Paternal Origins and Migratory Episodes of Domestic Sheep

Highlights
- Novel ovine SNPs of the male-specific region of Y chromosome were developed
- Y chromosome of domestic sheep contains four different paternal lineages
- Lineages C and B predominate in breeds of primitive traits and fat tail, respectively
- Expansions of sheep correlate with various phenotypic traits and breeding goals

Authors
Juan Deng, Xing-Long Xie, Dong-Feng Wang, ..., Hong-Ping Zhang, Li Li, Meng-Hua Li

Correspondence
menghua.li@cau.edu.cn

In Brief
Deng et al. show that domestic sheep harbor four Y chromosome lineages and early expansions of sheep were associated with the segregation of primitive and fat-tailed phenotypes as well as traits selected for different purposes.
Paternal Origins and Migratory Episodes of Domestic Sheep

Juan Deng,1,2,39 Xing-Long Xie,1,3,39 Dong-Feng Wang,1,3 Chao Zhao,1,4 Feng-Hua Lv,1,5 Xin Li,1,3 Ji Yang,1,5 Jia-Lin Yu,6 Min Shen,1,6 Lei Gao,7,8 Jing-Quan Yang,7,8 Ming-Jun Liu,7,8 Wen-Rong Li,9 Yu-Tao Wang,10 Feng Wang,11 Jin-Quan Li,12 Eer Hehua,13 Yong-Gang Liu,14 Zhi-Qiang Shen,14 Yan-Ling Ren,15 Guang-Jian Liu,16 Ze-Hui Chen,15 Neena A. Gorkhali,17 Hossam E. Rushdi,18 Hosein Salehian-Dehkordi,19 Ali Esmailizadeh,19 Maryam Nosrati,20

SUMMARY

The domestication and subsequent global dispersal of livestock are crucial events in human history, but the migratory episodes during the history of livestock remain poorly documented [1–3]. Here, we first developed a set of 493 novel ovine SNPs of the male-specific region of Y chromosome (MSY) by genome mapping. We then conducted a comprehensive genomic analysis of Y chromosome, mitochondrial DNA, and whole-genome sequence variations in a large number of 595 rams representing 118 domestic populations across the world. We detected four different paternal lineages of domestic sheep and resolved, at the global level, their paternal origins and differentiation. In Northern European breeds, several of which have retained primitive traits (e.g., a small body size and short or thin tails), and fat-tailed sheep, we found an overrepresentation of MSY lineages y-HC and y-HB, respectively. Using an approximate Bayesian computation approach, we reconstruct the demographic expansions associated with the segregation of primitive and fat-tailed phenotypes. These results together with archaeological evidence and historical data suggested the first expansion of early domestic hair sheep and the later expansion of fat-tailed sheep occurred ~11,800–9,000 years BP and ~5,300–1,700 years BP, respectively. These findings provide important insights into the history of migration and pastoralism of sheep across the Old World, which was associated with different breeding goals during the Neolithic agricultural revolution.

RESULTS AND DISCUSSION

The Neolithic agricultural revolution triggered the transition from a hunter-gatherer nomad lifestyle to a sedentary farming society, and a subsequent spreading of agriculture and pastoralism across Asia, Europe, and Africa during ~10,000–5,000 years before present (BP) [4, 5]. Previous molecular studies have contributed to the reconstruction of domestication and
migrations of farm animals in terms of time, place, and driving forces [1, 5–6]. Sheep was one of the first species domesticated in the Fertile Crescent ~12,000–10,000 years BP [9, 10] and initially was bred as a source of meat [11]. Domesticated sheep accompanied human migrations worldwide (Figure 1A; Data S1) and acted as a key factor in the social transformation between sedentary agrarian, seminomadic, and nomadic pastoralist societies [12–15]. In addition to the initial dispersal, archaeological records show at least other population expansions, when sheep were bred for wool and for fat tails as adaptation to a dry climate during the late Holocene (6,000–3,000 years BP) (Figure 1B; Data S1).

Autosomal and mitochondrial DNA (mtDNA) variations as well as whole-genome sequences have been widely used in revealing the domestication, migration, and genetic diversity of sheep [12, 16–18] and other farm animals [7, 8, 19, 20]. Variation of the male-specific region of the Y chromosome (MSY) can be particularly powerful because it represents the paternal lineage and does not recombine. However, it is hindered by the abundance of ampliconic regions with multicopy genes and by the heterochromatic repeat units [21, 22]. As a consequence, only few reports have studied variation in the MSY of sheep on the basis of three markers [23–26]. In this study, we used whole-genome sequences to generate a unique large set of novel MSY SNPs in ovine species. To elucidate the history of sheep domestication and their early dispersions, MSY SNP genotypes were analyzed in combination with mtDNA variations and whole-genome sequences in a large number of rams across the world (Data S2). These include populations from Northern Europe, Eastern Europe, UK, Southern Asia, Central Asia, Mongolia, Russia, Africa, and Southern America, which were underrepresented in earlier work but essential to assess ovine population expansions. More specifically, we aim to answer several key questions: (1) how many paternal lineages exist in domestic sheep, (2) are sheep with distinct phenotypic traits and their migration episodes associated with the segregation of different lineages, and (3) when did the expansion event(s) occur?

From the high-depth whole-genome sequences of 31 rams and 5 ewes (NCBI: PRJNA624020) [27] with a wide geographic origin ranging from China and Afghanistan to the Netherlands, Sweden, UK, and France (average coverage = 24.7 ×; Data S2), we extracted a reference panel of 467 reliable single-copy MSY (scpMSY) contigs (501 kb) covering 10 Mb assembly by using a modified method in Wallner et al. [28] (Figure 2). To detect variants, we mapped whole-genome sequences of 161 rams and 32 ewes from 88 domestic populations, and 18 rams from three wild species [nine Asiatic mouflons (O. orientalis), eight argali (O. ammon), and one urial (O. vignei)]; average coverage = 20.3 ×; NCBI: PRJNA624020 and PRJNA645671; Data S2) to the scpMSY reference. In total, we called 85 novel MSY SNPs in domestic sheep and 425 in the wild species (Data S3), 15 of which were shared by wild and domestic sheep. In the 179 domestic and wild rams, a total of 63 haplotypes were defined based on genotypes of the 495 SNPs including...
the two previously reported SNPs oY1 and oY2 [26] (Data S4). Different phylogenetic reconstruction methods yielded similar phylogenetic trees, in which two haplotypes of Asiatic mouflons are closely related to those of domestic sheep (Figure S1A). Diagnostic SNPs evidenced the segregation of four distinct haplogroups (y-HA, y-HB, y-HC, and y-HD) in domestic rams (Figure S1A; Data S4), all of which were supported by high bootstrap values.

For the 179 rams, we constructed Bayesian trees based on the MSY SNPs, mitogenomes, and a neighbor-joining (NJ) tree based on whole-genome sequences (Figure 3). Two MSY haplotypes from four Asiatic mouflons (TH.2, SH.7, 266, and 267) are similar to the y-HB haplotypes from domestic sheep (Figure S1B). However, the haplotypes of five other mouflons cluster with an urial haplotype, suggesting gene flow from urial into the Asiatic mouflon population in Iran [17]. Previous field and genetic evidence have shown the presence of hybrids between urial and Asiatic mouflon in this country [29, 30]. Inter-species introgressions have also been detected in several genera [31], including the wild and domestic sheep species [13, 32]. We observed different patterns of MSY, mtDNA, and whole-genome phylogenies among the wild and domestic sheep. Only mt-HA and mt-HB were detected in the individuals of MSY lineage y-HC; conversely, the sheep of mt-HC showed paternal lineages of y-HA and y-HB (Figures S1B, S1C, S1E, and S1F). A phylogenetic tree of the whole-genome sequences shows a significant phylogeographic pattern by clustering sheep according to their continents with the exception of six Asian fine-wool (Merino-like) sheep that are attached to the European cluster (Figures 3C and S1D).

The three major MSY haplogroups differ in Watterson $\theta$ ($\theta_W$) values: $\theta_W = 2.47 \times 10^{-8}$ for y-HB, $2.32 \times 10^{-8}$ for y-HA, and $1.04 \times 10^{-8}$ for y-HC, about 200 times lower than those found for the three major mtDNA lineages of the same samples (mt-HA, $\theta_W = 4.39 \times 10^{-8}$; mt-HB, $7.17 \times 10^{-9}$; mt-HC, $4.19 \times 10^{-8}$). Our estimates of $\theta_W$ for MSY haplogroups are similar to those reported for domestic Bactrian camels ($\theta_W = 1.17 \times 10^{-9}$) [33] and domestic stallions ($\theta_W = 0.79 \times 10^{-9}$) [28]. Using an empirical Bayesian approach [34], we detected more recent estimates of divergence times among MSY haplogroups than for mtDNA lineages from the same individuals (Figures 3A and 3B).

Using the $T_{\text{MRCA}}$ (time to the most recent common ancestor) of argali and urial with domestic sheep, we estimate that the overall mutation rate is $0.93 \times 10^{-10}$ mutations per generation per site [95% highest posterior density interval (95% HPDI): $0.83 \times 10^{-10}$–$1.03 \times 10^{-10}$], which is around 50-fold lower than that for the mtDNA (95% HPDI: $4.8 \times 10^{-9}$–$5.4 \times 10^{-9}$). The mutation rate of the sheep MSY is slightly lower than other species: $1.4$–$4.2 \times 10^{-10}$ for dog [35–38], $1.7 \times 10^{-8}$–$4.2 \times 10^{-10}$ for horse [28, 39], and $1.0$–$1.9 \times 10^{-9}$ for wolf [40, 41].
followed by a reduction after domestication (~10,000–5,000 years BP), and a more drastic decrease after 5,000 years BP (Figures 3A and 3B). During the last 2,000 years, the $N_e$ estimates based on mitogenomes from the same 161 sheep were 47–239-fold higher than those based on the Y chromosome, with a gradual increasing trend in the observed differences (Figures 3A and 3B).

We then analyzed an expanded sample set of 595 rams representing 118 local populations from across the world (Data S2). For the additional samples, the 81 MSY SNPs were genotyped
using the Hi-SNP genotyping platform (Data S4) and the mtDNA D-loop region (AF010406, 15,522–16,318) was sequenced (GenBank: MT768709–MT769136). We observed all paternal and maternal lineages and similar phylogenetic resolution as above (Figures S1E and S1F). The MSY y-HA showed the highest frequency (71.1%), followed by y-HB (17.5%) and y-HC (10.8%), while y-HD was only present in four animals (0.67%) of Djallonke and Cameroon populations from Western Africa (Data S4).
Lineage y-HA is the dominant lineage in most populations (Figure 4A), while y-HB haplotypes are mainly distributed throughout semi-desertic and steppe regions in the Near East, Central Asia, Indian subcontinent, Northern Asia, and Northern Africa. Also, y-HB has a high frequency in southwest Asian and east-Mediterranean fat-tailed populations such as Barki, Chios, and Kermani (Data S5). Notably, y-HB and fat-tailed sheep have similar distribution areas [42] (Figure 4A). Out of the 39 populations carrying y-HB, 27 are fat-tailed sheep, suggesting a significant association between y-HB and the fat-tailed phenotype (Pearson $\chi^2 = 7.64, df = 1, p = 0.006$) and a major influence of the colonization of fat-tailed sheep on the expansion of y-HB.

Lineage y-HC is present in Spain, England, Scandinavia, Western Russia, the Qinghai-Tibetan Plateau, and the Yunnan-Kweichow Plateau of China. The oY1 alleles in 3,056 modern samples from 214 populations reported in previous studies (Data S6) revealed a similar geographic pattern for oY1(G) (Figure 4B). Although we found oY1(G) in the two Catalanian Xisqueta and Ripollesa rams, the frequency of y-HC in these and other Spanish breeds is low (Data S6). The close similarity of the Chinese and European y-HC sequences (Figure S1B) indicate that the occurrence of y-HC in China may be partially attributed to recent introgression of English rams, which are worldwide popular breeding sires and have been used in the Yunnan Province of China [43]. In six ancient DNA samples from Finland (Medieval and post-Medieval) and Estonia (late-Chalcolithic and Medieval), only allele oY1(G) has been detected [44, 45], suggesting an ancient presence of y-HC in Northern Europe. In the same region, several sheep breeds have retained primitive features such as a small body size and short or thin tails [5], whereas a relatively high component of their genomes (up to 20%) originates from the European mouflon [46]. Thus, the range of y-HC suggests possibly an ancient expansion of primitive Northern European sheep, which became the paternal ancestors of the y-HC-carrying English mutton breeds [24], and several of them became transboundary breeds [47].

The pattern of geographic distribution may reflect ancient male founder effect and the small effective population size of males within breeds. This has led to a high frequency of y-HC in Northern Europe and of y-HB within the fat-tailed sheep. The founder effect may very well have been stimulated by selection and, for instance, have played a role in the spreading of the fat-tailed phenotype. The current evidence does not allow associating the secondary spread of wool sheep throughout the Old World with lineage y-HA or with haplotypes within this predominant lineage.

We analyzed the selected whole-genome sequences by approximate Bayesian computation (ABC) in order to obtain time estimates for the demographic expansions of the first domestic sheep and the fat-tailed sheep, which were assumed to be associated with the MSY haplogroups y-HC and y-HB, respectively. We used high-depth (~15.54–36.93 x coverage) whole-genome sequences from 15 Asiatic mouflons, 16 Northern European, and 18 native fat-tailed sheep (Figure S2A; Data S2). We tested two alternative models for their expansions (Figures 4D and S2) [48]. In the best fitting model, the ancestors of Northern European primitive breeds, most of which carry y-HC, spread ~10,500 years BP ($T_1$, 50% HPDI: 9,000–11,800; Figure S2F). This indicates that y-HC probably expanded as part of the dispersal of the first domestic sheep, which had a hair coat. A later spread was possibly associated with the expansion event of fat-tail sheep ~3,400 years BP ($T_2$, 50% HPDI: 1,700–5,300; Figure S2F) and involved y-HB. We found that the rate of gene flow from the fat-tailed sheep to the first hair-coat domesticates ($m_2 = 2.64 \times 10^{-5}$, 50% HPDI: $5.56 \times 10^{-6}$–1.68 $\times 10^{-4}$; Figure S2F) is higher than the rate in the opposite direction ($m_1 = 8.97 \times 10^{-6}$, 50% HPDI: 2.42 $\times 10^{-6}$–5.32 $\times 10^{-5}$), which indicated a net introgression into hair sheep introducing the coarse-wool trait in most of the y-HC-carrying breeds.

These findings provided genetic support to the dispersal episodes of the first domestic hair sheep and of the fat-tailed sheep as evidenced by archeological records (Figures 1A and 1B; Data S1) and previous genetic results [5]. Although both the estimates of time under the Bayesian approximate model should be interpreted with caution because of the potential biases in the MSY mutation rate applied, the ABC results show that the fat-tailed sheep dispersed long after the spreading of the first domestic sheep and also after the emergence of the wool sheep according to the archeological evidence (Figure 1B).

The first migratory wave of sheep populations, as inferred from the distribution of lineages y-HC and y-HD, spread to the borders of Europe and Africa as well as the highlands of Asia ~10,000 years BP. Nowadays, remnants of genetic material from the first expansion persist in Northern European sheep, which have retained a mouflon-like morphology [5], and have an Iron Age origin [49, 50] and a high proportion of their genomes originating from the European mouflon [46]. The topology of the whole-genome sequence tree (Figures 2C and S1D) suggests that these populations are related to other primitive sheep, such as the Dutch or German Heath sheep, the German Rhön, and the French Soloigne and Ouessant [43], which carry y-HA. The Y chromosome of the feral European mouflons, which descend from the first domestic sheep, does not carry y-HC (Data S6). This finding...
indicates that the first introduction of domestic sheep did not carry exclusively y-HC. Tibetan sheep also have primitive features but originated ~3,000 years BP [13], which is later than the earliest European sheep (Figure 1A). They carry y-HA in addition to y-HC that could have possibly been partly introduced by English rams [43]. Their whole-genome sequences cluster with those of the Asian fat-tailed sheep, suggesting gene flow between Asian breeds with different tail types. Given the multiple episodes of movement and interaction, modern populations are unlikely to act as direct proxies for the first wave of sheep that were introduced to a region. Admixture, extinction, turnover, drift, and cultural and biological processes have constantly shuffled the proportions of genetic markers in modern sheep populations in their history [51].

Our results provide genetic evidence for distinct introductions and dispersion histories for the African thin-tailed and fat-tailed populations, as previously suggested by archeological studies [52]. The fat-tailed Barki and Rahmani sheep (Egypt) and Barbary and Moboro sheep (Libya) possess the y-HB haplotypes, which are in agreement with evidence based on 50k SNP genotypes [53], indicating an immigration of Southwest Asian sheep [52]. The thin-tailed hair sheep (Mossi, Sahelian, Cameroon, and Djallonke sheep) in Western Africa can be inferred to be the first to migrate from their Near Eastern domestication center, entering the African continent via the Isthmus of Suez or across the Bab-El Mandeb strait and migrating to Western Africa overland [52]. In particular, the Djallonke and its derived population of Cameroon sheep showed an unusual y-HD haplotype, which is phylogenetically close to the haplotypes of Asiatic mouffons (Figures S1B, S1C, S1E, and S1F). Chessa et al. [5] found the same endogenous retroviral signature in the primitive European populations and the Djallonke sheep.

Overall, MSY haplotypes for the great majority of sheep populations grouped together and showed a high frequency of y-HA. Besides this, all the Merino-derived fine wool breeds carry the MSY y-HA (Data S3). Previous genetic evidence [5], together with archaeological data and historical records, suggested that the specialization for secondary products such as wool may have spurred a second expansion of sheep populations around 8,000–7,000 years BP, most likely from Southwest Asia at first [5, 54] (Figure 1B; Data S1). Thus, selection of sires from wool breeds may explain a high level of historic admixture [47, 55].

Different sheep varieties emerged over the following few thousand years. Fat-tailed sheep are known for their fat storage in their large tails and hindquarters, which provide a reserve of energy for survival in harsh production conditions with nutritional stress [36, 57]. Therefore, selection for the fat-tailed phenotype in domestic sheep and their expansion could have been initiated and accelerated by extreme climatic conditions and/or human preferences [58]. This has probably driven the third expansion in domestic sheep history from the Middle East to Northern Africa, Central and Eastern Asia, and the eastern edge of Europe, involving both y-HA and y-HB rams. The earliest archeological evidence for the history of fat-tailed sheep indicated the first description on an Uruk II stone vessel in the Near East (~5,000 years BP) [59] and the rock paintings in Ethiopia (~3,000–2,600 years BP) [60–62] (Figure 1B; Data S1). In addition, historical data record the presence of fat-tailed sheep in Central Asia since the Tsin Dynasty (266–420 AD) [63] (Data S1). In approximately the same time period, human-mediated widespread introgression of aridity-adapted zebu (Bos indicus) bulls has enhanced herd survival in tropical regions [8]. The high frequencies of y-HB in modern sheep from regions of extreme environments, such as the dryland regions in Iran, Pakistan, and Afghanistan and the cold regions of Altay, Mongolia, and Siberia, coupled with archeological evidence, suggest that selection for sheep with fat tails occurred most likely first in the Near East. Thus, our data provide the first genetic evidence for a Southwest-Asian origin of fat-tailed sheep, which subsequently spread into specific regions of Asia and Africa. The observation of one y-HB haplotype (s-H22; Figure 4C) associated with the fat-tail phenotype in four European thin-tailed populations (Leccese, White Mountain, Brown Mountain, and Sumavska) may suggest introgression of fat-tailed sheep from Northern Africa via the maritime Mediterranean route [64]. In future studies, genome data should be screened for the specific traits in sheep, overcoming the limitations of neutral markers.

In conclusion, the global distribution of MSY lineages coupled with demographic reconstruction based on whole-genome sequences reveals multiple and separate waves of expansions of domestic sheep, which possibly correlate with various phenotypic traits and breeding goals for different products. At least the first domestic expansion and the dispersal of the fat-tailed sheep could have started in Southwest Asia. Our results provide important insights into the history of sheep migration and pastoralism across the Old World.

STAR METHODS

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

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
- **METHOD DETAILS**
  - Archeological Records and Historical Data
  - Samples and Sequencing
  - Identification of MSY SNPs
  - Assembly of Mitogenomes
  - Whole-genome SNP Calling
  - MSY SNPs Genotyping
  - MIDNA D-Loop Sequencing
  - Phylogenetic Analyses
  - Geographic Patterns of the MSY Variability
  - Approximate Bayesian Computation Analysis
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2020.07.077.

ACKNOWLEDGMENTS

We are grateful to the late Zhi-Fa Wang, and Joaquim Casellas and Jordi Jordana from the Universitat Autònoma de Barcelona, and several others for...
sampling. We thank Xiao-Xiao Zhang and An-Ni Yu for laboratory work and Thor Blichfeldt and the Norwegian Association of Sheep and Goat Farmers (NSG) for providing samples of Norwegian sheep. Thanks are also due to Ama-
dou Traore (Institut de l’Environnement et de Recherches Agricoles [INERA], Ouagadougou, Burkina Faso), Masroor Ellahi Babar (Virtual University, Lahore, Pakistan), Pradeepa Silva (University of Peradeniya, Peradeniya, Sri Lanka), Seyed Abbas Rafat (University of Tabriz, College of Agriculture, Tabriz, Iran), Thirunvenkadan A.K. and Saravanan Ramasamy (Tamil Nadu Veterinary and Animal Sciences University, Chennai, India) for their help in sample collection. This study was funded by grants from the National Key Research and Devel-
opment Program-Key Projects of International Innovation Cooperation be-	ween Governments (2017YFE0117900), the National Natural Science Foun-
dation of China (nos. 31825024, 31661143014, 31760661, and 31972527), the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) (no. 2019QZKK0501), the External Cooperation Program of Chinese Academy of Sciences (152111KYSB20190027), the Taishan Scholars Pro-
gress Program of China (no. ts201511085), and Chinese Government contribution to CAAS-ILRI Joint Laboratory on Livestock and Forage Genetic Resources in Beijing and to ICARDA. The paper contributes to the CGIAR Research Program on Livestock and Fish.

AUTHOR CONTRIBUTIONS

M.-H.L. conceived the project and supervised the work; J.D. and X.-L.X. per-
fomed the majority of analysis and laboratory work with contribution from D.-F.W.; all the other co-authors provided the samples. J.D. wrote the manu-
script with contributions from M.-H.L., J.-L.H., and J.A.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES


lations reveals strong partitioning of diversity and highlights post-domes-

sity of modern, ancient and wild sheep (Ovis gmelinii anatolica) from Turkey: new insights on the evolutionary history of sheep. PLoS ONE 8, e19592.


## STAR METHODS

### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biological Samples</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole blood from urial</td>
<td>Iran</td>
<td>N/A</td>
</tr>
<tr>
<td>Whole blood from argali</td>
<td>China</td>
<td>N/A</td>
</tr>
<tr>
<td>Blood samples from modern domestic sheep</td>
<td>Asia, Europe, Africa, South America</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Critical Commercial Assays</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hi-SNP genotyping platform</td>
<td>Novogene Bioinformatics Institute</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>DEPOSITED DATA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheep reference genome Oar v4.0</td>
<td>[65]</td>
<td><a href="https://www.ncbi.nlm.nih.gov/assembly/GCF_000298735.2">https://www.ncbi.nlm.nih.gov/assembly/GCF_000298735.2</a></td>
</tr>
<tr>
<td>Whole genome next-generation sequences (NGS) of 3 rams analyzed for raw MSY assembly</td>
<td>[27]</td>
<td>GenBank/NCBI PRJNA624020</td>
</tr>
<tr>
<td>Whole genome next-generation sequences (NGS) of 31 rams and 5 ewes analyzed for scpMSY SNP calling</td>
<td>This paper and [27]</td>
<td>GenBank/NCBI PRJNA624020 and PRJNA645671</td>
</tr>
<tr>
<td>Whole genome next-generation sequences (NGS) of 161 rams, 56 ewes and 25 wild sheep analyzed for whole genome analysis</td>
<td>This paper and [27]</td>
<td>GenBank/NCBI PRJNA624020 and PRJNA645671</td>
</tr>
<tr>
<td><strong>Oligonucleotides</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single-copy MSY contigs</td>
<td>This paper</td>
<td>GenBank/NCBI MT768242–MT768708</td>
</tr>
<tr>
<td>Single-copy MSY SNPs</td>
<td>This paper</td>
<td>EVA/NCBI/dbSNP PRJEB39831</td>
</tr>
<tr>
<td>Mitogenome assembly</td>
<td>This paper</td>
<td>GenBank/NCBI MT768063–MT768241</td>
</tr>
<tr>
<td>mtDNA D-loop</td>
<td>This paper</td>
<td>GenBank/NCBI MT768709–MT769136 and MT768063–MT768241</td>
</tr>
<tr>
<td><strong>SOFTWARE AND ALGORITHMS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trimmomatic v0.38</td>
<td>[66]</td>
<td><a href="http://www.usadellab.org/cms/?page=trimmomatic">http://www.usadellab.org/cms/?page=trimmomatic</a></td>
</tr>
<tr>
<td>BWA v0.7.8</td>
<td>[67]</td>
<td><a href="http://bio-bwa.sourceforge.net/">http://bio-bwa.sourceforge.net/</a></td>
</tr>
<tr>
<td>SAMtools v0.1.19</td>
<td>[68]</td>
<td><a href="https://sourceforge.net/projects/samtools/files/samtools/0.1.19/">https://sourceforge.net/projects/samtools/files/samtools/0.1.19/</a></td>
</tr>
<tr>
<td>BEDTools v2.17.0</td>
<td>[69]</td>
<td><a href="https://sourceforge.net/projects/bedtools/">https://sourceforge.net/projects/bedtools/</a></td>
</tr>
<tr>
<td>Integrative Genomics Viewer (IGV) v2.4.5</td>
<td>[70]</td>
<td><a href="http://software.broadinstitute.org/software/igv/">http://software.broadinstitute.org/software/igv/</a></td>
</tr>
<tr>
<td>R v3.5.1</td>
<td>N/A</td>
<td><a href="https://www.R-project.org/">https://www.R-project.org/</a></td>
</tr>
<tr>
<td>Genome Analysis Toolkit (GATK) v4.0</td>
<td>[71]</td>
<td><a href="https://software.broadinstitute.org/gatk/download/">https://software.broadinstitute.org/gatk/download/</a></td>
</tr>
<tr>
<td>VCFtools v0.1.13</td>
<td>[72]</td>
<td><a href="http://vcftools.github.io/">http://vcftools.github.io/</a></td>
</tr>
<tr>
<td>MITObim v1.9.1</td>
<td>[73]</td>
<td><a href="https://github.com/chrishah/MITObim">https://github.com/chrishah/MITObim</a></td>
</tr>
<tr>
<td>MIRA v4.0.2</td>
<td>[73]</td>
<td><a href="https://sourceforge.net/projects/mira-assembler/">https://sourceforge.net/projects/mira-assembler/</a></td>
</tr>
<tr>
<td>Picard v2.18.12</td>
<td>N/A</td>
<td><a href="http://broadinstitute.github.io/picard/">http://broadinstitute.github.io/picard/</a></td>
</tr>
<tr>
<td>MEGA 7.0</td>
<td>[74]</td>
<td><a href="https://www.megasoftware.net/">https://www.megasoftware.net/</a></td>
</tr>
<tr>
<td>jModelTest v.2.1.4</td>
<td>[75]</td>
<td><a href="https://code.google.com/p/jmodeltest2">https://code.google.com/p/jmodeltest2</a></td>
</tr>
</tbody>
</table>

(Continued on next page)
RESOURCE AVAILABILITY

Lead Contact
Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Meng-Hua Li (menghua.li@cau.edu.cn).

Materials Availability
This study did not generate new unique reagents.

Data and Code Availability
The accession number for the whole genome sequencing reads of domestic and wild sheep reported in this paper is NCBI Short Reads Archive: PRJNA645671. The single-copy MSY contigs, complete mtDNA, and D-loop sequences have been deposited at GenBank under the accessions MT768242–MT768708, MT768063–MT768241, and MT768242–MT769136, respectively. SNPs identified in single-copy MSY have been submitted to NCBI/EVA/dbSNP: PRJEB39831.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

For NGS sequencing aliquots of blood samples from 164 healthy sheep (136 adult rams and 28 adult ewes) were collected as part of routine diagnostics at the sampling sites. All animal work was conducted according to a permit (no. IOZ13015) approved by the Committee for Animal Experiments of the Institute of Zoology, Chinese Academy of Sciences (CAS), China. For domestic sheep samples, animal sampling was also approved by local authorities from where the samples were taken. For the eight argali samples, sample collection was conducted under Law of the People’s Republic of China on the Protection of Wildlife and relevant regulations. For urial, we collected peripheral blood sample from one captive urial after receiving authorization for research from the Department of Environmental Protection in Iran (no. 93/34089).

METHOD DETAILS

Archeological Records and Historical Data
To investigate the distribution of domestic sheep in early historical times, we compiled archeological records and historical data of domestic sheep across the world from the published literature, including the information of archeological sites, time, ancient remains and historical records associated with sheep, in particular the wooly and fat-tailed sheep (Data S1). Specifically, we performed an extensive literature review by searching for relevant publications in “National Center for Biotechnology Information,” “Web of Science” and “China National Knowledge Infrastructure” original briefings that included the keywords (i) archeological, sheep; (ii) archeological, history, wool, sheep; and (iii) archeological, history, fat tail, sheep. These queries were deemed sufficient to find the majority of relevant archeological and historical data about sheep in the world. We adopted the upper limit of the estimated
time period for archaeological sites and historical records in the plots using the ArcMap program implemented in the ArcGIS v10.0 software (ESRI, Redlands, California, USA).

**Samples and Sequencing**

Blood or tissue samples were collected from a total of 598 individuals comprising 561 rams and 28 ewes of domestic sheep (O. aries), and 9 rams of wild sheep (8 argali (O. ammon) and one urial (O. vignei)). The domestic sheep represent 96 native populations (498 rams) and 22 improved populations (63 rams) from different geographic origins in Asia, Europe, Africa, America, and the Middle East (Data S2). The vast majority of the samples are from local populations that have been native to specific geographic areas for several hundred years and have not been admixed by commercial flocks. Genomic DNA was extracted following the standard phenol-chloroform extraction procedure or the QIAGEN’s DNeasy Blood and Tissue Kit.

For whole genome sequencing, at least 0.5 μg of genomic DNA from 164 unrelated animals representing 79 populations of domestic sheep (127 rams and 28 ewes) and 2 wild species (8 argali and one urial) (NCBI: PRJNA645671; Data S2) was used to construct a library with an insert size of approximately 350 bp. Paired-end sequencing libraries were constructed according to the manufacturer’s instructions (Illumina Inc., San Diego, CA, USA), and sequenced on the Illumina HiSeq X Ten sequencer (Illumina Inc.). Raw paired-end reads with a depth of 12.45–23.20 × (average depth = 18.60 ×) were sequenced (Data S2). Published whole-genome sequence data include 16 Asiatic mouflons (9 rams and 7 ewes) and 52 domestic sheep (34 rams and 18 ewes) representing 12 populations (dataset number PRJNA624020; Data S2).

Adaptors of all raw data were removed, and quality-based trimming was performed with Trimmomatic v0.38 [66] using settings “LEADING:5 || TRAILING:5 || SLIDINGWINDOW:4:15 || MINLEN:40.”

**Identification of MSY SNPs**

**MSY de novo assembly**

Reads generated through the whole-genome data of three rams from three different populations from different geographic origins (Afghanistan Wagir sample 2N of Wagir sheep, Xinjiang Duolang DLS309 sample, and Sweden sample Gotland24514; NCBI: PRJNA624020; Data S2) were mapped to the O. aries genome (OARv4.0, NCBI: GCA_000298735.2) [65] using BWA v0.7.8 [67]. The unmapped reads were then extracted by SAMtools v0.1.19 [68], CLC Genomics Workbench 11 (https://digitalinsights.qiagen.com/products-overview/analysis-and-visualization/qiagen-clc-genomics-workbench/) was used to assemble MSY raw contigs based on the unmapped reads. After filtering the short pieces of contigs with length less than 399 bp, we obtained a total of 12,303 raw contigs (N50 = 818 bp; Figure 2A).

**Identification of single-copy MSY contigs**

Whole genome next-generation sequences (NGS) of 31 rams and 5 ewes (NCBI: PRJNA624020; Data S2) were mapped to the MSY_raw_contigs using BWA v0.7.8 [67]. SAMtools v0.1.19 [68] was then used to remove PCR duplicates and filter for mapping quality. The coverage percentage (covperc) per contig was calculated with BEDTools v2.17.0 [69] for each sample. Contigs that met the following criteria were retained: covperc > 0.66 in males and < 0.25 in females.

A total of 92 randomly chosen contigs were then validated by PCR using DNA from rams and as a template (Figure 2B). All contigs were checked with the Integrative Genomics Viewer (IGV) v2.4.5 [70]. Afterward, 600 contigs coded by MSY_linked were retained.

To identify single-copy MSY (scpMSY) contigs, we estimated the coverage of 5 available Y chromosome genes in sheep (DBY, SRY, UTY, AMELY, and ZFY) [24]. The per-base coverage was computed using genomeCoverageBed in BEDTools v2.17.0 [69]. Average coverage (averagecov) for the five genes and MSY_linked in each sample genome was calculated using R v3.5.1 (https://www.R-project.org/). We obtain the average coverage range of 9.5–15.4 × for the four single-copy genes (DBY, SRY, UTY, and AMELY) (Figure 2C). We then normalized the averagecov of MSY contigs for each sample based on the coverage range of the 4 single-copy genes. Contigs with a mean normalized average coverage (MNAC) less than 1.5 across all the ram samples were retained, and 333 multiple-copy MSY (mcpMSY) contigs were filtered (Figure 2D). Finally, we obtained 467 single-copy MSY (scpMSY) contigs (GenBank: MT768242–MT768708). The scpMSY contigs of each sample showed that the averagecov (~17.69) was near half of the coverage of whole genome sequencing (~27.50) for the 31 rams (Figure 2E, Data S2).

**ScpMSY SNPs calling and filtering**

From a set of 211 samples, NGS reads were mapped to the MSY-linked contigs and were then realigned using the Genome Analysis Toolkit (GATK) v4.0 [71]. Variants were called in each sample separately with GATK’s HaplotypeCaller using the settings “–genotyping-mode DISCOVERY|–output-mode EMIT_ALL_SITES|–min-base-quality-score 10.” Joint genotyping was performed by merging all the samples using GATK’s GenotypeGVcf. Raw variants were filtered with VCFTools v0.1.13 [72] using the settings “–minDP 2|–minQ 20|–remove-indels”. Only the SNPs called in scpMSY contigs meeting the following criteria were retained: (i) only present in at least two males and not in females; (ii) hemizygous; (iii) absence of additional SNP within 100 bp upstream and downstream of the called SNP; and (iv) passing the visual inspection by IGV v2.4.5 [70]. After filtering, a novel set of 493 scpMSY SNPs including 85 SNPs in domestic sheep and 425 in the wild species were obtained (Data S3). Further, all the 85 scpMSY SNPs in domestic sheep and 18 randomly selected scpMSY SNPs in wild species were validated by independent PCR in three rams and three ewes.

**Assembly of Mitogenomes**

The clean paired-end reads were merged as one interleaved fastq file for the 161 rams of domestic sheep and 18 rams of wild sheep. The program MITObim v1.9.1 (stable-relies on MIRA v4.0.2) [73] was used to assemble the whole mitogenomes based on the...
Whole-genome SNP Calling

Clean paired-end reads of the 203 samples (179 rams and 24 ewes) (Data S2) were mapped to the sheep reference genome (Oar v4.0, NCBI: GCA_000298735.2) using the MEM module in the program BWA v0.7.8 [67] with the parameters “bwa -k 32 -M -R.” The SAMtools v0.1.19 [68] software was used to convert mapping results into the BAM format and filter the unmapped and non-unique reads. We used Picard v2.18.12 (http://broadinstitute.github.io/picard/) SortSam to sort the resulting bam files according to the order of chromosome coordinates, and the program Picard MarkDuplicates was used to remove duplicate reads. After the BWA alignment, a local realignment around indels was performed to correct misalignments due to the presence of indels with GATK v4.0 [71] in two steps. First, the RealignmentTargetCreator command was used to determine the regions affected by local realignment. Second, IndelRealigner was used to realign the regions found above, which produced a realigned BAM file for each sample. To obtain accurate bases, Base Quality Score Recalibration (BQSR) was applied to detect systematic sequencing errors by using GATK modules, BaseRecalibrator and ApplyBQSR.

The SNP calling followed the best practice workflow recommended by GATK v4.0 [71]. In brief, SNPs were called for each sample using the GATK HaplotypeCaller module, and then a joint genotyping step for the integrated variations union which contains all samples was performed on a combined gVCF file. Filtering parameters were set as QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 || HaplotypeScore > 13.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || QUAL < 30. SNPs with non-biallelic and > 40% missing calls were removed, which yielded a dataset of 31,343,261 SNPs for further analysis.

MTDNA D-Loop Sequencing

We sequenced the mtDNA D-loop region (749 bp) of all males using a pair of primers, MSD-F 5'-ACAACACGGACTTTCACCTC-3' (map positions 15,522–15,541 in GenBank: AF010406) and MSD-R 5'-CCAGCTCCCCACCCAAAATTA-3' (map positions 16,299–16,318 in AF010406) [16]. Each PCR mixture (30 µl total volume) contained 15 µl 10 x Taq Master Mix (ComWin Biotech, Beijing, China), 1 µl DNA (30 ng/µl), 1 µl of each primer (10 pmol/µl), and 12 µl ddH2O. PCR cycling conditions were 95°C for 15 min, 35 cycles at 95°C for 30 s, 60°C for 90 s, and 72°C for 30 s, and a final extension at 72°C for 7 min. PCR products were sequenced directly in both directions using the PCR primers on an ABI 3730 capillary sequencer (Applied Biosystems, Life Technologies, NY, USA). Following quality trimming and filtering, common fragments of 285 bp (15,541–15,652 and 15,960–16,132 of AF010406) of the D-loop were aligned and edited for analysis using MEGA 7.0 [74].

Phylogenetic Analyses

MSY and mitogenome haplotypes

In the 179 rams of domestic and wild sheep, 495 scpMSY SNPs (493 novel plus the two published SNPs, Data S3) were used for haplotype reconstruction. Overall, sixty-three MSY haplotypes were defined including 49 for domestic sheep (ds-H1–ds-H49), one for urial (u-H1), six for Asiatic mouflons (am-H1–am-H6), and seven for argali (a-H1–a-H7) (Data S4). In addition, 171 mtDNA haplotypes (mt-H1–mt-H171) were defined based on the 179 mitogenomes (GenBank: MT768063–MT768241).

Phylogenetics and scpMSY haplogroups

Hierarchical likelihood ratio tests were implemented to select a best-fit model of nucleotide substitution for phylogenetic analysis using jModelTest v2.1.4 [75]. Out of 88 candidate models, the best-fit model HKY+I was chosen for the scpMSY dataset, while HKY+I+G was selected for the mitogenomes. Phylogenetic trees were inferred using the maximum likelihood (ML), maximum parsimony (MP) and neighbor-joining (NJ) methods with the HKY model in the program MEGA v 7.0 [74] and Bayesian inference (BI) in the program BEAST v2.5.1 [34]. To compare the paternal and maternal phylogenetic relationships among the rams of domestic sheep, we constructed NJ trees based on the 81 scpMSY SNPs (Data S4) and mtDNA D-loop sequences (GenBank: MT768709–MT769136 and MT768063–MT768241) using MEGA v.7.0 [74]. The bootstrap consensus trees were inferred from 100 replications. Neighbor-joining trees based on data from the 179 rams as well as from the 49 sheep (25 rams and 24 ewes) involved in the ABC analysis were constructed based on whole-genome sequences. Single nucleotide polymorphisms with a minor allele frequency (MAF) below 0.05 and a missing call rate above 0.1 were removed (-maf 0.05 and -geno 0.1) using PLINK v 1.90 [76] from the two datasets. We then performed LD pruning by using PLINK with the commands “--indep-pairwise (50 5 0.5)” for the dataset of 179
rams and “–indep-pairwise (50 5 0.2)” for the dataset of 49 sheep. Finally, 4,262,880 SNPs and 1,689,814 SNPs segregating in 179 rams and 49 sheep, respectively, were retained for the phylogenetic reconstruction. The NJ trees were built using TreeBeST v.1.9.2 [77] with 100 bootstraps. FigTree v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize the NJ trees.

**Empirical Bayesian estimation of \( T_{MRCA} \) and \( N_e \)**

An empirical Bayesian approach that uses a prior distribution of \( T_{MRCA} \) (time to the most recent common ancestor), based on the coalescent theory, and conducts Markov chain simulations to estimate the likelihood of parameters was applied to estimate divergence times of MSY lineages. Phylogenetic reconstruction and molecular dating of lineage splitting, including estimation of \( T_{MRCA} \) for all the ovine lineages were implemented using BEAST v.2.5.1 [34]. Due to the absence of fossil records of species from the *Ovis* genus that could be used for time calibration, we used the divergence time of 2.93 million years ago (mya) between argali and domestic sheep and the time of 2.60 mya between urial and domestic sheep as a *priori* parameter, both of which were obtained from a comprehensive evolutionary analysis [16]. A strict clock was set to calculate the mutation rate. We set the Bayesian Markov chain Monte Carlo (MCMC) length to 20,000,000 steps and logging parameters every 1,000 steps. A Maximum Clade Credibility (MCC) tree was generated using a 10% burn-in with BEAST’s TreeAnnotator and drawn with FigTree v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/). Convergence was confirmed by effective sampling size (ESS) > 200 using the program Tracer v.1.5 (http://beast.bio.ed.ac.uk/Tracer). We simulated the posterior distribution of \( T_{MRCA} \), conditional on the two estimates of divergence time.

We performed a grid search to obtain the Watterson’s \( \theta \) (\( \theta_W \)) estimator for a scaled mutation rate, \( \theta_W = 2N_e\mu_{bg} \), where \( \mu_{bg} \) is the mutation rate per site per year [85]. R v3.2.1 (https://www.R-project.org/) was used to estimate \( \theta_W \),

\[
\theta_W = S/(L + \text{sum}(1/(1:(n-1))))
\]

where \( S \) is the expected number of segregating sites, \( L \) is the total number of sites, and \( n \) is the number of samples. We combined all the scpMSY contigs having the called SNPs and obtained a total length of 344,382 bp. Using the above equations, we calculated the effective population size (\( N_e \)) for the lineages of scpMSY and complete mitogenomes, separately. Also, the Bayesian skyline plot (BSP) curves for the lineages of scpMSY contigs and complete mitogenomes were drawn to reconstruct the paternal and maternal historical demographics, respectively.

**Geographic Patterns of the MSY Variability**

First, frequencies of the four MSY lineages in the 595 rams from 118 populations were visualized using the ArcMap program implemented in the ArcGIS v10.0 software (ESRI, Redlands, California, USA) (Data S5). A total of 61 MSY haplotypes (s-H1–s-H61) assigned to the four lineages were defined by the 81 scpMSY SNPs (Data S4). The MSY SNP oY1 alleles were collected in 3,056 modern samples from 214 domestic populations and 92 wild sheep from 7 populations (4 populations of European mouflons, 2 populations of argali and 1 population of urial) in previous studies (Data S6).

**Approximate Bayesian Computation Analysis**

**Data preparation and processing**

To gain insights into the demographic history of domestic sheep with distinctive phenotypic traits (“primitive” and fat-tailed populations), we used the approximate Bayesian computation (ABC) approach to test two alternative scenarios based on the whole genome sequences. In each scenario, we considered the Asiatic moufflon (*Ovis orientalis*) as the wild ancestor of domestic sheep based on archeological and genetic evidence [11] and only native populations were used in this analysis in order to minimize the impact of intense and recent artificial selection. We selected individuals representative from the “primitive” short-tailed Northern European populations and from the Near Eastern fat-tailed populations. In total, 49 individuals (16 “primitive”, 18 fat-tailed, and 15 Asiatic moufflons) were included in the ABC analysis (Figure S2A, Data S2).

To match the requirements of simulations, SNPs with a minor allele frequency (MAF) lower than 0.05 and with a missing call rate higher than 0.1 were excluded (-maf 0.05 and -geno 0.1 in PLINK). To mitigate the effect of LD, we implemented LD pruning using the command “–indep-pairwise (50 5 0.5)” in the program PLINK v.1.90 [76]. Then, SNPs located 150-kb away from genes and without missing genotypes were retained, establishing an average distance of 150-kb between SNPs to minimize the effects of linkage. Since the coalescent simulations underlying ABC inference assume neutrality, we used the Bayesian approach implemented in BAYESCAN v.2.1 [78] to detect outlier SNPs with default arguments. The SNPs with a q-value lower than 0.05 as well as those mapping to the X chromosome or to the mitochondrial genome, which would exhibit reduced \( N_e \) as compared to the autosomal regions, were eliminated. After filtering, 4,438 unlinked and neutral SNPs were kept in the final dataset for the modeling analyses.

**ABC inferences**

Backward coalescent simulations with recombination were performed using FASTSIMCOAL v.2.5.2.21 [79, 80] under 2 models (Figure S2D). In **Model-1**, the fat-tailed sheep (\( N_{FT} \)) first diverged from the wild ancestor of domestic sheep (\( N_{AM} \)) at time \( T_2 \), then the “primitive” sheep (\( N_{pr} \)) split from the fat-tailed sheep (\( N_{FT} \)) at time \( T_1 \). In **Model-2**, the “primitive” sheep (\( N_{pr} \)) first diverged from the wild ancestor of domestic sheep (\( N_{AM} \)) at time \( T_1 \), then the fat-tailed sheep (\( N_{FT} \)) split from the “primitive” sheep (\( N_{pr} \)) at time \( T_2 \). Black arrows represent migration rates that are simulated as independent, continuous parameters.

Model parameter prior distributions are shown in Figure S2F, which were established based on previous study [13, 80]. For testing the model, we ran \( 5 \times 10^5 \) simulations per model. A set of 24 summary statistics were computed by the program ARSLSUMSTAT v.3.5.2.2 (Figure S2B) [81] for each simulation and were used to describe genetic variations within (e.g., \( K \), mean number of alleles; \( H \), mean heterozygosity) and between (e.g., \( F_{ST} \)) sheep populations with different phenotypic traits (Data S2). Because correlated
summary statistics can make the model simulation redundant and even produce incorrect results [86, 87]. We calculated the correlations between each pair of the summary statistics using the Spearman’s rho statistic by the corr.test function implemented in the R package psych [88]. The graphical representation of correlation coefficients (Spearman’s rho) among the summary statistics was then obtained using a modified script of the corrplot function from the R package corrplot (Figure S2B) [82]. Ultimately, 7 informative and independent summary statistics were selected to compare the candidate models (Figures S2B and S2E).

We performed a cross-validation procedure to assess if the 7 summary statistics can provide enough statistical power to distinguish the two scenarios using the function cv4postpr in the R package abc [83]. We used 5,000 (1%) simulations closest to the observed data and randomly chose 1,000 sets of summary statistics for each model, using a single tolerance rate (tols) of 0.05 and the method ‘mnlogistic’ based on multinomial logistic regression. The results showed that the two models could be clearly classified in the corresponding simulations (Figure S2C).

In order to identify the best-supported model, we compared all models simultaneously using a standard ABC-GLM approach as implemented in ABCtoolbox [48] with the 5,000 (1%) simulations closest to the observed data for each model. The marginal distributions of each model were used to calculate the posterior probability of the model (PP) by using the function postpr in the R package abc with the same arguments as function cv4postpr, which corresponds to the proportion of the retained simulations that presented a smaller or equal likelihood under the estimated GLM as compared to the observed data [89]. We also used the marginal densities to calculate Bayes factors (BF) for each pairwise comparison between the two candidate models [48, 90–92]. The highest support was estimated for Model-2 using both the posterior probabilities and the Bayes Factors (Figure S2D). Furthermore, we calculated the probability (p value) of the observed data under the general linear model (GLM) used for the post-sampling regression with ABCtoolbox [48]. A small probability indicates that the inferred GLM does not fit the observed data well. We obtained a p value of 1, showing that Model-2 can reproduce the observed data well (Figure S2D) [93, 94]. As evidence, all the 7 selected observed summary statistics occurred within the 95% (i.e., 2.5%–97.5%) percentiles of the simulated summary statistics (Figure S2E).

For the demographic parameters inference, we re-ran 2 x 10^6 simulations under the best-supported model (Model-2). We retained 10,000 (0.5%) simulations that were closest to the observed data under the best-supported model (Model-2) (Figures 4D and S2D) following the methods described previously [48]. We then utilized the retained simulation data to calculate the median, mode, 50% and 95% highest posterior density (HPD) intervals for each demographic parameter under Model-2 (Figure S2F). Furthermore, we checked for the biased posterior distributions based on 1,000 pseudo-observed datasets randomly chosen from the simulated data. We then computed the coverage properties of the posterior distribution using our 10,000 closest simulations. Uniformity was assessed using a classical Kolmogorov-Smirnov test using the function ks.test in the R package for each parameter independently (Figure S2F) [48, 87, 95]. Deviations from uniformity indicate biased posterior distributions and the corresponding parameter estimates should be considered with caution.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

The Pearson’s Chi-square test was used to assess the association between y-HB and the fat-tail phenotype. \( \chi^2 (df = 1, p = 0.01) = 6.63; \)
\( \chi^2 (df = 1, p = 0.05) = 3.84. \) In the ABC analysis, Bayes factors (BF) were calculated to indicate the relative strength of evidence for two candidate models based on their own marginal densities: 1 < BF \( \leq 3 \) (non-significant evidence); 3 < BF \( \leq 10 \) (substantial evidence); BF > 10 (strong evidence). We used ABCtoolbox to compute a probability (p value) to compare the marginal density of the observed data with marginal densities obtained from the retained simulations, which measured the ability of the model to reproduce the observed data. \( ^* p \leq 0.01; \) \( ^* p \leq 0.05. \) Kolmogorov-Smirnov test was used to compare the posterior distributions of parameters of the best-supported model against a uniform distribution. \( ^* p \leq 0.01; \) \( ^* p \leq 0.05. \)
Supplemental Information

Paternal Origins and Migratory Episodes of Domestic Sheep

Juan Deng, Xing-Long Xie, Dong-Feng Wang, Chao Zhao, Feng-Hua Lv, Xin Li, Ji Yang, Jia-Lin Yu, Min Shen, Lei Gao, Jing-Quan Yang, Ming-Jun Liu, Wen-Rong Li, Yu-Tao Wang, Feng Wang, Jin-Quan Li, EEr Hehua, Yong-Gang Liu, Zhi-Qiang Shen, Yan-Ling Ren, Guang-Jian Liu, Ze-Hui Chen, Neena A. Gorkhali, Hossam E. Rushdi, Hosein Salehian-Dehkordi, Ali Esmailizadeh, Maryam Nosrati, Samuel R. Paiva, Alexandre R. Caetano, Ondrej Stěpánek, Ingrid Olsaker, Christina Weimann, Georg Erhardt, Ino Curik, Juha Kantanen, Joram M. Mwacharo, Olivier Hanotte, Michael W. Bruford, Elena Ciani, Kathiravan Periasamy, Marcel Amills, Johannes A. Lenstra, Jian-Lin Han, Hong-Ping Zhang, Li Li, and Meng-Hua Li
Figure S1. Phylogenetic trees based on MSY, mitogenome and whole-genome data. Related to Figure 3.

(A) Phylogenetic trees using Maximum likelihood (ML), maximum parsimony (MP), neighbor joining (NJ) with bootstrap values (1,000 replicates) and Bayesian inference (BI) with posterior probabilities indicated near the nodes based on the 49 MSY haplotypes, related to (see Data S4).

(B) The phylogenetic relationship of Y-chromosomal lineages based on 495 scpMSY SNPs from 179 rams (see Data S4).

(C) The phylogenetic relationship of mtDNA lineages based on 179 mitogenome sequences (GenBank MT768063–MT768241).

(D) Neighbor-joining tree with 100 bootstrap replicates based on the whole-genome sequences, related to Figure 3C.

(E) The neighbor-joining tree based on the 81 MSY SNPs from a total of 595 rams (see Data S4).

(F) The neighbor-joining tree based on mtDNA D-loop sequences from a total of 589 rams (GenBank MT768709–MT769136 and MT768063–MT768241). Divergence times [million(s) years ago, myr] for the lineages were estimated on the basis of data from the 179 individuals.
**Figure S2. Summary information of approximate Bayesian computation (ABC) analysis for demographic inference. Related Figure 4D.**

(A) The neighbor-joining tree of 34 domestic sheep and 15 Asiatic Mouflon constructed based on 1.69 million SNPs. The native sheep with different phenotypic traits and Asiatic Mouflon is highlighted using different colors. The corresponding MSY haplotypes of rams are displayed next to their samples. The individuals in gray are females.

(B) Spearman correlation plots of 24 simulated summary statistics used in the program ABCtoolbox. Ellipse glyphs are shaped to match the corresponding Spearman’s coefficient. Italic and red characters represent 7 selected summary statistics, which are utilized for the model choice and parameter inference. Population specific summary statistics: the mean number of alleles over loci for single genetic group (K), the standard deviation over loci of the number of allele for single genetic group (Ksd), mean number of alleles over loci and genetic groups (mean_K), standard deviation over genetic groups of the mean number of alleles (sd_K), mean total number of alleles over loci (tot_K), mean heterozygosity over loci for single genetic group (H), standard deviation over loci of the heterozygosity for single genetic group (Hsd), mean heterozygosity over loci and genetic groups (mean_H), standard deviation of the mean heterozygosity over genetic groups (sd_H), mean total heterozygosity (tot_H). Population pairwise comparisons: pairwise \( F_{ST} \) between genetic groups (\( F_{ST} \)), mean number of differences between genetic groups (PI). PR, “Primitive” Sheep; AM, Asiatic Mouflon; FT, Fat-Tailed Sheep.

(C) Misclassification of the two demographic models using the R package abc. The confusion matrix is based on 1,000 samples (nval) of each model. Each grey shade from dark to light corresponds to the models 1–2.

(D) The model-testing approach compared two candidate models. In this comparison, **Model-2** is the best-fitting model. The corresponding posterior probability (PP), Bayes factor (BF) and \( p \)-value are shown above the two models, all parameters are reported in Figure S2F.

(E) Summary statistics for the 7 selected observed dataset with corresponding 2.5 and 97.5 percentiles. Distributions are generated from the 5,000 (1%) simulations closest to the observed dataset of **Model-2**.

(F) Parameter estimation of the best supported model (**Model-2**) in the approximate Bayesian computation (ABC) analysis. GAMMA, gamma distribution of the mutation rate; MUTATION, rate per site per generation, with a generation time assumed to be 3 years; \( N_{AM} \), the estimated effective population size of Asiatic Mouflon, in haploids; \( N_{PR} \), the estimated effective population size of “Primitive” sheep, in haploids; \( N_{FT} \), the estimated effective population size of Fat-Tailed Sheep, in haploids; \( m_1 \), migration rate from “Primitive” sheep (\( N_{PR} \)) to Fat-Tailed sheep (\( N_{FT} \)); \( m_2 \), migration rate from Fat-Tailed Sheep (\( N_{FT} \)) to “Primitive” Sheep (\( N_{PR} \)); \( T_1 \), the split time of “Primitive” sheep, in years; \( T_2 \), the split time of Fat-Tailed sheep, in years; \( P_{value_{KS}} \), the \( P \) value of the Kolomogorov-Smirnov test of uniformity of the posterior quantiles; 50%-HPD, 50%-highest posterior density interval; 95%-HPD, 95%-highest posterior density interval. All other models use the same prior bound.
Supplemental References


bones. PloS one 12, e0178543.


