The origin of domestication genes in goats

Zheng, Zhuqing; Wang, Xihong; Li, Ming; Li, Yunjia; Yang, Zhirui; Wang, Xiaolong; Pan, Xiangyu; Gong, Mian; Zhang, Yu; Guo, Yingwei; Wang, Yu; Liu, Jing; Cai, Yudong; Chen, Qiuming; Okpeku, Moses; Colli, Licia; Cai, Dawei; Wang, Kun; Huang, Shisheng; Sonstegard, Tad S.; Esmailizadeh, Ali; Zhang, Wenguang; Zhang, Tingting; Xu, Yangbin; Xu, Naiyi; Yang, Yi; Han, Jianlin; Chen, Lei; Lesur, Joséphine; Daly, Kevin G.; Bradley, Daniel G.; Heller, Rasmus; Zhang, Guojie; Wang, Wen; Chen, Yulin; Jiang, Yu

Published in:
Science Advances

DOI:
10.1126/sciadv.aaz5216

Publication date:
2020

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC

Citation for published version (APA):
**GENETICS**

The origin of genomic genes in goats

Zhuing Zheng1*, Xihong Wang1+, Ming Li1*, Yunia Li1*, Zhirui Yang1*, Xiaolong Wang1*, Xiagyu Pan1, Mian Gong1, Yu Zhang1, Yingwei Guo1, Yu Wang1, Jing Liu1, Yudong Cai1, Qiuming Chen1, Moses Adipeko1,3, Licia Colli1,4, Dawei Cai6, Kun Wang7, Shisheng Huang1, Tad S. Sonstegard9, Ali Esmalizadeh9, Wenguang Zhang8, Tingting Zhang8, Yangbin Xu9, Najy Xu1, Yi Yang1, Jianlin Han12,13, Lei Chen12, Josephine Lesur14, Kevin G. Daly15, Daniel G. Bradley15, Rasmus Heller16, Guojie Zhang17,18,19, Wen Wang17,19, Yulin Chen1†, Yu Jiang1†

Goat domestication was critical for agriculture and civilization, but its underlying genetic changes and selection regimes remain unclear. Here, we analyze the genomes of worldwide domestic goats, wild caprid species, and historical remains, providing evidence of an ancient introgression event from a West Caucasian tur-like species to the ancestor of domestic goats. One introgressed locus with a strong signature of selection harbors the MUC6 gene, which encodes a gastrointestinal secreting muccin. Experiments revealed that the nearly fixed introgressed haplotype confers enhanced immune resistance to gastrointestinal pathogens. Another locus with a strong signal of selection may be related to behavior. The selected alleles at these two loci emerged in domestic goats at least 7200 and 8100 years ago, respectively, and increased to high frequencies concurrent with the expansion of the ubiquitous modern mitochondrial haplogroup A. Tracking these archaeologically cryptic evolutionary transformations provides new insights into the mechanisms of animal domestication.

INTRODUCTION

Domestication presents an extreme shift of physiological and behavioral stress for free-living animals (1) to a high-density and disease-prone anthropogenic environment, especially for herbivores. The goat (Capra hircus) was one of the first domesticated livestock species, demonstrating remarkable adaptability and versatility (2, 3), and it has been intimately associated with human dispersal (4). Recent studies have identified candidate targets of selection during goat domestication, including loci related to pigmentation, xenobiotic metabolism, and milk production (5, 6); however, the evolutionary dynamics of key genes involved in adaptation during the early phase of domestication remain unclear.

Goat domestication is believed to have begun ~11,000 years ago from a mosaic of wild bezoar populations (Capra aegagrus) (2, 6). However, other wild Capra species are widely distributed and many of them can hybridize with domestic goats (7–9). Their contribution to the goat domestication process remains unexplored. Introgres- sion of adaptive variants has been widely recognized as a notable evolutionary phenomenon in humans (10), sheep (11), and cattle (12), and it may be particularly effective for increasing fitness without negative pleiotropic effects as demonstrated in other species (13). Here, we conducted a comprehensive population genomic survey of modern goat populations distributed across the world, six wild Capra species, and previously published ancient goat genomes to investigate temporal shifts in key genetic variants under selection during goat domestication.

RESULTS

Population structure and origin of domestic goats

We sequenced 101 Capra genomes (coverage, 3 to 47 times; average, 12 times), including 88 domestic goats collected from three different continents, one bezoar, one Alpine ibex (Capra ibex), three Siberian ibex (Capra sibirica), three Markhors (Capra falconeri), one West Caucasian tur (Capra caucasica), and four Nubian ibex × domestic goat hybrids (Capra nubiana × C. hircus) (Fig. 1A and text S1). We also sequenced five ancient goat samples at ~0.04 to 13.44 times coverage (Fig. 1A and text S1), including the earliest known Chinese archaeological samples from the Neolithic time period (14). Together with the publicly available genomic sequences for modern goat and bezoar (table S1) (5, 15), as well as ancient genomes from (table S4) (6), we compiled a worldwide collection of 164 modern domestic goats, 52 ancient goats, 24 modern bezoars, and 4 ancient bezoars.

A whole-genome neighbor-joining phylogenetic tree reveals that all domestic goats form a monophyletic sister lineage to bezoars (Fig. 1B and fig. S4), confirming that modern domestic goats are the
descendants of bezoar-like ancestors (16). The other four wild Capra species (C. ibex, C. sibirica, C. falconeri, and C. caucasica), which are referred to as the ibex-like species (17), fall exclusively in another branch divergent from bezoar-goat (Fig. 1B and fig. S4). Principal components analysis (PCA) shows that the three bezoar populations, which were collected in the Middle East (Fig. 1A) near the domestication center (2, 18), are structured and cluster corresponding to their geographic origins (Fig. 1C). Within domestic goats, PCA and model-based clustering (k = 3) show that Asian goats are genetically distinct from European (EUR) and African (AFR) samples (Fig. 1, D and E). At k = 6, Asian goats further split into two geographic subgroups: Southwest Asia–South Asia (SWA-SAS) and East Asia (EAS) (Fig. 1E). This geographic structure is also supported by TreeMix (fig. S8) and haplotype-based statistics (fig. S9), in agreement with the scenario that the ancestors of present-day domestic goat populations followed distinct dispersal routes along the east-west axis of Afro-Eurasia (Fig. 2A and table S6) (4). This dispersal scenario was further supported by the observation that populations far from the domestication center (non-SWA) had elevated linkage disequilibrium and reduced genetic diversity (fig. S10).

Our demographic analyses using multiple sequentially Markovian coalescent (MSMC), SMC++, and $\partial_\alpha \partial_\iota$ indicate that the divergence times between modern Asian and European goat populations predated the archaeologically estimated domestication time (~11,000 years ago) (fig. S12 and text S2). In addition, D statistics revealed that the bezoars from Zagros display higher genetic affinity to eastern domestic populations, whereas the bezoars in Azerbaijan show higher affinity to western domestic populations (Fig. 2B and fig. S15A). Therefore, the deep coalescence times found between east-west population pairs can be explained by their origin in structured ancestral bezoar
populations (6) or by postdomestication recruitment from different local bezoar populations.

**Gene flow from ibex-like species to predomestication bezoars and modern goats**

*D statistics reveal that all four ibex-like species have significant signals of allele sharing with ancient and modern goats, indicative of admixture (Fig. 2C and table S8). We then examined this genome-wide pattern of admixture between ibex-like species and domestic goats using *D* statistics and identity by state in 20-kb sliding windows. We further verified candidate introgressed regions using Sprime and maximum likelihood (ML) phylogenetic trees. Using a conservative criterion (namely, only keeping putative introgressed haplotypes with a frequency higher than 0.1 in goats), we identified 112 genomic...
segments overlapping with 81 protein-coding genes with signatures of introgression from ibex-like species (Fig. 2D, fig. S16, and data file S1). A Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis for these genes shows that the most significantly enriched category is amoebiasis (hypergeometric test, adjusted $P < 5.28 \times 10^{-3}$; table S10), which is related to parasite invasion and immunosuppression, including four genes (SERPINB3, SERPINB4, CD1B, and COL4A4). Three additional genes (BPI, MAN2A1, and CD2AP) are also involved in immune function (19–21). In these segments, we observed a pronounced signature of putatively introgressed alleles from the West Caucasian tur (fig. S17), consistent with this species showing the greatest genome-wide allele sharing with domestic goats (Fig. 2C).

While investigating the history of West Caucasian tur admixture, we found that the West Caucasian tur shared more alleles with all four predomesticated bezoars (an Armenian bezoar dating to >47,000 years ago and three Anatolian bezoars dating to ~13,000 years ago) and especially Armenian bezoar, compared to present-day bezoars and domesticates (Fig. 2E). Further $F_\text{ST}$ statistics in the form of $f_2$ (modern bezoar, West Caucasian tur; target) support gene flow from the West Caucasian tur to ancient Armenian bezoar (table S9). Together with the fact that three predomestication Anatolian bezoars had a tur-like mitochondrial haplotype (6), our results provide evidence for a predomestication admixture event of bezoar with a tur-like species. This ancient admixture event is further supported by TreeMix analyses (fig. S8C) and the close geographical proximity between the West Caucasian tur and the ancient Armenian bezoar, with both being, in turn, closer to the goat domestication center than any other living ibex-like species (Fig. 2A). Although all modern goat genomes suggest admixture with tur, Neolithic Balkan and modern EUR and AFR goats show more allele sharing with tur compared to Asian goats (fig. S15B), probably because of their higher genetic affinity with the ancient Armenian and Anatolian bezoars (Fig. 2E) (6). Therefore, predomestication introgression from a tur-like species may have contributed alleles to ancient bezoars and thereby early domesticated goats and their derived modern populations.

Notably, one introgressed region on chromosome 29 has a nearly fixed introgressed haplotype (frequency = 95.7%) in the modern worldwide domestic goats (Fig. 2F and fig. S22A). This region contains one intact protein-coding gene, MUC6, which encodes a gastrointestinal secreted mucin that forms a protective glycoprotein coat involved in host innate immune responses to the invasion of multiple gastrointestinal pathogens (22–24). We searched for potential donor populations in our sequenced wild caprid genomes. The haplotype network constructed with the nonrepeated region of MUC6 shows that the most common domestic haplotype (MUC6$^D$) is similar to West Caucasian tur, differing by only one mutation, while being different from the minor frequency haplotypes of domestic goats and the common haplotype in bezoar (MUC6$^B$) (Fig. 2G). To further validate this introgression signal, we estimated the most recent common ancestor of the divergent haplotypes and investigated whether selection acting on either a de novo mutation or on standing variation could possibly lead to the pattern in this region within domestic goats. Coalescence time estimates (fig. S18A) and neutral simulations (fig. S18B) suggest that the observed pattern of haplotype differentiation is highly unlikely in the absence of interspecific gene flow. Therefore, the nearly fixed MUC6$^D$ in goats was most likely introgressed from a lineage close to the West Caucasian tur, consistent with the genome-wide signal.

### Domestication-mediated selection on immune and neural genes

To identify key selective sweeps in goat domestication, we compared the worldwide domestic goat populations with all 24 bezoars by estimating pairwise genetic differentiation ($F_{\text{ST}}$), differences in nucleotide diversity ($\pi$), and cross-population extended haplotype homozygosity (XP-EHH) in 50-kb sliding windows along the genome (figs. S19 and S20). We defined the windows with significant values ($Z$ test, $P < 0.005$) in all three statistics ($F_{\text{ST}} > 0.195$, $\pi$ ln-ratio $> 0.395$, and XP-EHH $> 2.10$) as putative selective sweeps. After merging consecutive outlier windows, 105 unique regions containing 403 protein-coding genes were identified (Fig. 3A and data file S2). Eighteen of these regions have been identified in previous domestications, including those associated to phenotypic effects related to immunity, neural pathways or processes, pigmentation, and productivity traits associated with milk composition and hair characteristics (data file S2). The modest overlap with previous studies is mainly due to differences in sample sets, selection scan methodology, and different versions of the reference genome (text S3).

A KEGG analysis showed that 9 of the 14 significantly enriched pathways (hypergeometric test, adjusted $P < 0.01$) are immune related (hypergeometric test, adjusted $P = 5 \times 10^{-3}$ to 2.35 $\times 10^{-7}$) (Fig. 3B and table S12), including influenza A, malaria, African trypanosomiasis, natural killer cell–mediated cytotoxicity, measles, herpes simplex infection, cytokine–cytokine receptor interaction, hepatitis C, and adipocytokine signaling pathway. We surveyed the literature and identified 40 additional genes with immune function (table S13), obtaining a total of 41 regions that contain 77 immune-related genes. Among these, $F_{\text{ST}}$ is particularly high in the MUC6 region, and we detected 16 missense single-nucleotide polymorphism (SNP) mutations showing $F_{\text{ST}} > 0.88$ in this gene (windowed $F_{\text{ST}} = 0.89$, $\pi$ ln-ratio = 1.01, and XP-EHH = 5.25; Fig. 3C and table S14). Two of the 16 missense mutations were predicted to be deleterious, and these deleterious alleles occur at very low frequency (mean = 0.038) in the domestic goat population (fig. S21). Overall, this region contains 228 SNPs nearly fixed for the derived allele in modern goats (frequency > 95%, absent in bezoars) accounting for 93.8% of such variants in the whole genome (a total of 243, data file S3), illustrating the singular nature of this selection signal.

The selection enrichment analysis furthermore pinpoints 12 genes involved in neuroactive ligand–receptor interaction (hypergeometric test, adjusted $P = 3.70 \times 10^{-5}$). We observed an additional 37 genes linked to other functions in the nervous system that were under selection during goat domestication (table S15). One genomic region located on chromosome 15 shows the strongest combined signal of selection ($F_{\text{ST}} = 0.90$, $\pi$ ln-ratio = 3.20, and XP-EHH = 8.09) (Fig. 3A). Evidence for large negative Tajima’s $D$ values, high composite likelihood ratio (CLR) scores, and extensive haplotype sharing in domestic goats all suggest strong positive selection at this locus (Fig. 3D and fig. S22B). This locus contains 15 SNPs almost fixed for the derived allele in domestic goats and harbor two protein-coding genes, STIM1 and RRM1 (data file S3). STIM1 is an endoplasmic reticulum calcium sensor involved in regulating Ca$^{2+}$ and metabotropic glutamate receptor signaling in the neural system (25, 26). It is known for its role in modulating the excitability of Purkinje neurons in the cerebellum, which play a key role in motor learning and the integration of sensory–motor information (27). RRM1 encodes ribonucleoside-diphosphate reductase large subunit, an enzyme essential for the production of deoxyribonucleotides,
In ratio ($\theta_{\text{Bezoar}} / \theta_{\text{Domestic}}$) between bezoars and domestic goats. The dashed vertical and horizontal lines indicate the significance threshold (corresponding to $Z_{0.01}$) used for extracting outliers. (B) KEGG pathways identified as significantly overrepresented (hypergeometric test, adjusted $P < 0.01$). (C and D) Selective sweep and distribution of the recombination (RHO) on chromosome 29 (46.22 to 46.31 Mb) (C) and selective sweep on chromosome 15 (32.24 to 32.37 Mb) (D). The putative sweep region is additionally validated by Tajima’s $D$, CLR test. Horizontal dashed lines represent the whole-genome mean for the corresponding parameters. Gene annotations in the sweep region and SNPs nearly fixed for derived alleles in domestic goats (C and D) are indicated at the bottom.

Introgressed $MUC6$ plays a role in pathogen resistance

The $MUC6$ region constitutes the only intersection between the selection and introgression scans. In sheep and cattle, $MUC6$ is associated with gastrointestinal parasite resistance (29, 30). The results of transcriptome sequencing, quantitative real-time polymerase chain reaction (qPCR), and immunohistochemistry revealed that $MUC6$ is specifically expressed in the abomasum and duodenum of goat (Fig. 4A and fig. S23). Structurally, $MUC6$ has a high and variable number of tandem repeats (VNTRs) rich in Pro, Thr, and Ser residues, which can influence the covalent attachment of O-glycans (31). We therefore used PacBio SMRT transcriptome sequencing to investigate the difference between the $MUC6^B$ and $MUC6^B$ at the transcriptional level. Besides a number of small indels, an 82–amino acid deletion containing three copies of VNTR after the 2789th amino acid was uniquely identified in $MUC6^B$ compared to $MUC6^B$ (figs. S24 and S25). The different number of VNTRs in these two haplotypes may influence the function of $MUC6$, which is key to the generation of gastrointestinal mucous gel (23, 32). To examine the potential difference of pathogen resistance of the $MUC6^B$ and $MUC6^B$ variants, we carried out an epidemiological survey in a polymorphic goat population. By evaluating fecal egg counts (FEC) for gastrointestinal nematodes in 143 $MUC6^B$ and $MUC6^B$ goats, we found that the goats carrying $MUC6^B$ exhibited lower FEC than those carrying the other two genotypes (Fig. 4B and data file S4), implying that the goats carrying $MUC6^B$ might be more resistant to gastrointestinal...
nematodes. To control for the genetic background, we also sequenced a subset of 117 animals to 5 average depth (data file S4). A genome-wide association analysis for FEC by incorporation of principal component covariates and pairwise relatedness found a strong association signal that contains MUC6 locus on chromosome 29 (42.77 to 51.33 Mb) (Fig. 4C). These results support that the introgressed MUC6D in domestic goats is most likely under selection due to its advantage in the host innate immune response toward potential gastrointestinal pathogens (23).

The origin and diffusion of domestic STIM1-RRM1 and MUC6 alleles

An in-depth genetic survey of ancient genomes throughout the Near East revealed that two ancient Balkan goats dating to ~8100 years ago carried STIM1-RRM1D (fig. S28 and data file S5). MUC6D appeared later at ~7200 years ago in Southwest Iran (fig. S28 and data file S6), a period characterized by an increased density of settlements in the Fertile Crescent (33). Herding goats at higher densities in a crowded and disease-prone anthropogenic environment would likely have exerted an increased selective pressure for livestock pathogen resistance (34). The first detected ancient animals having STIM1-RRM1D or MUC6D both carry mitochondrial haplogroup A, although this mitochondrial haplogroup had a low frequency and narrow distribution before 7500 years ago (6). By ~6500 years ago, the frequency of the STIM1-RRM1D and MUC6D increased to ~60% in the Near East (Fig. 5A) and spread to China ~3900 years ago (Fig. 5B), concomitant with the consolidation and diffusion of livestock-based economies throughout Eurasia (35, 36). The expansion of these two selected loci was also contemporaneous with the overwhelming spread of mitochondrial haplogroup A. In contrast, the frequencies of the two major Y-chromosome haplogroups remained relatively unchanged over time (Fig. 5B).

DISCUSSION

The present study generated genomic data from a substantial number of domestic and wild Capra species to characterize adaptive introgression and genetic changes during goat domestication. Collectively, both the selective sweep enrichment analyses and the two outstanding selection signals found in MUC6 and STIM1-RRM1 in domestic goats are consistent with the hypothesis that genes related to pathogen resistance (37) and behavior have been particularly targeted during goat domestication. Several studies have shown that adaptive introgression can provide beneficial alleles that allow the recipient populations to adapt to new environments (12, 38, 39). In humans, immune-related loci acquired substantially advantageous alleles by immigrant modern
humans admixing with archaic humans who were probably well adapted to local environments and pathogens through their long exposure to them (40–42). We here report numerous genomic segments putatively introgressed from an ibex-like species. These introgressed segments are enriched in genes with an immune function, suggesting that historic gene flow from wild species played an important role in shaping the diversity of immune system phenotypes in domestic goat and possibly increased its adaptive potential. When comparing the set of putatively introgressed alleles in each segment to our sampled ibex-like genomes, we find that most match rates range from 0.5 to 0.8 (fig. S17), indicating the degree of similarity between the introgressing and our sequenced ibex-like individuals. Additional wild Capra genome sequences in the future may clarify the origin of these alleles, although it is also possible that the introgression came from a now extinct branch on the Capra tree.

In particular, a number of lines of evidence support that gene flow between West Caucasian tur and goats occurred before the onset of domestication. The West Caucasian tur is distributed around the Black Sea coast (Fig. 2A), geographically close to the goat domestication center. It inhabits a humid, subtropical environment where the expected pathogen exposure and parasite load may be considerably higher than in more arid regions of Southwest Asia. The rigorous identification of introgression segments in domestic goats shows that the nearly fixed MUC6D was introgressed from a West Caucasian tur-like species. Although the earliest sign of introgression from West Caucasian tur is old (>47,000 years ago), we did not find the MUC6D in ancient and modern sampled bezoars. Two possibilities can explain this observation. First, given the pervasive gene flow between different Capra species (8, 9, 43), we cannot rule out the possibility of multiple introgression from ibex-like species before
and after goat domestication. Second, because of the scarcity of ancient and modern bezoars, it may be that the introgressed variants were segregating at low frequencies in these populations. Nonetheless, our results indicate that the introgression of advantageous MUC6<sup>D</sup> might enhance innate immune reactivity against potential gastrointestinal pathogens under a grazing agro-ecological niche (44) with previously unknown infectious disease pressures.

Our dataset including several ancient goat samples allowed us to tentatively track the emergence and spread of the advantageous variants in these two important domestication loci (Fig. 5). The first occurrences of the STIM1-RRM1<sup>D</sup> and MUC6<sup>D</sup> locus coincided with the otherwise rare mitochondrial DNA (mtDNA) haplogroup A, and the three increased in frequency roughly concurrently (Fig. 5A), becoming nearly fixed in modern goat populations. These results suggest that the global diffusion of variations central to the domestication process was possibly mediated by a subset of female goats carrying the mtDNA haplogroup A. Despite this, modern goat populations remain differentiated from each other. The most likely explanation we can think of is that even Neolithic goat husbandry entailed some kind of breeding strategy under which immigrant matriline containing globally advantageous alleles were deliberately crossbred with local populations, which probably carried local genetic adaptations, rather than simply replacing local populations.

Overall, our results provide evidence for adaptive introgression and the genetic basis of selected traits during the domestication of goats. We highlight that livestock domestication is a dynamic evolutionary process, with adaptive leaps driven by introgression and selection. Our results indicate that domestication may have profound impacts on neural traits and pathogen resistance, which helped manage herds to adapt to an anthropogenic environment.

**MATERIALS AND METHODS**

**Sample collection**

We collected 106 samples including 88 domestic goats (C. hircus), one bezoar (C. aegagrus), three Markhors (C. falconeri), three Siberian ibex (C. sibirica), one Alpine ibex (C. ibex), one West Caucasian tur (C. caucasica), five ancient goats spanning from the Neolithic to the Middle Ages, and four Nubian ibex hybrids (C. nubiana × C. hircus) for whole-genome sequencing. Among them, the Nubian ibex hybrids offer a potential source for determining whether the introgressed haplotypes could originate from the Nubian ibex or closely related ibex. All procedures involving sample collection were approved by the Northwest A&F University Animal Care Committee (permit number: NWAFAC1019), making all efforts to minimize invasiveness. The C. caucasica (tooth) sample was obtained from a zoo specimen donated to Ménagerie du Jardin des Plantes in 1976 of unknown provenance (wild or captive-born) but with no evidence of recent admixture (text S1). The individual died in 1982 and was subsequently sampled by the Muséum National d’Histoire Naturelle (MNHN, sample ID MNHN ZM AC 1982-1092). In addition, we also obtained whole-genome data and their geographical coordinates of the sampling sites from previous studies (5, 15). Details of the samples used in this study are given in tables S1 to S4.

**DNA extraction and sequencing**

For modern samples, genomic DNA was extracted from whole blood using the standard phenol-chloroform method and checked for quality and quantity on the Qubit 2.0 fluorometer (Invitrogen). Library construction for resequencing was performed with 1 to 3 μg of genomic DNA using standard library preparation protocols and insert sizes from 300 to 500 base pair (bp). All libraries were sequenced on either the Illumina HiSeq 2500 or X Ten platforms.

For the historical sample (C. caucasica), tooth pulverization, DNA extraction, and Illumina sequencing library preparation were performed in Trinity College Dublin, Ireland as described in (6). The double-stranded DNA sequencing library was constructed without uracil-DNA glycosylase treatment. The sample library was then sequenced using a HiSeq 2500 platform (single end, 1 × 101 bp). Details on the processing of five ancient samples are described in the Supplementary Materials.

**Read alignment and variant calling**

Filtered reads from all individuals were aligned to the latest goat reference genome by the Burrows-Wheeler Aligner (BWA v0.7.15). To obtain high-quality SNPs, we carried out SNP calling and genotyping at two separate stages. Variants were found on a population scale (without outgroup argali Ovis ammon) using a SAMtools model implemented in the analysis of next-generation sequencing data (ANGSD) (45) with “-only_proper_pairs_1 -uniqueOnly 1 -remove_bads 1 -minQ 20 -minMapQ 30 -C 50 -doMaf 1 -doMajorMinor 1 -GL 1 -setMinDepth [Total_read_depth/3] -setMaxDepth [Total_read_depth*3] -doCounts 1 -dospstat 1 -SNP_pval 1.” The sites that could be called in at least 90% of the samples and had a strand bias score below 90% were retained. A P value threshold of 1 × 10<sup>−6</sup> and minor allele frequency value > 1/2n were used as SNP discovery criteria, where n was the sample size. We also kept the sites at which the sampled individuals were homozygous for the alternative alleles. To obtain the hard-called genotypes in all Capra species and outgroups, we used a workflow adapted from GATK v3.7.0 HaplotypeCaller (46) best practice. Briefly, the genome variants, in genomic variant call format (gVCF) for each sample, were identified with the HaplotypeCaller. After all the gVCF files were merged, a raw population genotype file with the SNPs and indels was created. Optional variant hard filter values were applied for SNPs using GATK VariantFiltration as Quality by Depth (QD) < 2.0, root mean square of Mapping Quality (MQ) < 40.0, Fisher Strand (FS) > 60.0, HaplotypeScore > 13.0, and MQRankSum ≤ 12.5. Last, only biallelic SNPs identified by both GATK and ANGSD that all show no more than 10% missing data were extracted using VCFtools v0.1.15 (47) with “--min-alleles 2 --max-alleles 2 --max-missing 0.9.” Then, all sample sets of filtered variant calls were used for imputation and phasing using Beagle v4.1 (48, 49) with default parameters, except for “niterations,” which was set to 10.

To retrieve the nucleotides that were accessible to variant discovery (used in the demographic estimation), hard filtering of the invariant sites was carried out with thresholds based on sequencing coverage (one-third–to threefold of corresponding mean depth) and missing data rate (no more than 10%) following the same strategies adopted for variant sites.

**Population structure and phylogenetic analysis**

Neighbor-joining tree was constructed for the whole-genome SNP set (table S5) using MEGA v6.0 based on pairwise genetic distances matrix calculated by PLINK v1.9. We also made use of the 5,043,096 fourfold degenerate sites to build an ML tree using RAxML v8.2.9. PCA was performed using smartpc, part of the EIGENSOFT v6.1. A Tracy-Widom test was used to determine the significance level of the eigenvectors. ADMIXTURE was used to perform an unsupervised...
clustering analysis. We increased the number of predefined genetic clusters from \( k = 2 \) to \( k = 7 \) and ran the analysis 20 times for each \( k \).

We further compared the individual-level haplotype similarity using fineSTRUCTURE v2.1.3. TreeMix v1.12 was also used to infer a population-level phylogeny using the ML approach.

**Demographic reconstruction**

To infer patterns of effective population size and population separations over historical time, we used MSMC2, using the statistically phased autosomal genotypes. We also used the SMC++, which does not rely on haplotype phase information. In addition, we calculated the likelihood of the observed site frequency spectrum conditioned on several demographic models using \( \delta \alpha \delta \) (fig. S13). For the best model, nonparametric bootstrapping (100 replicates) was performed to determine the variance of each parameter (fig. S14 and table S7).

To obtain reliable time estimates, we used a phylogenetic comparison (50) of goat and sheep (Ovis aries) to calibrate the mutation rate for goats. Using an average generation time of 2 years for goats (51), we estimated a mutation rate of \( 4.32 \times 10^{-9} \) per site per generation, which is similar to that obtained in (52).

**Gene flow analysis**

We computed \( f_2 \) statistics and \( D \) statistics to formally test hypotheses of gene flow. To identify regions introgressed into modern domestic goats from ibex-like species, we calculated \( D \) statistics using non-overlapping 20-kb sliding windows across the genome. Windows within top 5% of the \( D \) statistics were considered as outliers and were further examined. An introgressed segment should have low sequence divergence to the putative donor species, whereas sequence divergence between the donor species and bezoar should be higher.

Therefore, we also calculated the identity by state distance matrix. Differences in sequence divergence were determined by a \( t \)-test–based approach using a \( P \) value of 0.01 as cutoff. The windows which passed the above two filtering steps were merged as the significantly diverged regions.

We then applied Sprime (53) for investigating whether introgression can explain the significantly diverged regions. The first step of Sprime is to read the phased genotypes of the target and outgroup individuals. The outgroup is a population that is closely related to the ancestor(s) of the target population but that is not expected to have experienced admixture from ibex-like species that contributed to the target population. The 24 modern bezoars are used as the outgroup population, although they may themselves contain introgressed sequence from ibex-like species. This may lead to the occurrence of false negatives but will minimize the risk of false positives. We consider this an acceptable trade-off given our objective. Sprime first groups variants into three classes: those common in the outgroup, those not seen in the outgroup, and those uncommon but present in the outgroup. The threshold frequency used to distinguish common versus uncommon variants is 0.01 (53) for the analyses presented here. The common alleles (those with frequency > 0.01 in 24 modern bezoars) are excluded from further consideration because these variants in domestic goats are more likely inherited from their wild progenitor (bezoar). We assumed a score threshold of 150,000 (53) and a mutation rate of \( 4.32 \times 10^{-9} \) per site per generation to identify introgressed alleles in domestic goat genomes. The next step is to find putatively introgressed segments (i.e., sets of alleles) for each significantly diverged region. We process one region at a time and find the set of alleles with the highest score. These comprise the putative introgressed segment. The segment boundaries were defined by the position of first and last introgressed alleles. We consider only segments with at least 100 putatively introgressed alleles that are of nonbezoar origin. To determine whether a target individual carries the introgressed haplotype, at least 50% of introgressed alleles must be found in that haplotype. We primarily focus on the most common introgressed regions (i.e., introgressed haplotype frequency > 0.1) from ibex-like species to domestic goats.

We further checked the introgression status by constructing ML trees with MEGA v6.0 and searching for domestic goats located within the ibex-like clade (fig. S16). For each region, we also reported a match if the genotype of ibex like species included the putative introgressed allele and a mismatch otherwise. The match rate was calculated as the number of matches divided by the number of compared sites (matches and mismatches) (fig. S17). Note that the match rate should not imply that the sampled ibex-like genomes represent the direct source population(s), but the degree of similarity between the introgressing and sampled ibex-like species.

**Selective sweep analysis**

We screened for sweeps selected during goat domestication by parsing specific 50-kb windows that showed low diversity in domestic goats and had high divergence (a high fixation index, \( F_{ST} \), and haplotype differences, XP-EHH) between domestic goats and wild goats. After calculating all tests, the windows with \( P \) values less than 0.005 (Z test) were considered significant signals. Candidate genes under selection were defined as those overlapped by sweep regions or within 20 kb of the signals. Two additional statistics, the Tajima’s \( D \) and CLR test were applied to confirm the top signals.

**Functional enrichment analyses**

We characterized the most relevant functions of the protein-coding genes overlapping with the introgressed regions and selective sweeps by searching for overrepresented KEGG pathways. Goat protein sequences were used to conduct functional enrichment tests on the target genes using the KOBAS 3.0 server (54). The \( P \) value was calculated using a hypergeometric distribution. False discovery rate (FDR) correction was performed to adjust for multiple testing. Pathways with an FDR-corrected \( P \) value of <0.01 were considered statistically significantly enriched (tables S10 and S12).

**Genotyping loci underlying domestication in ancient genomes**

To investigate the genotypes of the \( STIM1-RRM1 \) and \( MUC6 \) loci in ancient samples, we used a total of 243 SNPs (15 SNPs within \( STIM1-RRM1 \) locus and 228 SNPs within \( MUC6 \) locus, respectively), which showed derived allele frequency > 0.95 in 164 modern domestic goats (defined as domestic haplotypes: \( STIM1-RRM1^{D} \) and \( MUC6^{D} \)) and was absent in 24 modern bezoars and 4 ancient bezoars (Hovk1, Direkli1-2, Direkli6, and Direkli5) (defined as bezoar haplotypes: \( STIM1-RRM1^{B} \) and \( MUC6^{B} \)). Because of the low coverage for some ancient genomes, we used allele reads count at SNP positions to determine the genotypes of ancient goats at \( STIM1-RRM1 \) locus and \( MUC6 \) locus. For each ancient goat and SNP, we determined the numbers of reads corresponding to the ancestral and derived allele using GATK v3.7.0 UnifiedGenotyper (--min_base_quality_score 15 --output_mode EMIT_ALL_SITES --standard_min_confidence_threshold_for_calling 20) (46). For each locus in each ancient goat, we calculated the proportion of the summed derived allele reads.
count versus total allele reads count. If the proportion was lower than 10%, the genotype of ancient goat was considered to be homozygous for ancestral allele (bezoar like). The genotype was set to heterozygous if the proportion was between 10 and 90%. If the proportion was more than 90%, it was classified as homozygous for derived allele (domestic like). We set the genotype to be missing when the ancient goats only had one single informative read (mapping to the predefined SNPs) mapped to the locus.

Expression and epidemiologic survey related to MUC6 locus

To further test the genotype-phenotype associations, we first examined the genotypes at the MUC6 locus using blood samples from breeding rams from a company located in Inner Mongolia, China. This company runs more than 9000 cashmere goats, mainly for superfine cashmere (fiber diameter < 14 μm). The animals are dewormed routinely with oxendazole, ivermectin, avermectin, and levamisole three times annually. Of the 20 tested breeding rams, we identified 4 animals with MUC6D/MUC6B genotype. We sampled their offspring partly on the basis of the breeding records. Blood samples of 495 animals were collected in January 2019.

Genotypic and phenotypic analysis

DNA was extracted from the blood samples. The MUC6 locus was genotyped by PCR (forward primer: 5′-CACGTATCTCCCATACATC-3′ and reverse primer: 5′-ATAGCCCCCTCCAGATGC-3′) was used as internal reference. Gene expression results were calculated using the delta-delta cycle threshold (2−ΔΔCT) method.

Immunohistochemistry analysis

For histological analysis, dissected goat tissues (abomasum pyloric and duodenal bulb) were fixed in 4% paraformaldehyde and embedded in paraffin for sectioning (5 μm thick). The tissues sections were deparaffinized, and rehydration was followed by antigen retrieval using a citrate-buffered solution in a microwave at 100°C for 15 min. After cooling down to room temperature, quenching of endogenous peroxidase, and protein block, the sections were treated with a primary antibody (cat. no. D161001, Sangon Biotech, Shanghai, China) overnight at 4°C. Negative control sections were obtained by incubating with the primary antibody diluted buffer. Subsequently, the secondary antibody (cat. no. D110058, Sangon Biotech, Shanghai, China) was used to detect primary antibody. Specific protein immunoreactivity was visualized with the substrate chromogen dianobenzidine. Last, the slides were rinsed in water, counterstained with hematoxylin, and mounted with cover slips.

Full-length transcript sequencing

To determine the sequence differences between the domestic (MUC6D) and bezoar (MUC6B) MUC6 haplotypes, total RNA was isolated from the abomasum pyloric with high expression of MUC6 from two heterozygous goats and sequenced by PacBio Sequel. Pacbio SMRTbell libraries were sequenced on two separate SMRT cells (Annroroad Gene Technology Co. Ltd., Beijing, China). A total of 23,338,976 subreads were generated with a mean accuracy of 80% and an average length of 1755 nt. High-quality circular consensus sequences (CCSs) were obtained using the Iso-Seq 3 application in the PacBio SMRT Analysis v6.0.0 (https://github.com/PacificBiosciences/IsoSeq), with parameters "--noPolish --minPasses 1." Last, we got a total of 960,271 CCSs. These CCSs were aligned to the latest goat reference genome using Minimap2 (55), and we picked out 251 CCSs that mapped to the MUC6 mRNA sequence. We used the CCS that belonged to the bezoar haplotype to manually assemble the mRNA sequence of the MUC6B. Then, we realigned the MUC6B mRNA sequence to MUC6D mRNA sequence (XM_018042766.1) by MEGA v6.0 and carefully checked the indels. The abundance of tandem repeats in MUC6 (XP_017898255.1) was examined by using the rapid automatic detection and alignment of repeat finding program (https://www.ebi.ac.uk/Tools/pfa/radar/) (56). Three major types of repeat units with distinct sequence features were observed. We detected a 246-bp insertion in the MUC6B mRNA sequence located at the 32nd exon, containing three copies of type III units between the 2789th amino acid and the 2871th amino acid (fig. S24). We further validated this insertion by aligning short reads from West Caucasian tur to the MUC6B mRNA sequences (fig. S25).

Test population

To determine the sequence differences between the domestic (MUC6D) and bezoar (MUC6B) MUC6 haplotypes, total RNA was isolated from the abomasum pyloric with high expression of MUC6 from two heterozygous goats and sequenced by PacBio Sequel. Pacbio SMRTbell libraries were sequenced on two separate SMRT cells (Annroroad Gene Technology Co. Ltd., Beijing, China). A total of 23,338,976 subreads were generated with a mean accuracy of 80% and an average length of 1755 nt. High-quality circular consensus sequences (CCSs) were obtained using the Iso-Seq 3 application in the PacBio SMRT Analysis v6.0.0 (https://github.com/PacificBiosciences/IsoSeq), with parameters "--noPolish --minPasses 1." Last, we got a total of 960,271 CCSs. These CCSs were aligned to the latest goat reference genome using Minimap2 (55), and we picked out 251 CCSs that mapped to the MUC6 mRNA sequence. We used the CCS that belonged to the bezoar haplotype to manually assemble the mRNA sequence of the MUC6B. Then, we realigned the MUC6B mRNA sequence to MUC6D mRNA sequence (XM_018042766.1) by MEGA v6.0 and carefully checked the indels. The abundance of tandem repeats in MUC6 (XP_017898255.1) was examined by using the rapid automatic detection and alignment of repeat finding program (https://www.ebi.ac.uk/Tools/pfa/radar/) (56). Three major types of repeat units with distinct sequence features were observed. We detected a 246-bp insertion in the MUC6B mRNA sequence located at the 32nd exon, containing three copies of type III units between the 2789th amino acid and the 2871th amino acid (fig. S24). We further validated this insertion by aligning short reads from West Caucasian tur to the MUC6B mRNA sequences (fig. S25).
null


7. S. J. Atienza, Demographic and genetic history of sheep and goats suggested by nuclear, mtDNA and Y chromosome markers in The Role of Biotechnology for the Conservation and Improvement of Capra hircus (Artiodactyla). Byulleten Moskovskogo Obshchestva Ispytatelei Prirody, Otdel biologicheskui


Acknowledgments: We thank members of the NextGen project for sharing data. We thank L. Mohammad for providing Nubian ibex hybrid blood samples. We thank High-Performance Computing (HPC) of NWAU for providing computing resources. Funding: This project was supported by grants from the National Natural Science Foundation of China (31822052 and 31572381 to Y.J. and 31572369 to Y.Ch.), the Talents Team Construction Fund of Northwestern Polytechnical University (NPU) and Strategic Priority Research Program of CAS (XDB13000000 to W.W., R.H., G.Z., K.G.D., D.G.B., L.Co., and T.S.S. revised the manuscript. From Z.Y., Y.L., Y.Ca., Q.C., J.Li., K.W., X.P., Y.W., S.H., M.G., T.Z., Y.Z., Y.G., Y.X., N.X., and Y.J. and Y.Ch. conceived the project and designed the tree-based partition. A. R. M. Albrechtsen, Y. Jiang, N. Kong, J. Wang, Q. Zhang, T. Li, R. Zhao, G. Jia, T. L. Faraut, J. Womack, Y. Zhang, J. Kijas, N. Cockett, X. Xu, S. Zhao, J. Wang, W. Wang, S. Servin, B. Sayre, B. Zhu, S. Dreyney, R. Moore, W. Nie, Y. Shen, R. Zhao, G. Jia, L. Liu, T. C. Freeman, K. M. Summers, A. L. Archibald, D. A. Hume, A high resolution atlas of gene expression in the domestic sheep (Ovis aries). PLOS Genet. 13, e1006997 (2017).


Downloaded from http://advances.sciencemag.org/ on June 15, 2020
The origin of domestication genes in goats

Zhuqing Zheng, Xihong Wang, Ming Li, Yunjia Li, Zhirui Yang, Xiaolong Wang, Xiangyu Pan, Mian Gong, Yu Zhang, Yingwei Guo, Yu Wang, Jing Liu, Yudong Cai, Qiuming Chen, Moses Okpeku, Lucia Coll, Dawei Cai, Kun Wang, Shisheng Huang, Tad S. Sonstegard, Ali Esmaeilzadeh, Wenguang Zhang, Tingting Zhang, Yangbin Xu, Naiyi Xu, Yi Yang, Jianlin Han, Lei Chen, Joséphine Lesur, Kevin G. Daly, Daniel G. Bradley, Rasmus Heller, Guojie Zhang, Wen Wang, Yulin Chen and Yu Jiang

Sci Adv 6 (21), eaaz5216, DOI: 10.1126/sciadv.aaz5216