	28
10	6,500,000	7,500,000	5	20	6	20
10	55,000,000	56,000,000	15	32	2	27
11	22,500,000	23,500,000	7	16	7	22
11	23,000,000	24,000,000	5	9	9	23
11	31,000,000	32,000,000	11	29	12	29

11	34,500,000	35,500,000	5	20	14	32
11	35,000,000	36,000,000	1	10	28	32
11	60,500,000	61,500,000	1	9	1	10
13	41,500,000	42,500,000	4	20	5	20
14	41,000,000	42,000,000	14	31	2	8
15	53,000,000	54,000,000	1	2	1	25
15	53,500,000	54,500,000	2	13	2	10
16	34,000,000	35,000,000	11	30	2	17
16	34,500,000	35,500,000	29	32	9	32
16	70,000,000	71,000,000	4	29	6	31
16	86,500,000	87,500,000	28	32	18	31
16	87,000,000	88,000,000	17	29	19	31
21	15,500,000	16,500,000	7	18	7	16
21	16,000,000	17,000,000	10	25	1	8
21	45,500,000	46,500,000	4	14	4	16
21	46,000,000	47,000,000	2	8	3	8
23	53,500,000	54,500,000	1	3	1	4
23	54,000,000	55,000,000	12	27	26	31
24	14,500,000	15,500,000	1	11	1	11
24	21,500,000	22,500,000	25	32	1	12

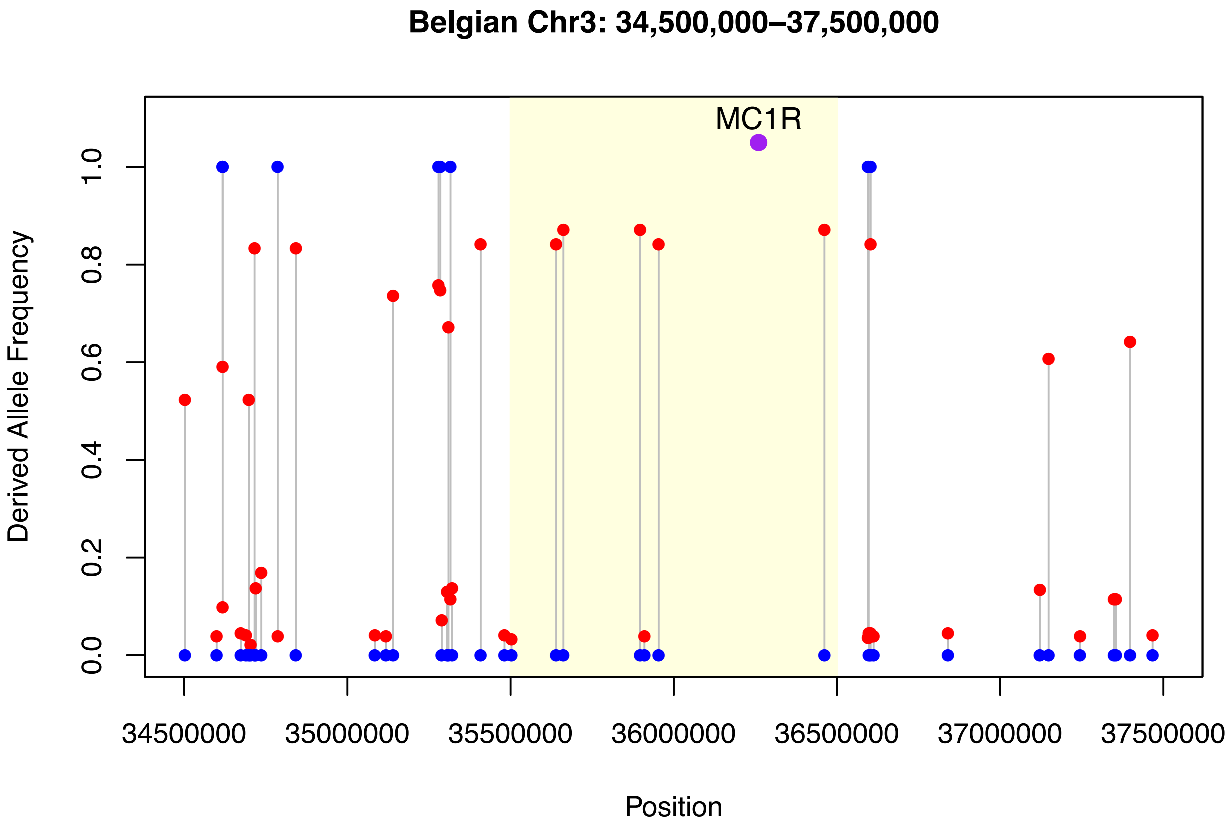
Supplementary Table S39. Regions found in the top 1% of both sets of comparisons

Chr: Chromosome. Overlapping regions of two window types are not merged. The numbers for each set tabulates the number of breeds detected at a given threshold. Definitions of sets 1 and 2 are provided in Supplementary Figure S47*.*



Supplementary Figure S50. A candidate region (highlighted in yellow) enclosing the KIT gene in the Clydesdale breed

Red dots represent the expected frequency of derived alleles in CGG10022, and blue dots denote the observed frequency of derived alleles in CGG10022 at each site. Coordinates for the KIT gene are Chr3: 77,730,011 - 77,809,756; this region is highlighted by a purple horizontal bar.



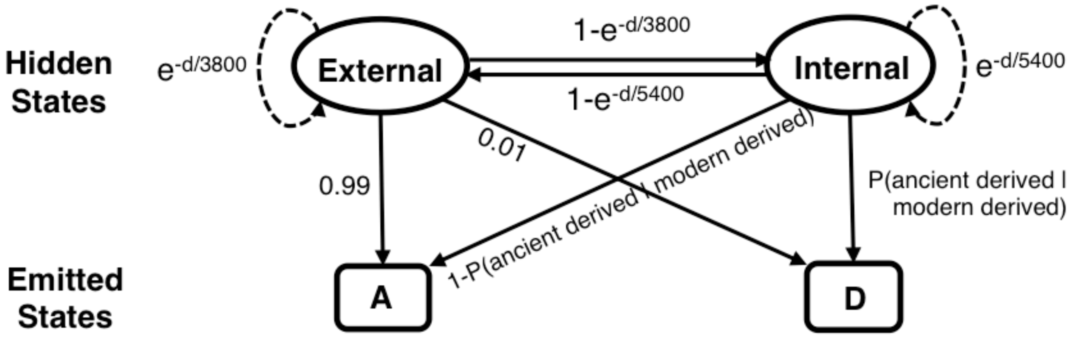
Supplementary Figure S51. A candidate region (highlighted in yellow) enclosing the MC1R gene in the Belgian breed

Red dots represent the expected frequency of derived alleles in CGG10022, and blue dots denote the observed frequency of derived alleles in CGG10022 at each site. Coordinates for the MC1R gene are Chr3: 36,259,276-36,260,354, and this gene is highlighted by a purple dot in the figure.

S4.4 Selective sweep scans using a Hidden Markov Model

S4.4.1 Background

We used a Hidden Markov Model (HMM) to identify signatures of selection in genes or regions of the horse genome, based on the genetic variation of modern horse genomes and the state of the pre-domesticated genome. This model is modified from the HMM model developed by Prüfer *et al.* (119), which identified genomic regions under selection in chimpanzees in comparison with the Bonobo. The horse genome is viewed as a sequential Markov chain of external and internal states (as defined in Supplementary Figure S47) going along the genome sequence as defined in Supplementary Figure S52. The state of each locus depends on those of the neighboring loci because of linkage and recombination, and such a relationship is modeled by 'transition probabilities'. The state at each locus, though hidden, emits observable states, i.e. alleles of each locus being derived (D) or ancestral (A) in modern horses, with certain probabilities (termed as 'emission probability'). This Hidden Markov Model, combined with the coalescent theory, allows for inference of the most probable sequential distribution of external and internal states in the horse genome. Regions of extended and continuous external states represent candidates for artificial selection.



Supplementary Figure S52. Diagram of the Hidden Markov Model

Values '3800' and '5400' in the equations of transition probabilities are the average lengths of external and internal regions estimated by coalescent simulations. Details are described in the Supplementary Section S4.4.3.

S4.4.2 Initial data processing

We haploidized the complete genome sequences of six modern horses (Arabian, Norwegian Fjord, Icelandic (unnamed), Icelandic (P5782), Standardbred and Thoroughbred (Twilight)) and the outgroup the Domestic donkey (Willy) following the same approach as described in Supplementary Section S4.3.2. We discovered SNPs from 'haploidized' genomes of six modern horses, and retrieved bases at these positions from the 'haploidized' genomes of the outgroup species.

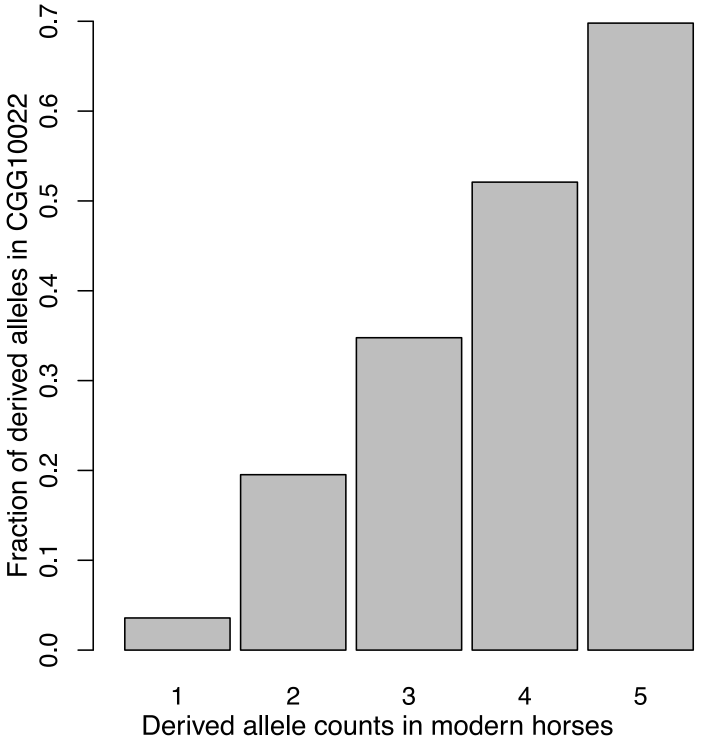
'Haploidization' of the ancient horse genome leads to inclusion of DNA damage in the final dataset; to avoid this we performed genotype calling on CGG10022 by GATK UnifiedGenotyper (14) with the option of 'emit all sites', obtained genotypes at the SNP positions, and filtered these genotypes by read depths (DP) and genotype qualities (GQ) (8 ≤ DP ≤ 59 and GQ ≥ 20). The cutoff of DP and

GQ were determined from genome-wide distributions of these two variables – less than 95% of positions possess GQ ≥ 20 and about 85% of SNP positions possess DP between 8 and 59. Then we took a random base at heterozygous sites that passed filters at each position, and consequentially haploidized the genome of CGG10022. We combined SNPs of outgroups and horses, filtered them with the criteria of a) autosomal, b) bi-allelic, c) with known alleles in all horses, d) with known ancestral alleles in donkey Willy, and polarized these SNPs by viewing the allele in donkey as ancestral and the alternative one in modern horses as derived.

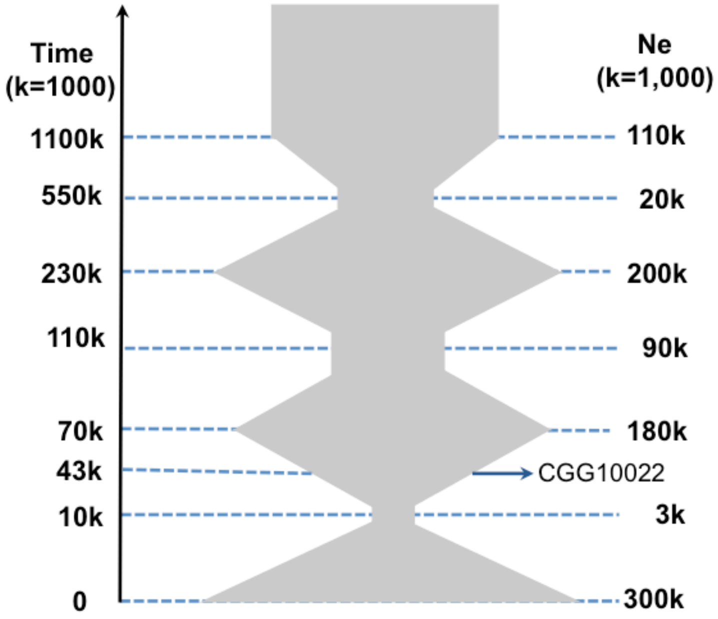
S4.4.3 Model training and coalescent simulation

For the HMM model we trained two sets of parameters; emission probabilities of internal states and transition probabilities between the internal and external states (Supplementary Figure S52). We estimated percentages of derived alleles in the ancient horse CGG10022 conditioned on derived allele frequencies of modern horses, $P(ancient\ derived|modern\ derived)$ (Supplementary Figure S53), and used this function as the emission probabilities of derived alleles given the hidden state of 'internal'.

We used an exponential distribution, a memory-less function, to model the transition probabilities between hidden states, $P(d|\beta) = 1 - e^{-d/\beta}$, where *d* is the physical distance (quantified in base pairs) between two adjacent SNPs. The parameter of the exponential distribution, β, equals to the expected values of d, i.e. average lengths of external and internal regions. We performed simulations with the coalescent simulator *fastsimcoal2* (92) to estimate lengths of external and internal regions under the null hypothesis of neutrality. We used a scenario of horse demography (Supplementary Figure S54) inferred from results of PSMC (Supplementary Figure S28) and previous studies of horse mitochondrial haplotypes (91). We assumed that the mutation rate is 7.242×10^{-9} per generation per site (with a generation time of 8 years), and that the recombination rate is 1cM/Mb (1, 91, 120). We simulated 100 10 Mb regions under this scenario, sampled one individual from the simulated outgroup, one from the ancient horse 43 kyr BP and six from the modern horses, determined states (external or internal) of every recombination block based on the phylogenetic relationships of ancient and modern horses at that region. We estimated average lengths of external (3,883 base pairs, rounded down to 3,800 for model simplicity) and internal regions (5,407 base pairs, rounded down to 5,400) in the simulated datasets. All trained parameters for transition and emission probabilities are illustrated in Supplementary Figure S52.



Supplementary Figure S53. Conditional probabilities of derived alleles in CGG10022 given the numbers of derived alleles in modern horses

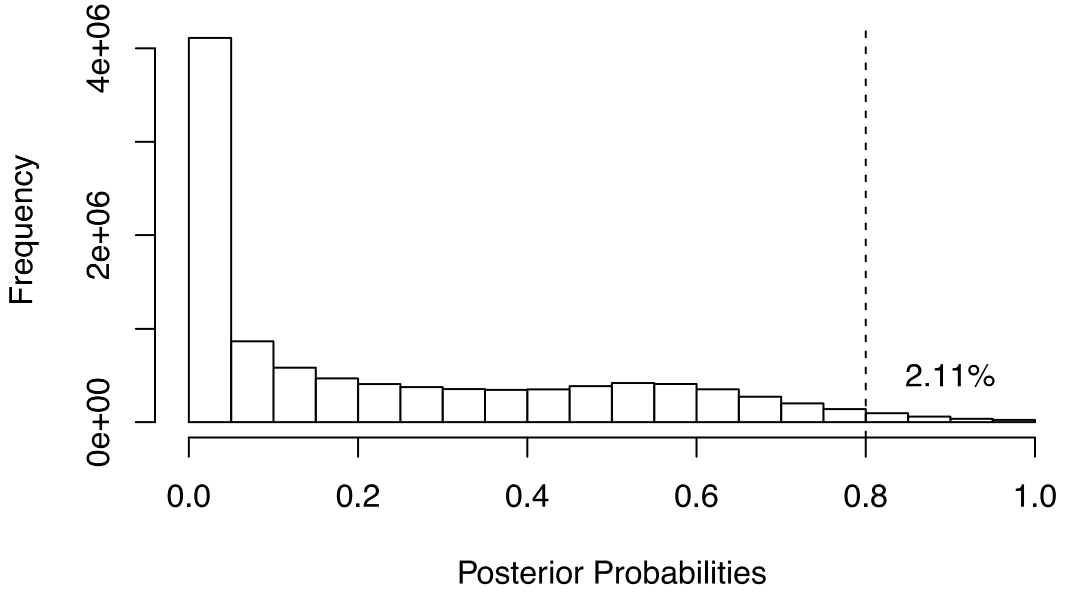


Supplementary Figure S54. Horse demography used for simulation

Timing of population expansion/bottleneck in the past is listed on the left next to the time arrow (from present to the past) and the population sizes (Ne) are shown on the right. One ancient horse (CGG10022) was sampled at 43 kyr BP (labeled with an arrow).

S4.4.4 Selective sweep scan with HMM

We used and modified the C++ program developed by Prüfer *et al.* (119), to perform the selection sweep scan with HMM. This program uses the forward-backward algorithm to estimate posterior probabilities of each site being external or internal through forward and backward propagations and scaling (121). We then performed posterior decoding, assigned sites as external/internal with a posterior probability cutoff of 0.8 (external if the posterior probability being in that state is greater than 0.8, and vice versa). We designated regions with at least two adjacent and continuous external SNPs as 'external' regions, and ranked these regions based on their lengths. We annotated genes that overlap with these regions, examined the distribution of external sites in coding sequences, determined non-synonymous and synonymous substitutions with Ensembl Variant Effect Predictor (19), and examined the ancient horse genotype and derived allele frequencies in modern horses at these sites. We also performed gene ontology analyses in DAVID with horse and human as backgrounds (the same approach as described in supplementary section S4.3.4).



Supplementary Figure S55. Histogram of the posterior probabilities of being in an external state

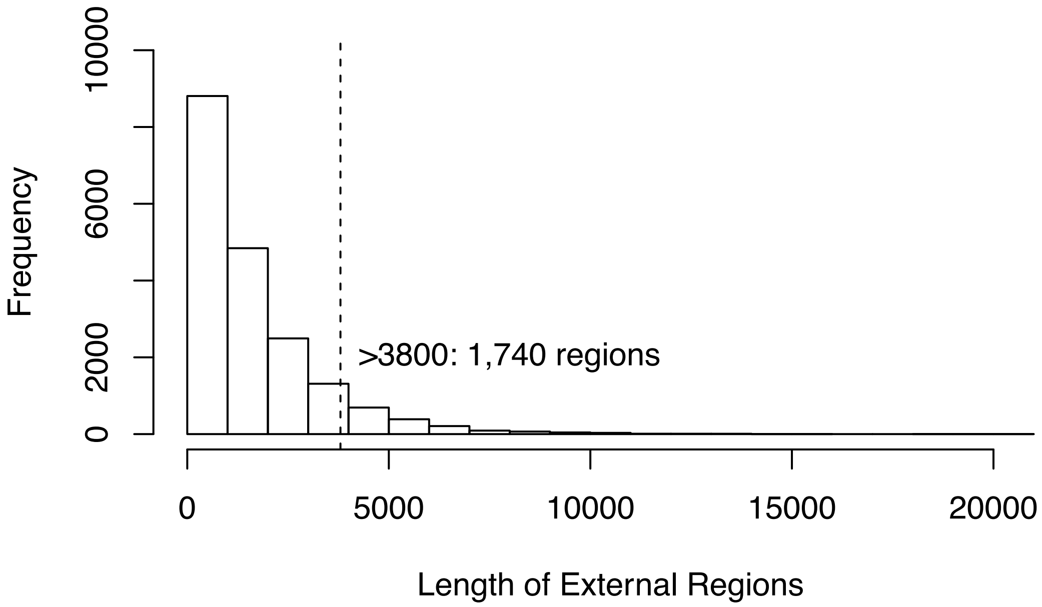
The vertical dashed line represents the probability cutoff of 0.8 for posterior decoding of external sites.

S4.4.5 Results

A total of 10,255,648 SNPs were used in the analyses with the HMM model after discovery, screening and filtering, out of which 216,467 sites were classified as in the 'external' states based on the posterior probability cutoff of 0.8. This cutoff of 0.8 selected about two percent of all SNP sites (Supplementary Figure S55). We found 19,007 external regions, among which 9.15% of these regions are longer than 3.8kb (Supplementary Figure S56). The longest external region is of 20,855 bp and contains 125 SNP loci (Chromosome 15: 88,729,884 - 88,750,738) (Supplementary Table S40).

A total of 262 protein-coding genes annotated in Ensembl v72 (122) overlap with these ‘external’ regions. Functional categories such as calcium signaling pathway (*P*-value=0.0034), gonadotropin releasing hormone signaling pathway (*P*-value=0.0068), arrhythmogenic right ventricular cardiomyopathy (*P*-value=0.017), hypertropic cardiomyopathy (*P*-value=0.025), dilated cardiomyopathy (*P*-value=0.032), MAPK signaling pathway (*P*-value=0.025), colorectal cancer (*P*-value=0.032) and ABC transporters (*P*-value=0.035) are enriched with the horse background by DAVID *(123, 124)*, but none of these functional categories was specifically enriched after the Benjamini correction for multiple tests. The longest ‘external’ region overlaps with the gene *TRAPPC12*, a traffic protein that may be involved in autophagy (125) (Supplementary Table S40; Supplementary Figure S57). The 3rd longest region overlaps with the *CNTN6* gene (Supplementary Figure S57). *CNTN6* encodes contactin 6, a glycosylphosphatidylinositol-anchored neuronal membrane protein that affects the neuro-development (126). Mice with this gene knocked out are lack of motor coordination and balance (127), traits that are essential for farming and equestrian in modern horse breeds. Exon 2 of *CNTN6*, which is highly conserved among outgroup, ancient and modern samples, falls within this external region (Supplementary Figure S57).

A total of 857,416 SNPs fall within exons of horse genes, among which 2.1% (18,165 SNPs) possess posterior probabilities no less than 0.8 and 11.9% (102,446 SNPs) no less than 0.6. A total of 77 genes contain SNP sites at which posterior probabilities are larger than 0.8, substitutions between ancient and modern alleles are non-synonymous, and CGG10022 is homozygous ancestral and all modern horses possess the derived alleles (Supplementary Table S41).



Supplementary Figure S56. Size distributions of external regions

External regions are defined by SNP sites with more than two adjacent and continuous external SNPs.

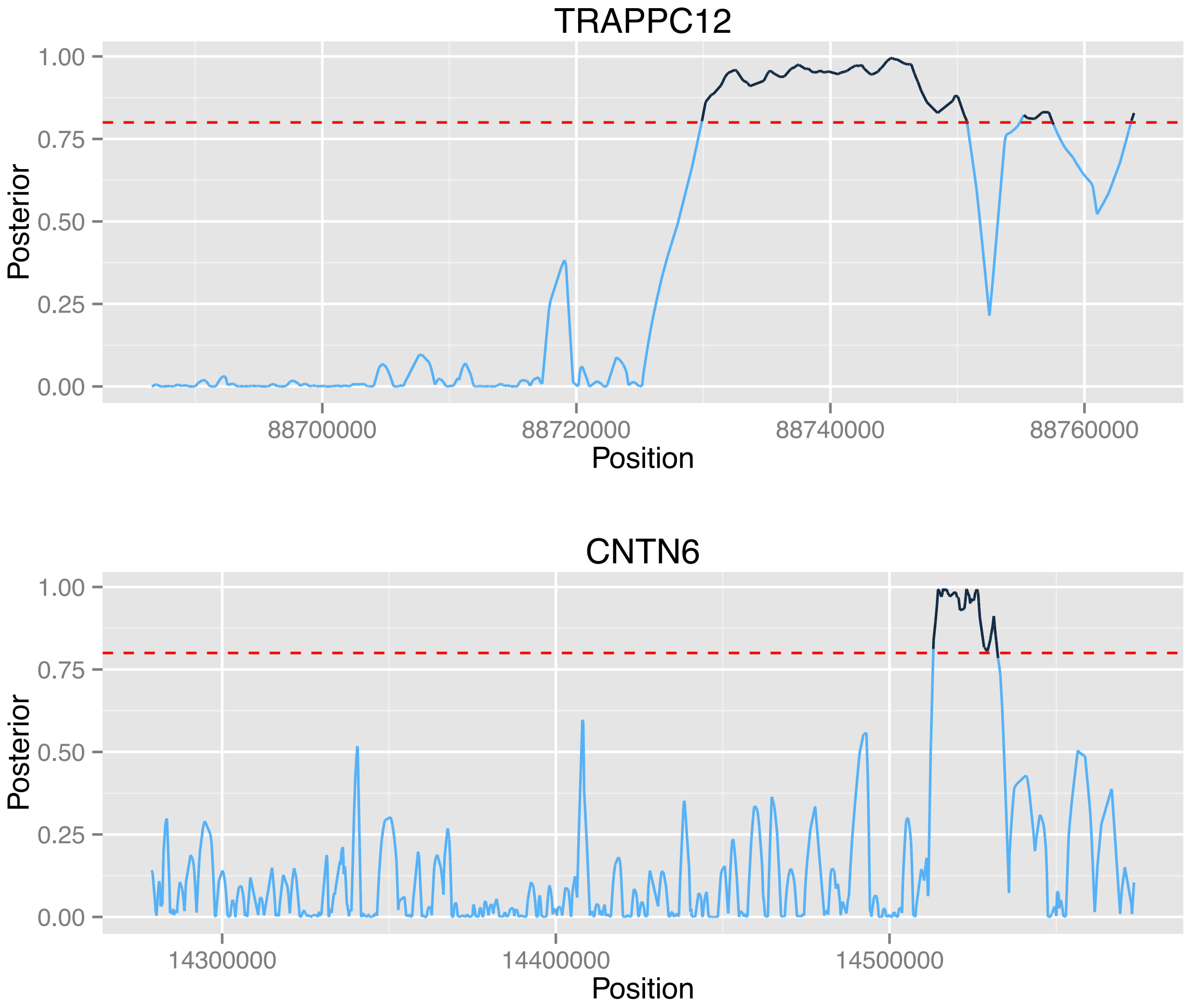
Chr	Start	End	Region Length	Number of SNPs	Gene Name
15	88,729,884	88,750,738	20,855	125	*TRAPPC12*

11	31,453,385	31,472,818	19,434	78	-
16	14513,141	14,531,992	18,852	91	*CNTN6*
7	39,863,290	39,878,613	15,324	83	-
31	9,658,624	9,673,463	14,840	103	*C6orf70*
11	9,742,613	9,756,392	13,780	103	-
18	7,541,178	7,554,733	13,556	51	-
6	31,993,094	32,006,490	13,397	66	-
18	59,619,755	59,633,117	13,363	59	*SSFA2*
17	54,645,639	54,658,758	13,120	53	*ENSECAG00000011378*
1	120,022,080	120,035,159	13,080	50	*C15orf60*
3	47,486,981	47,499,990	13,010	53	-
10	60819,378	60,832,379	13,002	49	-
3	24,328,847	24,341,498	12,652	71	
8	38,601,299	38,613,846	12,548	71	*TMEM241*
7	49,723,029	49,735,459	12,431	61	*S1PR2*
22	22,929,858	22,942,271	12,414	54	*PDRG1*
1	57,626,587	57,638,750	12,164	63	-
8	71,130,295	71,142,367	12,073	48	*DCC*
26	426,210	438,261	12,052	77	-
31	13,699,018	13,711,046	12,029	44	-
5	26,150,505	26,162,506	12,002	57	-
14	33,949,120	33,960,832	11,713	59	-
10	31,065,551	31,077,248	11,698	54	-
1	185,544,480	185,556,090	11,611	59	*NID2*
4	15,595,732	15,607,328	11,597	41	*MYO1G*
3	34,741,949	34,753,531	11,583	39	-
2	69,822,523	69,834,022	11,500	50	-
7	42,482,989	42,494,334	11,346	42	*ENSECAG00000019629*
7	31,634,201	31,645,500	11,300	83	-
14	83,331,068	83,342,202	11,135	45	-
2	50,910,787	50,921,860	11,074	45	*XPO7*
20	28,401,756	28,412,679	10,924	66	-
7	42,184,994	42,195,906	10,913	45	-
16	46,422,485	46,433,355	10,871	45	*SCN11A*
14	41,365,033	41,375,880	10,848	56	-
2	69,071,828	69,082,589	10,762	47	-
19	39,185,459	39,196,191	10,733	43	*PLA1A*
10	43,090,882	43,101,586	10,705	44	-
17	16,939,028	16,949,721	10,694	47	-
20	63,841,718	63,852,400	10,683	81	-
30	4,944,161	4,954,816	10,656	76	*KIF26B*
15	7,002,998	7,013,541	10,544	56	*SLC9A2*
13	28,225,492	28,235,998	10,507	48	*ARL6IP1*
7	39,194,652	39,205,152	10,501	67	-
15	27,567,169	27,577,626	10,458	55	*GCFC2*
4	18,228,390	18,238,802	10,413	61	*ABCA13*
5	74,222,430	74,232,836	10,407	55	-
14	26,283,600	26,293,980	10,381	46	-
15	45,285,072	45,295,414	10,343	43	*EML6*
22	48,245,690	48,255,970	10,281	46	*LAMA5*
9	44,092,487	44,102,731	10,245	38	-
21	9,621,567	9,631,800	10,234	56	-
15	52,180,722	52,190,938	10,217	51	-
1	115,914,913	115,925,115	10,203	39	*WDR61*
1	68,399,204	68,409,347	10,144	41	*ACTA1, NUP133*
2	47,715,889	47,726,030	10,142	36	*PRKCZ*
1	177,662,240	177,672,376	10,137	58	-
20	45,494,600	45,504,713	10,114	40	-

3	34,012,659	34,022,762	10,104	47	*FBXO31*
19	50,460,897	50,470,987	10,091	44	-
7	51,924,403	51,934,449	10,047	39	-
6	59,806,946	59,816,974	10,029	41	*CNTN1*
5	44,780,308	44,790,322	10,015	51	-

Supplementary Table S40. External regions longer than 10 kb and their annotated genes

Chr: chromosome.



Supplementary Figure S57. Genes within the longest external regions

Red dashed lines represent posterior probability cutoff of 0.8. Light blue lines connect sites with posterior probabilities lower than 0.8 and the dark blue lines connect external sites. Coordinates for TRAPPC12 are chr15: 88,686,290 – 88,763,968 and for CNTN6 gene are chr16: 14,278,424 - 14,573,446.

Chr	Position	Anc	Dir	PP	Gene ID	Gene Name	Existing Variation
1	32,196,665	A	G	0.81	ENSECAG00000011973	*SLIT1*	-

1	69,377,219	C	G	0.98	ENSECAG00000016463	*TMEM72*	-
1	145,351,508	C	T	0.81	ENSECAG00000008753		-
2	34,327,524	G	A	0.93	ENSECAG00000020925	*TMCO4*	-
3	28,944,191	G	A	0.94	ENSECAG00000015980	*PKD1L2*	rs68558449
3	64,591,106	C	G	0.80	ENSECAG00000001758	*PROL1*	-
3	76,359,038	A	C	0.88	ENSECAG00000023644		-
3	78,172,372	A	C	0.91	ENSECAG00000018176	*PDGFRA*	-
3	80,073,927	A	G	0.98	ENSECAG00000024870	*CWH43*	-
3	81,294,891	T	C	0.92	ENSECAG00000012518	*ATP10D*	rs68545218
4	49,061,788	G	T	0.82	ENSECAG00000013150	*LRRC72*	-
4	58,126,022	G	C	0.80	ENSECAG00000021767	*HOXA1*	-
4	63,423,815	A	G	0.83	ENSECAG00000022057	*BBS9*	-
4	97,188,977	A	G	0.87	ENSECAG00000007025		rs69508703
5	7,304,333	G	A	0.90	ENSECAG00000012649	*FMO3*	rs69489114
5	12,977,015	C	A	0.92	ENSECAG00000007213	*SEC16B*	-
5	22,947,298	C	A	0.84	ENSECAG00000020648	*PRG4*	-
5	44,466,120	G	C	0.89	ENSECAG00000016108		-
5	51,988,150	A	C	0.94	ENSECAG00000010259	*TRIM45*	-
6	15,587,266	T	C	0.89	ENSECAG00000012814	*TM4SF20*	-
6	62,260,667	C	G	0.94	ENSECAG00000011182	*PUS7L*	-
6	62,260,858	G	A	0.95	ENSECAG00000011182	*PUS7L*	-
6	66,500,093	C	A	0.89	ENSECAG00000007947		-
6	71,862,711	G	A	0.81	ENSECAG00000003338		-
6	72,153,032	C	A	0.83	ENSECAG00000006061	*OR6C65*	-
6	74,855,576	C	T	0.92	ENSECAG00000012166	*LRP1*	-
6	80,161,661	T	G	0.96	ENSECAG00000016235	*C12orf56*	-
7	14,580,827	G	A	0.84	ENSECAG00000008723	*CASP12*	-
7	14,585,594	G	C	0.87	ENSECAG00000008723	*CASP12*	-
7	72,132,267	G	A	0.92	ENSECAG00000007735		-
7	73,725,837	G	A	0.87	ENSECAG00000022991		-
7	75,081,420	A	G	0.80	ENSECAG00000005439		-
7	75,991,240	T	C	0.81	ENSECAG00000002575		-
8	5,425,949	G	C	0.90	ENSECAG00000022168		-
8	5,426,669	T	C	0.90	ENSECAG00000022168		-
8	19,404,659	G	A	0.90	ENSECAG00000009544	*RASAL1*	-
8	29,671,395	T	G	0.95	ENSECAG00000024483	*ANKLE2*	-
8	76,122,206	C	T	0.88	ENSECAG00000016124		-
9	70,972,206	G	C	1.00	ENSECAG00000022355		rs68793394
9	70,972,711	T	C	1.00	ENSECAG00000022355		rs68793401
9	81,970,731	C	T	0.82	ENSECAG00000013129	*TOP1MT*	-
10	7,232,948	G	C	0.90	ENSECAG00000007992	*OVOL3*	-
10	14,844,129	G	T	0.95	ENSECAG00000018247	*ZNF404*	rs68953542
10	14,844,388	C	T	0.96	ENSECAG00000018247	*ZNF404*	-
11	21,935,117	T	C	0.90	ENSECAG00000016177	*KRT24*	rs68807490
11	31,535,971	T	C	0.88	ENSECAG00000009841	*AKAP1*	-
11	37,760,548	G	A	0.87	ENSECAG00000017186	*SLFN5*	-
11	45,861,463	C	T	0.82	ENSECAG00000006519	*SMG6*	-
11	49,746,950	T	A	0.95	ENSECAG00000018406	*CXCL16*	-
12	13,028,368	C	T	0.86	ENSECAG00000007685		-
12	13,693,270	G	A	1.00	ENSECAG00000005518		-
12	13,693,306	A	G	1.00	ENSECAG00000005518		-
12	13,693,615	T	A	0.98	ENSECAG00000005518		-
12	15,802,361	G	A	0.97	ENSECAG00000000882	*OR5R1*	-
12	15,802,362	A	G	0.97	ENSECAG00000000882	*OR5R1*	-
12	15,802,374	G	A	0.96	ENSECAG00000000882	*OR5R1*	-
12	15,802,751	A	G	0.83	ENSECAG00000000882	*OR5R1*	-
12	17,127,298	C	A	0.87	ENSECAG00000004473		-
12	17,545,049	C	T	0.95	ENSECAG00000008427	*SLC43A1*	-

12	27,198,319	C	T	0.85	ENSECAG00000016385	*GPR152*	-
13	8,610,746	C	T	0.81	ENSECAG00000023227	*LRCH4*	rs68916110
13	38,162,743	G	A	0.86	ENSECAG00000019336	*UBN1*	-
14	1,678,865	T	C	0.98	ENSECAG00000016709		-
14	41,749,494	C	G	0.92	ENSECAG00000017207	*C5orf15*	-
14	42,827,327	A	G	0.88	ENSECAG00000022349	*Sept8*	-
14	45,250,148	T	G	0.89	ENSECAG00000023686	*KIAA1024L*	-
14	72,018,840	C	T	0.81	ENSECAG00000003636	*ERAP1*	-
14	89,034,578	A	G	0.87	ENSECAG00000013824	*POC5*	rs68994456
14	89,034,621	A	G	0.87	ENSECAG00000013824	*POC5*	rs68994457
15	55,264,660	G	A	0.89	ENSECAG00000020504	*THADA*	-
15	65,988,896	G	A	0.88	ENSECAG00000024027	*CAPN13*	-
15	77,062,420	T	C	0.91	ENSECAG00000020295	*GEN1*	-
15	82,767,951	T	G	0.99	ENSECAG00000010555	*ATP6V1C2*	-
15	82,768,351	T	C	1.00	ENSECAG00000010555	*ATP6V1C2*	-
16	41,452,836	A	G	0.94	ENSECAG00000025146	*CDCP1*	rs69080137
16	41,453,020	C	A	0.96	ENSECAG00000025146	*CDCP1*	rs69080139
18	59,599,035	C	G	0.93	ENSECAG00000012220	*SSFA2*	-
18	59,612,355	C	A	0.83	ENSECAG00000012220	*SSFA2*	-
19	38,746,008	C	G	0.96	ENSECAG00000024669	*GPR156*	-
20	23,740,992	G	T	0.82	ENSECAG00000017259	*SLC17A1*	-
20	42,327,584	C	T	0.82	ENSECAG00000001119	*ZNF318*	-
20	42,508,228	C	A	0.89	ENSECAG00000000507	*XPO5*	-
23	20,639,345	C	T	0.94	ENSECAG00000015934	*MAMDC2*	rs69256779
26	37,639,216	G	A	0.82	ENSECAG00000013946	*UMODL1*	-
26	40,592,293	G	A	0.92	ENSECAG00000016361	*COL18A1*	-
28	776,789	G	T	0.91	ENSECAG00000001320		rs69415182
28	36,540,089	G	C	0.92	ENSECAG00000018973	*ENTHD1*	-
29	28,372,040	C	A	0.97	ENSECAG00000006393	*CALML3*	-
30	24,873,853	C	T	0.81	ENSECAG00000020645	*ASPM*	-
31	2,460,558	C	T	1.00	ENSECAG00000021017	*SLC22A2*	-
31	2,460,756	G	A	0.99	ENSECAG00000021017	*SLC22A2*	-
31	9,661,000	G	A	0.99	ENSECAG00000010933	*C6orf70*	-

Supplementary Table S41. List of non-synonymous sites that are homozygous ancestral in the ancient horse CGG10022, derived in five modern horses and with posterior probabilities of at least 0.8

***Chr:** chromosome; **Position:** zero-based position of each site; **Anc:** ancestral allele; **Dir:** derived allele; **PP:** posterior probability; **Gene ID:** Ensembl gene ID of the overlapping genes. **Gene Name:** gene name, if known; **Existing Variation:** reference cluster ID of the existing SNPs in dbSNP.*

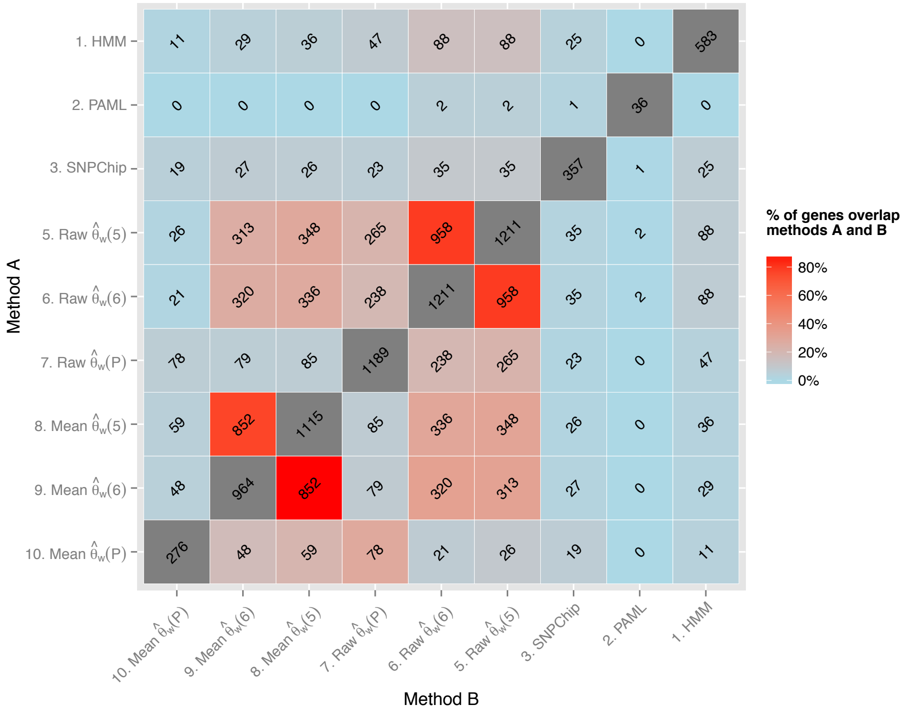
S4.5 Comparison between different selection scans

S4.5.1 Pairwise comparisons

To assess the overall consistency between the genomic selection scans, we performed a pairwise comparison of the Ensembl gene IDs detected in all methods described above. All pairwise comparisons performed are shown in Supplementary Figure S58, using the following abbreviations:

1. **HMM**, genes detected using the Hidden Markov Model (section S4.4).
2. **PAML**, genes detected using PAML (section S4.1).
3. **SNPChip**, top 1% of genes detected using the SNPChip (section S4.3) in strictly more than 16 breeds out of 32 for either CGG10022 or CGG10023 (in either window).
4. **Raw** $\boldsymbol{\widehat{\theta}_w(5)}$, top 5% of genes detected using the raw TWLR (S4.2.6), comparing the pre-domesticated horses to 5 domestic horses, excluding Icelandic (P5782).
5. **Raw** $\boldsymbol{\widehat{\theta}_w(6)}$, top 5% of genes detected using the raw TWLR (S4.2.6), comparing the pre-domesticated horses to all 6 domestic horses.
6. **Raw** $\boldsymbol{\widehat{\theta}_w(P)}$, top 5% of genes detected using the raw TWLR (S4.2.6), comparing Przewalski's horse to 5 domestic horses (excluding Icelandic (P5782)).
7. **Mean** $\boldsymbol{\widehat{\theta}_w(5)}$, top 5% of genes detected using the region-averaged $\widehat{\theta}_w$ (S4.2.2), comparing the pre-domesticated horses with 5 domestic horses, excluding Icelandic (P5782).
8. **Mean** $\boldsymbol{\widehat{\theta}_w(6)}$, top 5% of genes detected using the region-averaged $\widehat{\theta}_w$ (S4.2.4), comparing the pre-domesticated horses with all 6 domestic horses.
9. **Mean** $\boldsymbol{\widehat{\theta}_w(P)}$, top 5% of genes detected using the region-averaged $\widehat{\theta}_w$ (S4.2.5), comparing the pre-domesticated horses with 5 domestic horses, excluding Icelandic (P5782).

In the case of $\widehat{\theta}_w$ sets, we choose to disregard overlap (or lack of overlap) with the Tajima's D regions detected in conjunction with these. This was motivated by the observation that no deviations were observed on the QQ-plots, suggesting that there was no signal on which to base this selection (Supplementary Figure S40 and Supplementary Figure S43).



Supplementary Figure S58. Relative overlap of Ensembl gene IDs detected as potential candidates for positive selection across all methods

The gradient color corresponds to the percentage of overlap between two methods. The numbers inside each tile refers to the absolute number of gene IDs detected. See text for the description of the methods. Percentages were calculated # genes ($M_A \cap M_B$) / # genes (M_A), for the method M_A specified by row and for the method M_B specified by column. The number of genes overlapping is calculated in a similar fashion.

S4.6 Weighted selection of candidate genes

We next determined which genes were detected independently by at least two of the selection scan methods implemented by relying on a weighting scheme. For tests with an associated score (q-value for PAML, length of the external region for HMM, the number of individuals for individuals for the SNP chip) the set of unique scores were collected and ranked from 1 to N according to their significance, with the rank 0 assigned to genes that were not detected by a given test. Subsequently, each a score was calculated for each gene as $S_{t,i} = R_{t,i}/N_t$, where $R_{t,i}$ is the rank of gene i for test t, and N_t is the number of ranks for test t. For the $\hat{\theta}_w$ tests, the rank was calculated as the number of tests for which the gene was detected (0 to 6), and the score calculated as describe above with $N_t = 6$. The sum of scores was calculated for each gene, and 125 genes with an aggregated score greater than 1 were selected (Supplementary Table S42). The full table of genes is provided as Supplementary Table S43.

Of the 125 genes, 113 had known human orthologs and these were analyzed through the use of IPA (Ingenuity® Systems, www.ingenuity.com), and statistically significant functional clusters were determined following Benjamini-Hochberg correction for multiple tests (Supplementary Table S44). In addition to annotations relating to cancer, this includes two annotations relating to brain development, including “development of telencephalon” and “recognition of neurons”, as well as one annotation relating to lipid metabolism ("accumulation of lysobisphosphatidic acid"), the latter of which is important for the performance characteristics of the domestic horse.

Finally, functional clustering of all genes detected by more than one methodology (HMM, SNPChip, PAML, or $\hat{\theta}_w$, Supplementary Figure S59) was carried out, yielding 697 candidate genes, of which 571 had known human orthologs. Statistically significant results, following Benjamini-Hochberg correction for multiple tests, are reported in Supplementary Table S45. In addition to annotations relating to and to the nervous system, several annotations for other diseases were found (hypolipoproteinemia and hypobetalipoproteinemia).

A number of genes have previously been identified in regions determined to have undergone positive selection in the Thoroughbred, including *ACTA1*, *ACTN2*, *FOXO1*, *GRB2*, *IRS1*, *PIK3C3*, *PIK3R1*, *PTPN1*, *SOCS3*, *SOCS7*, and *STXBP4*. In addition, several genes have been identified as important to racing performance, including *COX4I2*, *MSTN*, and *PDK4* (reviewed in Hill et al. 2013 (128)). Of these, *ACTA1* and *PIK3C3* are detected using the weighting criteria described above, while *STXBP4* was detected by at least two types of selective scan, and *FOXO1*, *SOCS7*, *MSTN*, *PDK4*, and *PIK3R1* was detected by one type of scan.

While our samples includes just one Thoroughbred horse, it is notable that we were able to detect most of the genes listed above, suggesting the validity of our approach. However, the fact that only two of the eight genes detected exceeded the threshold value we selected also indicates that the set of 125 genes constitute a conservative set of candidates for positive selection in the domestic horse.

Weight	Ensembl Gene ID	Gene Name	Description
1.855	ENSECAG00000016246	*ASAP1*	ArfGAP with SH3 domain, ankyrin repeat and PH domain 1
1.850	ENSECAG00000019629	-	Beta-galactosidase
1.846	ENSECAG00000014267	*NINJ1*	ninjurin 1
1.790	ENSECAG00000002579	*MATN2*	matrilin 2
1.780	ENSECAG00000011435	*MSI2*	musashi RNA-binding protein 2
1.691	ENSECAG00000012105	*IGSF9B*	immunoglobulin superfamily, member 9B
1.656	ENSECAG00000023838	-	-
1.626	ENSECAG00000020760	*NUP133*	nucleoporin 133kDa
1.594	ENSECAG00000007574	*NTM*	Neurotrimin
1.588	ENSECAG00000007880	*ABCA5*	ATP-binding cassette, sub-family A (ABC1), member 5
1.540	ENSECAG00000009608	*B3GALTL*	beta 1,3-galactosyltransferase-like
1.530	ENSECAG00000015121	*ACSF3*	acyl-CoA synthetase family member 3
1.491	ENSECAG00000008971	*LEPREL1*	leprecan-like 1
1.480	ENSECAG00000010682	*GNPTAB*	N-acetylglucosamine-1-phosphate transferase, alpha and beta subunits
1.460	ENSECAG00000000207	*ACTA1*	actin, alpha 1, skeletal muscle
1.446	ENSECAG00000000050	*WASF3*	WAS protein family, member 3
1.442	ENSECAG00000013957	*PDE5A*	phosphodiesterase 5A, cGMP-specific
1.422	ENSECAG00000006788	*PCSK5*	proprotein convertase subtilisin/kexin type 5
1.404	ENSECAG00000000711	*COIL*	coilin
1.381	ENSECAG00000023160	*TCTN1*	tectonic family member 1
1.371	ENSECAG00000012556	*ZC3H3*	zinc finger CCCH-type containing 3
1.363	ENSECAG00000011163	*FCHSD2*	FCH and double SH3 domains 2
1.347	ENSECAG00000021446	*GRID1*	glutamate receptor, ionotropic, delta 1
1.335	ENSECAG00000020117	*SLC22A15*	solute carrier family 22, member 15
1.333	ENSECAG00000006392	*SEC63*	SEC63 homolog (S. cerevisiae)
1.330	ENSECAG00000021760	*DCC*	deleted in colorectal carcinoma
1.326	ENSECAG00000018233	*NIPBL*	Nipped-B homolog (Drosophila)
1.326	ENSECAG00000021838	-	-
1.313	ENSECAG00000008625	*MYBPC1*	myosin binding protein C, slow type
1.306	ENSECAG00000017883	*ABCB10*	ATP-binding cassette, sub-family B (MDR/TAP), member 10

1.303	ENSECAG00000010657	*PIK3C3*	phosphatidylinositol 3-kinase, catalytic subunit type 3
1.292	ENSECAG00000016277	*SGCD*	sarcoglycan, delta (35kDa dystrophin-associated glycoprotein)
1.278	ENSECAG00000012338	*URB2*	URB2 ribosome biogenesis 2 homolog (S. cerevisiae)
1.271	ENSECAG00000003020	-	-
1.271	ENSECAG00000010509	-	Uncharacterized protein
1.270	ENSECAG00000009233	*OPCML*	opioid binding protein/cell adhesion molecule-like
1.268	ENSECAG00000016963	*COMMD1*	copper metabolism (Murr1) domain containing 1
1.250	ENSECAG00000012724	*MAP3K4*	mitogen-activated protein kinase kinase kinase 4
1.238	ENSECAG00000014869	*PPM1D*	protein phosphatase, Mg2+/Mn2+ dependent, 1D
1.238	ENSECAG00000014849	-	-
1.232	ENSECAG00000018241	*C15orf60*	chromosome 15 open reading frame 60
1.228	ENSECAG00000016605	*NR3C2*	nuclear receptor subfamily 3, group C, member 2
1.214	ENSECAG00000024236	*POP1*	processing of precursor 1, ribonuclease P/MRP subunit (S. cerevisiae)
1.214	ENSECAG00000009432	*PDRG1*	p53 and DNA-damage regulated 1
1.203	ENSECAG00000017930	*PHF2*	PHD finger protein 2
1.199	ENSECAG00000017207	-	-
1.199	ENSECAG00000022613	*PPP2CA*	protein phosphatase 2, catalytic subunit, alpha isozyme
1.199	ENSECAG00000012365	*VDAC1*	voltage-dependent anion channel 1
1.199	ENSECAG00000019667	*KLHDC4*	kelch domain containing 4
1.196	ENSECAG00000024479	*LCLAT1*	lysocardiolipin acyltransferase 1
1.193	ENSECAG00000009841	*AKAP1*	A kinase (PRKA) anchor protein 1
1.190	ENSECAG00000002769	-	-
1.184	ENSECAG00000023361	*RHPN1*	rhophilin, Rho GTPase binding protein 1
1.183	ENSECAG00000021124	*PRMT3*	protein arginine methyltransferase 3
1.181	ENSECAG00000007481	*ASTN1*	astrotactin 1
1.178	ENSECAG00000023493	*VPS26B*	vacuolar protein sorting 26 homolog B (S. pombe)
1.168	ENSECAG00000024214	*FANCA*	Fanconi anemia, complementation group A
1.167	ENSECAG00000023888	*ALDH1L2*	aldehyde dehydrogenase 1 family, member L2
1.167	ENSECAG00000008427	*SLC43A1*	solute carrier family 43, member 1
1.166	ENSECAG00000015892	*CACNA1D*	calcium channel, voltage-dependent, L type, alpha 1D subunit
1.165	ENSECAG00000007080	*ALK*	anaplastic lymphoma receptor tyrosine kinase
1.161	ENSECAG00000000323	*ABCA10*	ATP-binding cassette, sub-family A (ABC1), member 10
1.159	ENSECAG00000025396	-	-
1.159	ENSECAG00000005781	*SCPEP1*	serine carboxypeptidase 1
1.156	ENSECAG00000019990	*WNK2*	WNK lysine deficient protein kinase 2
1.156	ENSECAG00000008495	*FAF1*	Fas (TNFRSF6) associated factor 1
1.150	ENSECAG00000018811	*ARL6IP1*	ADP-ribosylation factor-like 6 interacting protein 1
1.150	ENSECAG00000022891	*PDE4DIP*	phosphodiesterase 4D interacting protein
1.149	ENSECAG00000015369	*NID2*	nidogen 2 (osteonidogen)
1.141	ENSECAG00000000705	*MARCH10*	membrane-associated ring finger (C3HC4) 10, E3 ubiquitin protein ligase
1.138	ENSECAG00000023360	*KCNK10*	potassium channel, subfamily K, member 10
1.138	ENSECAG00000020012	*KIAA0556*	KIAA0556
1.133	ENSECAG00000017280	*MAPK10*	mitogen-activated protein kinase 10
1.132	ENSECAG00000015975	*CNTN6*	contactin 6
1.130	ENSECAG00000002321	*SEC24A*	SEC24 family, member A (S. cerevisiae)
1.130	ENSECAG00000027254	-	-
1.130	ENSECAG00000018013	*CDKL3*	cyclin-dependent kinase-like 3
1.130	ENSECAG00000016419	*UBE2B*	ubiquitin-conjugating enzyme E2B
1.121	ENSECAG00000016313	*PTPN4*	protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte)
1.118	ENSECAG00000024475	*CRTC3*	CREB regulated transcription coactivator 3
1.114	ENSECAG00000013739	*DLGAP1*	discs, large (Drosophila) homolog-associated protein 1
1.104	ENSECAG00000003504	-	-
1.104	ENSECAG00000011286	-	-
1.101	ENSECAG00000016684	*NUMB*	numb homolog (Drosophila)
1.098	ENSECAG00000007889	-	Uncharacterized protein
1.096	ENSECAG00000023832	*GAK*	cyclin G associated kinase
1.096	ENSECAG00000018382	*COL22A1*	collagen, type XXII, alpha 1
1.094	ENSECAG00000019890	*VRK1*	vaccinia related kinase 1
1.086	ENSECAG00000011570	-	Uncharacterized protein
1.081	ENSECAG00000018366	*JPH3*	junctophilin 3
1.080	ENSECAG00000007262	*EEA1*	early endosome antigen 1
1.075	ENSECAG00000015151	-	-

1.075	ENSECAG00000015522	-	-
1.075	ENSECAG00000026144	-	-
1.067	ENSECAG00000019952	*TRIO*	trio Rho guanine nucleotide exchange factor
1.063	ENSECAG00000017222	*LMF1*	lipase maturation factor 1
1.051	ENSECAG00000012543	*C-SKI*	ski oncogene
1.047	ENSECAG00000004789	-	-
1.047	ENSECAG00000001970	-	-
1.047	ENSECAG00000001912	-	-
1.047	ENSECAG00000008665	*FOXJ3*	forkhead box J3
1.046	ENSECAG00000001926	*FBXO31*	F-box protein 31
1.043	ENSECAG00000016835	*STXBP6*	syntaxin binding protein 6 (amisyn)
1.042	ENSECAG00000023789	*AMBRA1*	autophagy/beclin-1 regulator 1
1.038	ENSECAG00000025159	*DNAH9*	dynein, axonemal, heavy chain 9
1.038	ENSECAG00000009860	*STAB1*	stabilin 1
1.038	ENSECAG00000015835	*NT5DC2*	5'-nucleotidase domain containing 2
1.034	ENSECAG00000011659	*KIAA1549*	KIAA1549
1.034	ENSECAG00000022937	-	-
1.032	ENSECAG00000003931	*SKP1*	S-phase kinase-associated protein 1
1.032	ENSECAG00000027356	-	-
1.029	ENSECAG00000022378	*IGSF3*	immunoglobulin superfamily, member 3
1.028	ENSECAG00000024820	*PHF20*	PHD finger protein 20
1.027	ENSECAG00000023415	*PSMB7*	proteasome (prosome, macropain) subunit, beta type, 7
1.026	ENSECAG00000017284	*ANKDD1A*	ankyrin repeat and death domain containing 1A
1.022	ENSECAG00000006519	*SMG6*	smg-6 homolog, nonsense mediated mRNA decay factor (C. elegans)
1.019	ENSECAG00000010370	*CDK5RAP1*	CDK5 regulatory subunit associated protein 1
1.014	ENSECAG00000002746	*BRAF*	v-raf murine sarcoma viral oncogene homolog B1
1.013	ENSECAG00000014312	*PRKCZ*	protein kinase C, zeta
1.011	ENSECAG00000019584	-	-
1.011	ENSECAG00000018370	*NCAPD3*	non-SMC condensin II complex, subunit D3
1.011	ENSECAG00000011007	*ACAD8*	acyl-CoA dehydrogenase family, member 8
1.011	ENSECAG00000001596	*THYN1*	thymocyte nuclear protein 1
1.011	ENSECAG00000026872	-	-
1.011	ENSECAG00000015550	*JAM3*	junctional adhesion molecule 3

Supplementary Table S42. Weighted selection of candidate genes (weight > 1)

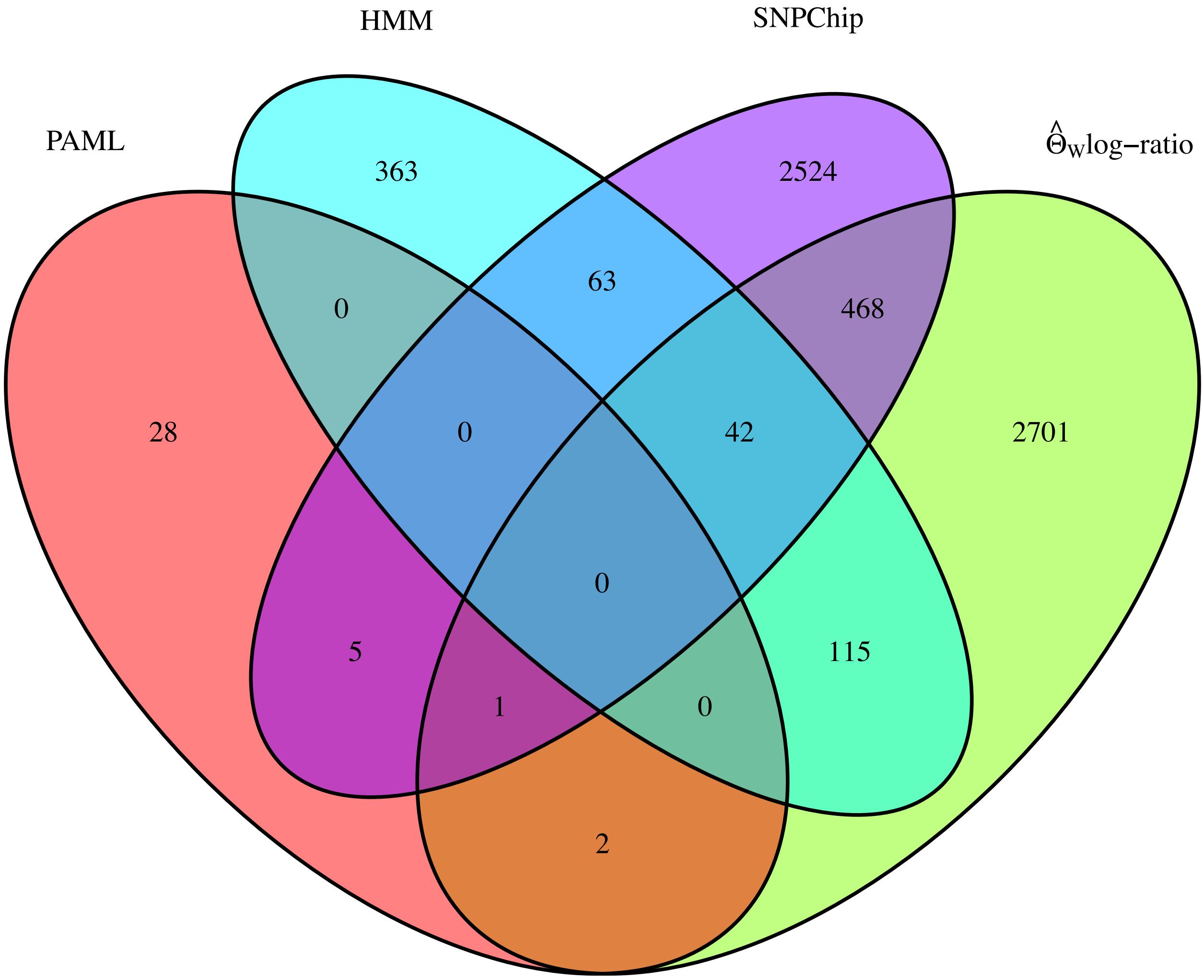
Please see SI Dataset 1.

Supplementary Table S43. Weighted and un-weighted selection of candidate genes

Annotation	***q*-value**	**# Molecules**	**Categories**
endometrioid carcinoma	3.28E-03	38	Cancer
aneuploidy of breast cell lines	1.82E-02	2	Cell Cycle; Reproductive System Development and Function
Cancer	1.82E-02	71	Cancer
colon cancer	1.82E-02	38	Cancer; Gastrointestinal Disease
colon carcinoma	1.82E-02	36	Cancer; Gastrointestinal Disease
colorectal cancer	1.82E-02	41	Cancer; Gastrointestinal Disease
development of telencephalon	1.82E-02	6	Embryonic Development; Nervous System Development and Function; Organ Development; Organismal Development; Tissue Development
gastrointestinal carcinoma	1.82E-02	37	Cancer; Gastrointestinal Disease

gastrointestinal tract cancer	1.82E-02	44	Cancer; Gastrointestinal Disease
digestive organ tumor	2.88E-02	45	Cancer; Gastrointestinal Disease
accumulation of lysobisphosphatidic acid	3.94E-02	2	Lipid Metabolism; Molecular Transport; Small Molecule Biochemistry
recognition of neurons	3.94E-02	2	Nervous System Development and Function; Cell-To-Cell Signaling and Interaction
breast or colorectal cancer	4.40E-02	46	Cancer

Supplementary Table S44. Enriched functions in genes with a weight > 1



Supplementary Figure S59. Venn diagram of genes detected by method category

Annotation	*q*-value	# Molecules	Category
adenocarcinoma	1.41E-04	186	Cancer
breast or colorectal cancer	1.41E-04	182	Cancer
Cancer	1.41E-04	288	Cancer
carcinoma	1.41E-04	235	Cancer
colon adenocarcinoma	1.41E-04	128	Cancer; Gastrointestinal Disease
colon cancer	1.41E-04	141	Cancer; Gastrointestinal Disease
colon carcinoma	1.41E-04	134	Cancer; Gastrointestinal Disease
endometrioid carcinoma	1.41E-04	124	Cancer
gastrointestinal adenocarcinoma	1.41E-04	131	Cancer; Gastrointestinal Disease
gastrointestinal carcinoma	1.41E-04	138	Cancer; Gastrointestinal Disease
solid tumor	1.41E-04	237	Cancer
epithelial neoplasia	1.76E-04	242	Cancer
gastrointestinal tract cancer	1.80E-04	162	Cancer; Gastrointestinal Disease
colorectal cancer	2.58E-04	152	Cancer; Gastrointestinal Disease
digestive organ tumor	1.77E-03	168	Cancer; Gastrointestinal Disease
malignant neoplasm of abdomen	3.25E-03	185	Cancer
morphology of artery	1.92E-02	14	Cardiovascular System Development and Function; Organismal Development; Tissue Morphology
recurrent medullary thyroid carcinoma	3.69E-02	5	Cancer; Endocrine System Disorders
scattered medullary thyroid carcinoma	3.69E-02	5	Cancer; Endocrine System Disorders
hypobetalipoproteinemia	3.78E-02	3	Developmental Disorder; Hematological Disease; Hereditary Disorder; Metabolic Disease
advanced rectal cancer	4.23E-02	5	Cancer; Gastrointestinal Disease
astrocytoma	4.26E-02	26	Cancer; Neurological Disease
development of central nervous system	4.84E-02	34	Nervous System Development and Function
hypolipoproteinemia	4.84E-02	4	Metabolic Disease

Supplementary Table S45. Enriched functions in genes detected by more than one method

References

1. Orlando L, *et al.* (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* 499(7456):74-78.
2. Andersson LS, *et al.* (2012) Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. *Nature* 488(7413):642-646.
3. Wade CM, *et al.* (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326(5954):865-867.
4. Stuiver M & Reimer PJ (1993) Extended C-14 Data-Base And Revised CALIB 3.0 C-14 Age Calibration Program. *Radiocarbon* 35(1):215-230.
5. Outram AK, *et al.* (2009) The earliest horse harnessing and milking. *Science* 323(5919):1332-1335.
6. Seguin-Orlando A, *et al.* (2013) Ligation bias in illumina next-generation DNA libraries: implications for sequencing ancient genomes. *PloS one* 8(10):e78575.
7. Der Sarkissian C, *et al.* (2014) Shotgun Microbial Profiling of Fossil Remains. *Molecular Ecology*:n/a-n/a.
8. Meyer M & Kircher M (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols* 2010(6):pdb prot5448.
9. Pedersen JS, *et al.* (2014) Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome research* 24(3):454-466.
10. Schubert M, *et al.* (2014) Characterization of ancient and modern genomes by sequencing, SNP detection, phylogenomic and metagenomic analysis. *Nature protocols* 9(5):1056-1082.
11. Lindgreen S (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC research notes* 5:337.
12. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
13. Schubert M, *et al.* (2012) Improving ancient DNA read mapping against modern reference genomes. *BMC genomics* 13:178.
14. McKenna A, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20(9):1297-1303.
15. Li H, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
16. Briggs AW, *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* 104(37):14616-14621.
17. Jónsson H, Ginolhac A, Schubert M, Johnson PL, & Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29(13):1682-1684.
18. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical population biology* 7(2):256-276.
19. McLaren W, *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26(16):2069-2070.
20. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841-842.
21. Der Sarkissian C, *et al.* (2014) Shotgun Microbial Profiling of Fossil Remains. *Mol Ecol* 23(7):1780-1798.
22. Segata N, *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* 9(8):811-814.
23. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* 9(4):357-359.

24. Green RE*, et al.* (2010) A draft sequence of the Neandertal genome. *Science* 328(5979):710-722.
25. Quail MA*, et al.* (2008) A large genome center's improvements to the Illumina sequencing system. *Nature methods* 5(12):1005-1010.
26. R-Core-Team (2013) *R: A Language and Environment for Statistical Computing*.
27. Suzuki R & Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22(12):1540-1542.
28. Fierer N*, et al.* (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences of the United States of America* 109(52):21390-21395.
29. Consortium HMP (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207-214.
30. Sawyer S, Krause J, Guschanski K, Savolainen V, & Paabo S (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PloS one* 7(3):e34131.
31. Keane TM, Creevey CJ, Pentony MM, Naughton TJ, & McLnerney JO (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 evolutionary biology* 6:29.
32. Huelsenbeck JP & Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754-755.
33. Tamura K*, et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution* 28(10):2731-2739.
34. Wallner B*, et al.* (2013) Identification of genetic variation on the horse y chromosome and the tracing of male founder lineages in modern breeds. *PloS one* 8(4):e60015.
35. Lippold S*, et al.* (2011) Discovery of lost diversity of paternal horse lineages using ancient DNA. *Nature communications* 2:450.
36. Librado P & Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25(11):1451-1452.
37. Bandelt HJ, Forster P, & Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution* 16(1):37-48.
38. McCue ME*, 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(1):e1002451.
39. Petersen JL*, et al.* (2013) Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genet* 9(1):e1003211.
40. Purcell S*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81(3):559-575.
41. Patterson N, Price AL, & Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.
42. Dixon P (2003) VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* 14(6):927-930.
43. Doan R*, et al.* (2012) Whole-genome sequencing and genetic variant analysis of a Quarter Horse mare. *BMC genomics* 13:78.
44. Nicholas FW (2003) Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. *Nucleic acids research* 31(1):275-277.
45. Signer-Hasler H*, et al.* (2012) A genome-wide association study reveals loci influencing height and other conformation traits in horses. *PloS one* 7(5):e37282.
46. Pruvost M*, et al.* (2011) Genotypes of predomestic horses match phenotypes painted in Paleolithic works of cave art. *Proceedings of the National Academy of Sciences of the United States of America* 108(46):18626-18630.

47. Bellone RR*, 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(10):e78280.

48. Hill EW, Gu J, McGivney BA, & MacHugh DE (2010) Targets of selection in the Thoroughbred genome contain exercise-relevant gene SNPs associated with elite racecourse performance. *Animal genetics* 41 Suppl 2:56-63.

49. Bellone RR*, et al.* (2010) Fine-mapping and mutation analysis of TRPM1: a candidate gene for leopard complex (LP) spotting and congenital stationary night blindness in horses. *Briefings in functional genomics* 9(3):193-207.

50. Tryon RC, White SD, & Bannasch DL (2007) Homozygosity mapping approach identifies a missense mutation in equine cyclophilin B (PPIB) associated with HERDA in the American Quarter Horse. *Genomics* 90(1):93-102.

51. Brooks SA*, et al.* (2010) Whole-genome SNP association in the horse: identification of a deletion in myosin Va responsible for Lavender Foal Syndrome. *PLoS Genet* 6(4):e1000909.

52. Brault LS, Cooper CA, Famula TR, Murray JD, & Penedo MC (2011) Mapping of equine cerebellar abiotrophy to ECA2 and identification of a potential causative mutation affecting expression of MUTYH. *Genomics* 97(2):121-129.

53. Marklund L, Moller MJ, Sandberg K, & Andersson L (1996) A missense mutation in the gene for melanocyte-stimulating hormone receptor (MC1R) is associated with the chestnut coat color in horses. *Mammalian genome : official journal of the International Mammalian Genome Society* 7(12):895-899.

54. Wagner HJ & Reissmann M (2000) New polymorphism detected in the horse MC1R gene. *Animal genetics* 31(4):289-290.

55. Brooks SA & Bailey E (2005) Exon skipping in the KIT gene causes a Sabino spotting pattern in horses. *Mammalian genome : official journal of the International Mammalian Genome Society* 16(11):893-902.

56. Brooks SA, Lear TL, Adelson DL, & Bailey E (2007) A chromosome inversion near the KIT gene and the Tobiano spotting pattern in horses. *Cytogenetic and genome research* 119(3-4):225-230.

57. Makvandi-Nejad S*, et al.* (2012) Four loci explain 83% of size variation in the horse. *PloS one* 7(7):e39929.

58. Wijnberg ID*, et al.* (2012) A missense mutation in the skeletal muscle chloride channel 1 (CLCN1) as candidate causal mutation for congenital myotonia in a New Forest pony. *Neuromuscular disorders : NMD* 22(4):361-367.

59. Spirito F*, et al.* (2002) Animal models for skin blistering conditions: absence of laminin 5 causes hereditary junctional mechanobullous disease in the Belgian horse. *The Journal of investigative dermatology* 119(3):684-691.

60. Hauswirth R*, et al.* (2012) Mutations in MITF and PAX3 cause "splashed white" and other white spotting phenotypes in horses. *PLoS Genet* 8(4):e1002653.

61. Hauswirth R*, et al.* (2013) Novel variants in the KIT and PAX3 genes in horses with white-spotted coat colour phenotypes. *Animal genetics* 44(6):763-765.

62. Brunberg E*, et al.* (2006) A missense mutation in PMEL17 is associated with the Silver coat color in the horse. *BMC genetics* 7:46.

63. Graves KT, Henney PJ, & Ennis RB (2009) Partial deletion of the LAMA3 gene is responsible for hereditary junctional epidermolysis bullosa in the American Saddlebred Horse. *Animal genetics* 40(1):35-41.

64. Shin EK, Perryman LE, & Meek K (1997) A kinase-negative mutation of DNA-PK(CS) in equine SCID results in defective coding and signal joint formation. *J Immunol* 158(8):3565-3569.

65. Aleman M*, et al.* (2004) Association of a mutation in the ryanodine receptor 1 gene with equine malignant hyperthermia. *Muscle & nerve* 30(3):356-365.

66. Gu J, *et al.* (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 veterinary journal. Supplement* (38):569-575.
67. Cannon SC, Hayward LJ, Beech J, & Brown RH, Jr. (1995) Sodium channel inactivation is impaired in equine hyperkalemic periodic paralysis. *Journal of neurophysiology* 73(5):1892-1899.
68. Christopherson PW, van Santen VL, Livesey L, & Boudreaux MK (2007) A 10-base-pair deletion in the gene encoding platelet glycoprotein IIb associated with Glanzmann thrombasthenia in a horse. *Journal of veterinary internal medicine / American College of Veterinary Internal Medicine* 21(1):196-198.
69. Orr N, *et al.* (2010) Genome-wide SNP association-based localization of a dwarfism gene in Friesian dwarf horses. *Animal genetics* 41 Suppl 2:2-7.
70. Cook D, Brooks S, Bellone R, & Bailey E (2008) Missense mutation in exon 2 of SLC36A1 responsible for champagne dilution in horses. *PLoS Genet* 4(9):e1000195.
71. Hansen M, Knorr C, Hall AJ, Broad TE, & Brenig B (2007) Sequence analysis of the equine SLC26A2 gene locus on chromosome 14q15-->q21. *Cytogenetic and genome research* 118(1):55-62.
72. Yang GC, *et al.* (1998) A dinucleotide mutation in the endothelin-B receptor gene is associated with lethal white foal syndrome (LWFS); a horse variant of Hirschsprung disease. *Human molecular genetics* 7(6):1047-1052.
73. Tozaki T, *et al.* (2010) A genome-wide association study for racing performances in Thoroughbreds clarifies a candidate region near the MSTN gene. *Animal genetics* 41 Suppl 2:28-35.
74. Hill EW, *et al.* (2012) MSTN genotype (g.66493737C/T) association with speed indices in Thoroughbred racehorses. *J Appl Physiol (1985)* 112(1):86-90.
75. Mariat D, Taourit S, & Guerin G (2003) A mutation in the MATP gene causes the cream coat colour in the horse. *Genetics, selection, evolution : GSE* 35(1):119-133.
76. Rieder S, Taourit S, Mariat D, Langlois B, & Guerin G (2001) Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus). *Mammalian genome : official journal of the International Mammalian Genome Society* 12(6):450-455.
77. Fox-Clipsham LY, *et al.* (2011) Identification of a mutation associated with fatal Foal Immunodeficiency Syndrome in the Fell and Dales pony. *PLoS Genet* 7(7):e1002133.
78. Revay T, Villagomez DA, Brewer D, Chenier T, & King WA (2012) GTG mutation in the start codon of the androgen receptor gene in a family of horses with 64,XY disorder of sex development. *Sexual development : genetics, molecular biology, evolution, endocrinology, embryology, and pathology of sex determination and differentiation* 6(1-3):108-116.
79. Towers RE, *et al.* (2013) A Nonsense Mutation in the IKBKG Gene in Mares with Incontinentia Pigmenti. *PloS one* 8(12):e81625.
80. Flicek P, *et al.* (2013) Ensembl 2013. *Nucleic acids research* 41(Database issue):D48-55.
81. Sharon N (1986) IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature of glycoproteins, glycopeptides and peptidoglycans. Recommendations 1985. *European journal of biochemistry / FEBS* 159(1):1-6.
82. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688-2690.
83. Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19(2):301-302.
84. Durand EY, Patterson N, Reich D, & Slatkin M (2011) Testing for ancient admixture between closely related populations. *Molecular biology and evolution* 28(8):2239-2252.
85. Busing FTA, Meijer E, & Leeden R (1999) Delete-m Jackknife for Unequal m. *Statistics and Computing* 9(1):3-8.
86. Cahill JA, *et al.* (2013) Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. *PLoS Genet* 9(3):e1003345.

87. Rasmussen M*, et al.* (2011) An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* 334(6052):94-98.

88. Danecek P*, et al.* (2011) The variant call format and VCFtools. *Bioinformatics* 27(15):2156-2158.

89. Pickrell JK & Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8(11):e1002967.

90. Csilléry K, François O, & Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution* 3(3):475-479.

91. Achilli A*, et al.* (2012) Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proc. Natl. Acad. Sci.* 109(7):2449-2454.

92. Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, & Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genet* 9(10):e1003905.

93. Gutenkunst RN, Hernandez RD, Williamson SH, & Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5(10):e1000695.

94. Li H & Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature* 475(7357):493-496.

95. Karolchik D*, et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic acids research* 42(Database issue):D764-770.

96. Cooper GM*, et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* 15(7):901-913.

97. Davydov EV*, et al.* (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology* 6(12):e1001025.

98. Wang K, Li M, & Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* 38(16):e164.

99. Cruz F, Vila C, & Webster MT (2008) The legacy of domestication: accumulation of deleterious mutations in the dog genome. *Molecular biology and evolution* 25(11):2331-2336.

100. Moray C, Lanfear R, & Bromham L (2014) Domestication and the mitochondrial genome: comparing patterns and rates of molecular evolution in domesticated mammals and birds and their wild relatives. *Genome biology and evolution* 6(1):161-169.

101. Nabholz B*, et al.* (2014) Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (Oryza glaberrima). *Mol Ecol* 23(9):2210-2227.

102. Lu J*, et al.* (2006) The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in genetics : TIG* 22(3):126-131.

103. Koenig D*, et al.* (2013) Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proceedings of the National Academy of Sciences of the United States of America* 110(28):E2655-2662.

104. Rokas A (2009) The effect of domestication on the fungal proteome. *Trends in genetics : TIG* 25(2):60-63.

105. Lau AN*, et al.* (2009) Horse domestication and conservation genetics of Przewalski's horse inferred from sex chromosomal and autosomal sequences. *Molecular biology and evolution* 26(1):199-208.

106. Prüfer K*, et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505(7481):43-49.

107. Cieslak M*, et al.* (2010) Origin and history of mitochondrial DNA lineages in domestic horses. *PloS one* 5(12):e15311.

108. Wakefield S, Knowles J, Zimmermann W, & Van Dierendonck M (2002) Status and action plan for the Przewalski's horse (Equus ferus przewalskii). *Equids: Zebras, Asses and Horses - Status Survey and Conservation Action Plan* (IUCN, Gland, Switzerland and Cambridge, UK), pp p82-92.

109. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* 24(8):1586-1591.

110. Sandve GK, Ferkingstad E, & Nygard S (2011) Sequential Monte Carlo multiple testing. *Bioinformatics* 27(23):3235-3241.
111. Amaral AJ*, et al.* (2011) Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PloS one* 6(4):e14782.
112. Luciano M*, et al.* (2012) Longevity candidate genes and their association with personality traits in the elderly. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 159B(2):192-200.
113. Lopez LM*, et al.* (2012) Evolutionary conserved longevity genes and human cognitive abilities in elderly cohorts. *European journal of human genetics : EJHG* 20(3):341-347.
114. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585-595.
115. Korneliussen TS, Moltke I, Albrechtsen A, & Nielsen R (2013) Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC bioinformatics* 14:289.
116. Nielsen R*, et al.* (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS biology* 3(6):e170.
117. Aasland M, Klungland H, & Lien S (2000) Two polymorphisms in the bovine mast cell growth factor gene (MGF). *Animal genetics* 31(5):345.
118. Lindahl T (1993) Instability and decay of the primary structure of DNA. *Nature* 362(6422):709-715.
119. Prüfer K*, et al.* (2012) The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486(7404):527-531.
120. Dumont BL & Payseur BA (2008) Evolution of the genomic rate of recombination in mammals. *Evolution* 62(2):276-294.
121. Durbin R, Eddy S, Krogh A, & Graeme Mitchison S (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids* (Cambridge University Press, Cambridge, UK).
122. Curwen V*, et al.* (2004) The Ensembl automatic gene annotation system. *Genome Res.* 14(5):942-950.
123. Huang da W, Sherman BT, & Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* 4(1):44-57.
124. Huang da W, Sherman BT, & Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* 37(1):1-13.
125. Behrends C, Sowa ME, Gygi SP, & Harper JW (2010) Network organization of the human autophagy system. *Nature* 466(7302):68-76.
126. Zuko A*, et al.* (2013) Contactins in the neurobiology of autism. *Eur. J. Pharmacol.* 719(1,Äì3):63-74.
127. Takeda Y*, et al.* (2003) Impaired motor coordination in mice lacking neural recognition molecule NB-3 of the contactin/F3 subgroup. *J. Neurobiol.* 56(3):252-265.
128. Hill EW, Katz LM, & MacHugh DE (2013) Genomics of performance. *Equine Genomics*, ed Chowdhary BP (Wiley-Blackwell, Oxford, UK), pp 265-283.