Supporting Information

Warmuth et al. 10.1073/pnas.1111122109

SI Materials and Methods

Model Prediction of Heterozygosities of Wild and Domesticated Horses. This section describes the analytic model predicting pairwise within- and between-population homozygosities. The corresponding heterozygosities are calculated as $H = 1 - F$, where H is the heterozygosity and F is the homozygosity. Let $F_{ij}^{ww}(t,\mu)$ denote the expected homozygosity of a pair of alleles drawn randomly from two wild horses in demes i and j in generation t, under the infinite-alleles model with mutation probability μ per locus and generation (i.e., when each mutation gives a new allelic variant). Given the migration matrix $M_{ik}^{w}(t)$ (the probability that, in generation t, an individual in deme k emigrates to deme i) and the population size $N_i^w(t)$ of deme i in generation t, we can write a recursion for $F_{ij}^{ww}(t,\mu)$,

$$F_{ij}^{ww}(t+1) = (1-\mu)^2 \sum_{kl} M_{ik}^{w} M_{jl}^{w} \left[\frac{\delta_{kl}}{2N_k^w} + \left(1 - \frac{\delta_{kl}}{2N_k^w}\right) F_{kl}^{ww} \right], \quad [S1]$$

where the right-hand side is evaluated in generation t, and $\delta_{kl} = 1$ if $k = l$ and 0 otherwise.

When comparing two domesticated horses, or one wild and one domesticated horse, we describe the effect of migration and colonization in two stages. The homozygosity after migration in generation t is

$$\begin{aligned} \tilde{F}_{ij}^{dd}(t) &= (1-\mu)^2 \sum_{kl} M_{ik}^{d} M_{jl}^{d} \left[\frac{\delta_{kl}}{2N_k^d} + \left(1 - \frac{\delta_{kl}}{2N_k^d}\right) F_{kl}^{dd} \right] \\ \tilde{F}_{ij}^{wd}(t) &= (1-\mu)^2 \sum_{kl} M_{ik}^{w} M_{jl}^{d} F_{kl}^{wd}. \end{aligned} \quad [S2]$$

Here N_i^d is the effective population size in the domesticated horse population of deme i, and M_{ij}^d is the migration rate between the domesticated populations in demes i and j. The effect of establishment of newly colonized demes on the homozygosity of two domesticated horses is described by the following relations:

When $i = j$ and i is being colonized from deme k_i,

$$\begin{aligned} F_{ij}^{dd}(t+1) = \frac{(1-\mu)^2}{2c_d K_d} + \left(1 - \frac{1}{2c_d K_d}\right) \\ \times \left[q^2 \tilde{F}_{k_i k_i}^{dd} + 2q(1-q)\tilde{F}_{ik_i}^{wd} + (1-q)^2 \tilde{F}_{ii}^{ww} \right]. \end{aligned} \quad [S3]$$

When i is being colonized from deme k_i, and j is already colonized,

$$F_{ij}^{dd}(t+1) = q\tilde{F}_{k_i j}^{dd} + (1-q)\tilde{F}_{ij}^{wd}. \quad [S4]$$

When $i \neq j$ and both demes are being colonized (from demes k_i and k_j, respectively),

$$F_{ij}^{dd}(t+1) = q^2 \tilde{F}_{k_i k_j}^{dd} + q(1-q)\left(\tilde{F}_{ik_j}^{wd} + \tilde{F}_{jk_i}^{wd}\right) + (1-q)^2 \tilde{F}_{ij}^{ww}. \quad [S5]$$

Otherwise, $F_{ij}^{dd}(t+1) = \tilde{F}_{ij}^{dd}(t+1)$.

The corresponding relation for the homozygosity of a wild and a domesticated horse is

$$F_{ij}^{wd}(t+1) = \begin{cases} q\tilde{F}_{ik_j}^{wd} + (1-q)\tilde{F}_{ij}^{ww} & \text{when deme } j \text{ is being colonized} \\ \tilde{F}_{ij}^{wd} & \text{otherwise.} \end{cases} \quad [S6]$$

When the first domestic deme is colonized, we take $q = 0$, because in this case all horses must come from the local wild population.

To calculate the corresponding recursions for the stepwise mutation model (SMM) of microsatellite loci (1, 2), let t_{ij} be the number of generations to the most recent common ancestor of a pair of individuals from demes i and j. Under the SMM model the difference Δ in repeat count of two alleles is then the sum of $2t_{ij}$ independent identically distributed random variables, each of which is -1, 0, or 1 with probabilities $\mu/2$, $1 - \mu$, and $\mu/2$, respectively. Hence, the characteristic function for the difference in repeat number, $\langle e^{i\omega\Delta} \rangle$, is $(1 - \mu + \mu\cos\omega)^{2t_{ij}}$. It follows that the homozygosity under the SMM model is

$$F_{ij}^{\mathrm{SMM}}(t,\mu) = \frac{1}{2\pi} \int_0^{2\pi} \left\langle (1 - \mu + \mu\cos\omega)^{2t_{ij}} \right\rangle d\omega, \quad [S7]$$

where the angular brackets denote expectation over gene genealogies (3). Thus, it follows that the SMM homozygosity is related to the infinite-alleles homozygosity as

$$F_{ij}^{\mathrm{SMM}}(t,\mu) = \frac{1}{2\pi} \int_0^{2\pi} F_{ij}(t, \mu(1 - \cos\omega)) d\omega, \quad [S8]$$

where F_{ij} is any of F_{ij}^{ww}, F_{ij}^{wd}, or F_{ij}^{dd}. We used the following numerical approximation to evaluate the integral:

$$F_{ij}^{\mathrm{SMM}}(t,\mu) \approx \frac{1}{n} \sum_{k=1}^{n} F_{ij}\left(t, \mu\left(1 - \cos\frac{\pi(k-1/2)}{n}\right)\right). \quad [S9]$$

This approximation is very accurate when n is large enough that the probability of observing a difference of more than n repeat units can be ignored. Using $n = 50$ was enough to obtain machine precision for the parameters used in this study.

Validation of the Statistical Method. Because of the need to efficiently generate sample predictions, we could not use a fully stochastic method. We therefore approximated the stochastic samples by adding noise to the matrix of expected predicted homozygosities within and between populations. The noise is independent between parameter combinations but matrix elements within the matrix are correlated. We estimated this correlation structure using 10,000 bootstrap samples of the real data and generated the noise by sampling from the multivariate Gaussian distribution with these correlations and zero mean.

To validate our approach, we implemented a fully stochastic version of the part of the model describing the wild progenitors of domestic horses in our model and generated a synthetic dataset with the same number of loci, sample sizes, and population locations as in the real data, for parameter values $t = 10{,}000$, $r = 0.05$, $cK = 20$, $mK = 100$, $K = 50{,}000$, and $K_0 = 1{,}000$. We then performed the same analysis as for the full model: We estimated the correlations between matrix elements of pairwise homozygosity in the synthetic dataset, using 10,000 bootstrap iterations; ran a uniform parameter sweep of the parameter space, generating pairwise homozygosities from the predicted expected values (calculated using F_{ij}^{ww} values in the previous section, using the mutation rate $\mu = 1.5\ 10^{-4}$) and the correlated noise; and finally performed the same rejection sampling and GLM-ABC analyses using the ABCtoolbox (4) as for the full model (*Materials and Methods* in main text). Fig. S5 shows the resulting marginal distributions of parameter values from the rejection stage, the posterior distribution estimated by GLM-ABC, and the true parameter values. Some differences between true and fitted parameters are expected due to the limited amount of genetic data (26 nuclear microsatellite markers in 12 populations);

nevertheless, for most parameters the mode of the distribution is quite close to the true value of the parameter (except where a large range of values are compatible with the generated data, for example for the carrying capacity K).

1. Kimura M, Ota T (1975) Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proc Natl Acad Sci USA* 72:2761–2764.
2. Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc Natl Acad Sci USA* 75:2868–2872.
3. Eriksson A, Manica A (2011) Detecting and removing ascertainment bias in microsatellites from the HGDP-CEPH panel. *G3* 1:479–488.
4. Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: A versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* 11:116.

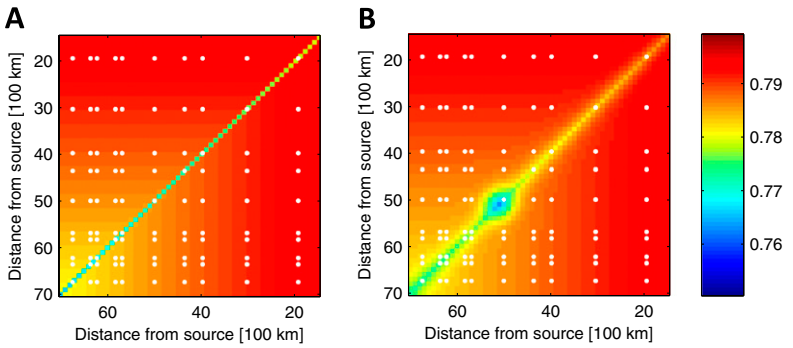

Fig. S1. Within- and between-population heterozygosity in wild and domestic horses as a function of distance from the origin of the expansion in East Asia. (*A*) Between-population heterozygosity in wild horses (off-diagonal elements) corresponds to a pattern of isolation-by-distance (IBD). The decline in within-population heterozygosity (on-diagonal elements) is relatively weak. (*B*) The demic component in the spread of horse domestication accentuated the east-to-west decline in within-population diversity (on diagonal), whereas the extensive incorporation of wild horses into domestic stock means that the original pattern of IBD (off diagonal) has been preserved in modern domestic horses from the steppes. The dip in within-population heterozygosity around 5,000 km reflects the strong bottleneck associated with the initial domestication of horses in the western steppe.

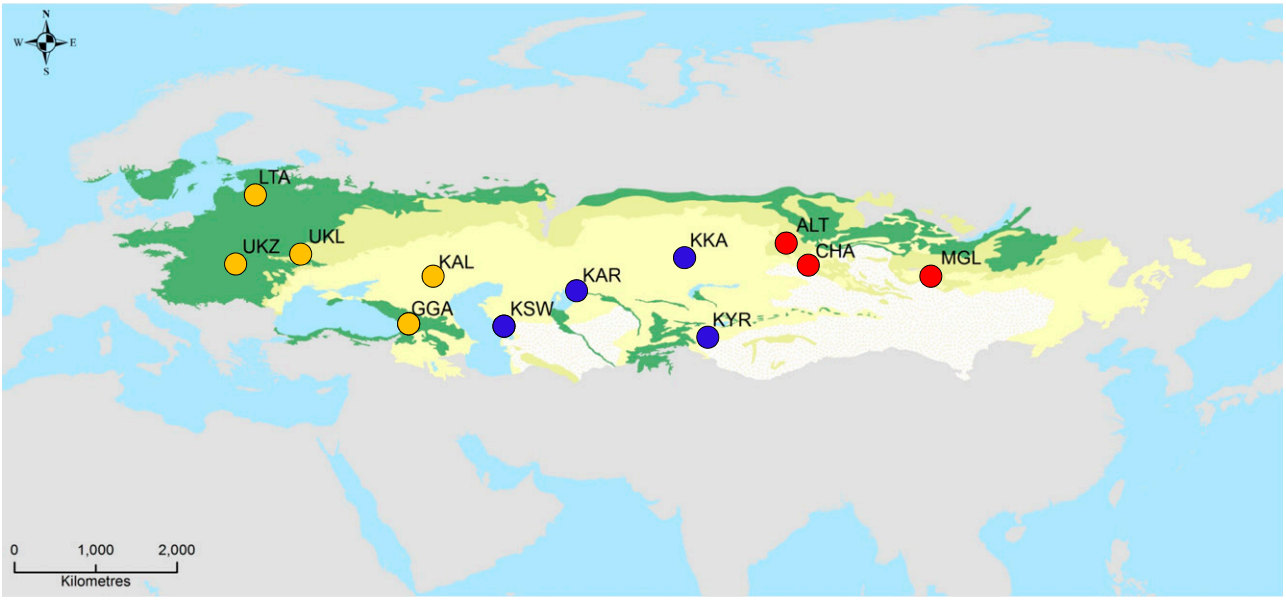

Fig. S2. Partitioning of sample populations into the three areas for calculating the summary statistics for ABC: Western Eurasia (orange circles), Central Eurasia (blue circles), and Eastern Eurasia (red circles). See Fig. 1 of the main text for definitions of abbreviations.

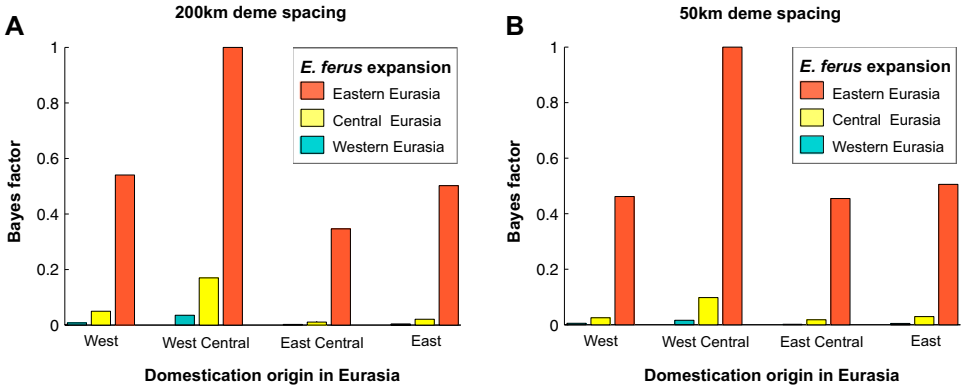

Fig. S3. Bayes factors for 12 scenarios describing the origin of the wild progenitor of domestic horses, *Equus ferus*, and the origin of horse domestication in Eurasia, for 200-km (*A*) and 50-km (*B*) deme spacing. Bayes factors are relative to the best-supported scenario, which posits an origin of *E. ferus* in eastern Eurasia and a domestication origin in western central Eurasia. The scenarios are grouped by domestication origin and colored by the origin of *E. ferus*.

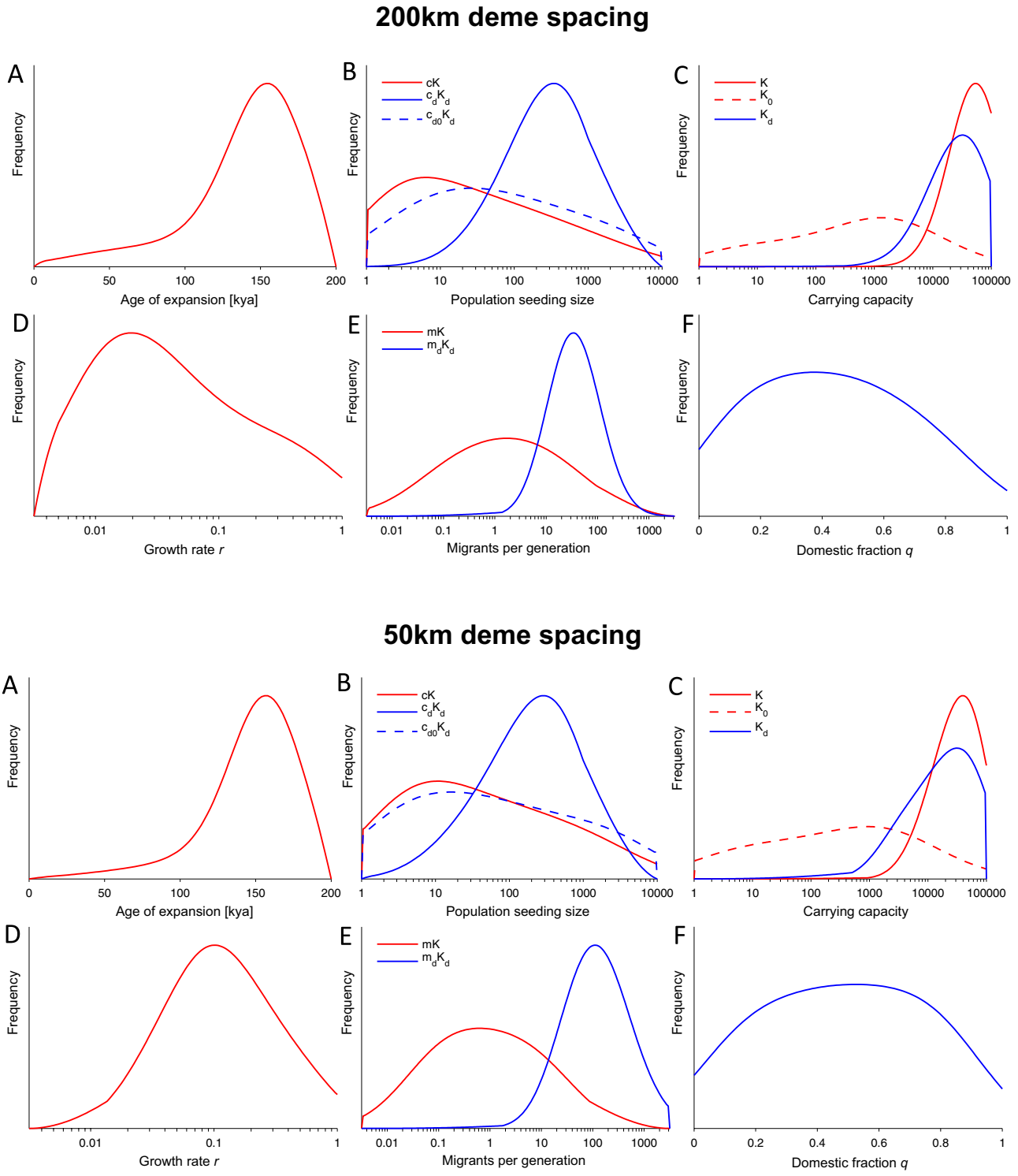

Fig. S4. Posterior distributions of the demographic parameters of wild and domestic populations, from ABC analysis, for demes spaced 200 km (*Upper*) and 50 km (*Lower*) apart. For both spacings, the parameters are shown in *A*–*F*, as follows: (*A*) timing (*t*) of the expansion of *E. ferus* out of eastern Eurasia; (*B*) effective founder population sizes of wild horse (*cK*, red line), the first domestic population ($c_{d0}K_d$, blue dashed line), and subsequent domestic populations (c_dK_d, blue solid line); (*C*) carrying capacity (effective population sizes) of the ancestral (K_0, dashed red line), wild (*K*, solid red line), and domestic populations (K_d, blue line); (*D*) growth rate, *r*, of wild horse populations; (*E*) migration between populations of wild (*mK*, red line) and domestic horses (m_dK_d, blue line); and (*F*) the proportion of domestic (*q*) and wild (1 − *q*) horses in domestic horse founder populations.

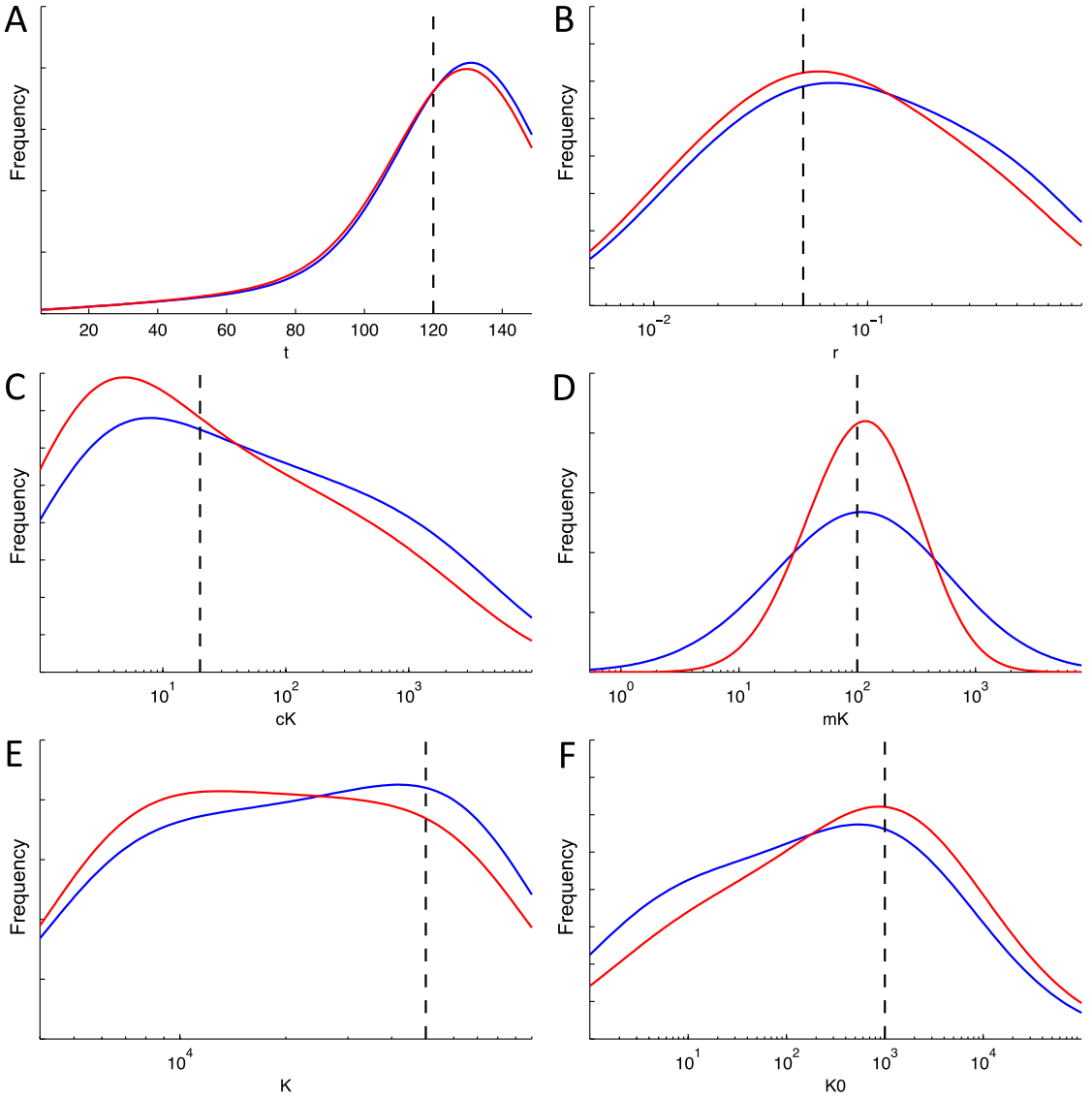

Fig. S5. Rejection sampling and posterior distributions from GLM-ABC. Blue curves are marginal distributions from rejection sampling for the model parameters from a uniform parameters sweep, red curves are posterior distributions from GLM-ABC, and the dashed black lines show the parameter values used to produce the synthetic dataset, using the stochastic implementation of the model. (*A*) Timing, *t*, of the expansion; (*B*) growth rate, *r*; (*C*) number of colonists, *cK*; (*D*) number of migrants per generation, *mK*; (*E*) carrying capacity, *K*; and (*F*) ancestral carrying capacity, K_0.

Table S1. Details of the sampled populations

Country	Region	ID	Latitude	Longitude	d_a, km*	$H^{\dagger}$	$N^{\ddagger}$
Mongolia	Övörkhangai	MGL	48.0	101.0	1,951	0.791	44
China	Xinjiang	CHA	48.7	87.0	3,035	0.790	34
Russia	Altai	ALT	51.6	85.0	3,041	0.776	40
Russia	Kalmykia	KAL	47.5	45.3	5,833	0.778	22
Kyrgyzstan	Naryn	KYR	41.1	75.7	4,359	0.786	20
Kazakhstan	Mangystau	KSW	42.3	53.2	5,687	0.769	24
Kazakhstan	Kyzylorda	KAR	46.0	61.3	4,997	0.769	35
Kazakhstan	Karagandy	KKA	50.0	73.0	3,981	0.775	25
Ukraine	Lviv	UKL	50.3	30.9	6,359	0.782	21
Ukraine	Zakarpattia	UKZ	49.2	23.6	6,739	0.772	18
Lithuania	Vilnius	LTA	56.9	25.4	6,223	0.756	21
Georgia	Samegrelo	GGA	42.3	42.3	6,361	0.777	24

See Fig. 1 for definitions of abbreviations.
*Distance to the easternmost deme (deme 0), in kilometers.
†Expected within-population heterozygosity.
‡Sample size.

Table S2. Summary statistics for the posterior distributions of each estimated model parameter

			95% CI*		95% HPD†	
Parameter	Mode	Median	Lower	Upper	Lower	Upper
T, kya	160	150	51	180	79	180
r	0.052	0.063	0.0076	0.72	0.0078	0.69
cK	7.7	28	1.2	4,000	1.0	2,300
mK	1.0	1.1	0.012	59	0.02	79
K	54,000	39,000	7,900	94,000	11,000	100,000
K_0	950	360	1.6	42,000	1.4	31,000
c_dK_d	340	180	8	900	15	1,000
$c_{d0}K_d$	6.6	19	1.2	640	1.0	430
m_dK_d	72	72	21	260	21	260
K_d	6,400	4,700	1,100	9,600	1,500	10,000
q	0.47	0.46	0.04	0.91	0.01	0.88

*Credibility Intervals.
†Highest posterior density intervals, the shortest continuous intervals with an integrated posterior density of a certain value.

Table S3. List of microsatellite loci amplified in two multiplex PCR reactions

Locus	ECA*	Primer 5′–3′	Reference	Size range	Multiplex
VHL20	30	CAAGTCCTCTTACTTGAAGACTAG AACTCAGGGAGAATCTTCCTCA	(1)	82–102	1
HTG4	9	CTATCTCAGTCTTGATTGCAGGAC GCTCCCTCCCTCCCTCTGTTCTC	(2)	123–137	1
AHT4	24	AACCGCCTGAGCAAGGAAGT GCTCCCAGAGAGTTTACCCT	(3)	151–169	1
HMS7	1	CAGGAAACTCTCATGTTGATACCATC GTGTTGTTGAAACATACCTTGACTGT	(4)	172–186	1
COR18	25	AGTCTGGCAATATTGAGGATGT AGCAGCTACCCTTTGAATACTG	(5)	263–277	1
AHT5	8	ACGGACACATCCCTGCCTGC GCAGGCTAAGGAGGCTCAGC	(3)	125–141	1
HMS6	4	CTCCATCTTGTGAAGTGTAACTCA GAAGCTGCCAGTATTCAACCATTG	(4)	159–171	1
ASB23	3	ACATCCTGGTCAAATCACAGTCC GAGGGCAGCAGGTTGGGAAGG	(6)	183–215	1
TKY312	6	AACCTGGGTTTCTGTTGTTG GATCCTTCTTTTTATGGCTG	(7)	100–126	1
TKY343	11	TAGTCCCTATTTCTCCTGAG AAACCCACAGATACTCTAGA	(8)	143–173	1
LEX33	4	TTTAATCAAAGGATTCAGTTG GGGACACTTTCTTTACTTTC	(9)	191–217	1
HMS3	9	CCAACTCTTTGTCACATAACAAGA GCCATCCTCACTTTTTCACTTTGTT	(10)	151–171	1
COR58	12	CACCAGGCTAAGTAGCCAAG GGGAAGGACGATGAGTGAC	(10)	210–234	1
HMS5	5	TAGTGTATCCGTCAGAGTTCAAGG GCAAGGAAGTCAGACTCCTGGA	(4)	98–104	2
EB2E8	26	TTCTGTGTTAGGGGTTGTG GTATGAGCCAGTTCTTGAT	(11)	125–139	2
TKY321	20	TTGTTGGGTTTAGGTATGAAGG GTGTCAATGTGACTTCAAGAAC	(7)	182–208	2
ASB2	15	CACTAAGTGTCGTTTCAGAAGG GCACAACTGAGTTCTCTGATAGG	(6)	216–248	2
TKY301	23	AATGGTGGCTAATCAATGGG GTGTATGATGCCCTCATCTC	(7)	149–169	2
TKY337	4	AGCAGGGTTTAATTACCGAG TAGATGCTAATGCAGCACAG	(8)	169–189	2
TKY374	1	CTGGTCCCTCTGGATGGAAG TCCCAAGAGGGAGTACAATC	(7)	197–225	2
HTG7	4	CCTGAAGCAGAACATCCCTCCTTG ATAAAGTGTCTGGGCAGAGCTGCT	(12)	113–123	2
UM11	20	TGAAAGTAGAAAGGGATGTGG GTCTCAGAGCAGAAGTCCCTG	(13)	162–184	2
TKY394	24	GCATCATCGCCTTGAAGTTG CCTTTCTGGTTGGTATCCTG	(8)	232–258	2
UM32	14	AAATGGTCAGCCTCTCCTC TGTCTCTCTAGTCCCACTCCTC	(14)	140–150	2
HMS1	15	CATCACTCTTCATGTCTGCTTGG TTGACATAAATGCTTATCCTATGGC	(4)	170–182	2
TKY294	27	GATCTATGTGCTAGCAAACAC CTAGTGTTTCAGATAGCCTC	(7)	216–230	2

**Equus caballus* chromosome number.

1. van Haeringen H, Bowling AT, Stott ML, Lenstra JA, Zwaagstra KA (1994) A highly polymorphic horse microsatellite locus: VHL20. *Anim Genet* 25:207.
2. Ellegren H, Johansson M, Sandberg K, Andersson L (1992) Cloning of highly polymorphic microsatellites in the horse. *Anim Genet* 23:133–142.
3. Binns MM, Holmes NG, Holliman A, Scott AM (1995) The identification of polymorphic microsatellite loci in the horse and their use in thoroughbred parentage testing. *Br Vet J* 151: 9–15.
4. Guérin G, Bertaud M, Amigues Y (1994) Characterization of seven new horse microsatellites: HMS1, HMS2, HMS3, HMS5, HMS6, HMS7 and HMS8. *Anim Genet* 25:62.
5. Hopman TJ, et al. (1999) Equine dinucleotide repeat loci COR001-COR020. *Anim Genet* 30:225–226.
6. Breen M, et al. (1997) Genetical and physical assignments of equine microsatellites—first integration of anchored markers in horse genome mapping. *Mamm Genome* 8:267–273.
7. Tozaki T, et al. (2001a) Population study and validation of paternity testing for Thoroughbred horses by 15 microsatellite loci. *J Vet Med Sci* 63:1191–1197.
8. Tozaki T, et al. (2001b) Characterization of equine microsatellites and microsatellite-linked repetitive elements (eMLREs) by efficient cloning and genotyping methods. *DNA Res* 8: 33–45.

9. Coogle L, Reid R, Bailey E (1996) Equine dinucleotide repeat loci from LEX025 to LEX033. *Anim Genet* 27:289–290.
10. Ruth LS, et al. (1999) Equine dinucleotide repeat loci COR041-COR060. *Anim Genet* 30:320–321.
11. Gralak B, Zurkowski M, Niemczewski C, Coppieters W (1994) The preliminary study on linkage between two horse microsatellites and genes coding for blood groups and some blood proteins. *Anim Genet* 25:285–286.
12. Marklund S, Ellegren H, Eriksson S, Sandberg K, Andersson L (1994) Parentage testing and linkage analysis in the horse using a set of highly polymorphic microsatellites. *Anim Genet* 25:19–23.
13. Meyer AH, Valberg SJ, Hillers KR, Schweitzer JK, Mickelson JR (1997) Sixteen new polymorphic equine microsatellites. *Anim Genet* 28:69–70.
14. Swinburne J, et al. (2000) First comprehensive low-density horse linkage map based on two 3-generation, full-sibling, cross-bred horse reference families. *Genomics* 66:123–134.