

SCIENTIFIC REPORTS

OPEN

Identification of selection signals by large-scale whole-genome resequencing of cashmere goats

Xiaokai Li¹, Rui Su^{1,2,3,4,5}, Wenting Wan⁶, Wenguang Zhang¹, Huaizhi Jiang⁷, Xian Qiao¹, Yixing Fan¹, Yanjun Zhang^{1,2,3,4}, Ruijun Wang^{1,2,3,4}, Zhihong Liu^{1,2,3,4}, Zhiying Wang^{1,2,3,4}, Bin Liu⁸, Yuehui Ma⁹, Hongping Zhang¹⁰, Qianjun Zhao⁹, Tao Zhong¹⁰, Ran Di⁹, Yu Jiang¹¹, Wei Chen^{12,14}, Wen Wang⁵, Yang Dong^{12,13,14} & Jinquan Li^{1,2,3,4}

Inner Mongolia and Liaoning cashmere goats are two outstanding Chinese multipurpose breeds that adapt well to the semi-arid temperate grassland. These two breeds are characterized by their soft cashmere fibers, thus making them great models to identify genomic regions that are associated with cashmere fiber traits. Whole-genome sequencing of 70 cashmere goats produced more than 5.52 million single-nucleotide polymorphisms and 710,600 short insertions and deletions. Further analysis of these genetic variants showed some population-specific molecular markers for the two cashmere goat breeds that are otherwise phenotypically similar. By analyzing F_{ST} and θ_{π} outlier values, we identified 135 genomic regions that were associated with cashmere fiber traits within the cashmere goat populations. These selected genomic regions contained genes, which are potential involved in the production of cashmere fiber, such as *FGF5*, *SGK3*, *IGFBP7*, *OXTR*, and *ROCK1*. Gene ontology enrichment analysis of identified short insertions and deletions also showed enrichment in keratinocyte differentiation and epidermal cell differentiation. These findings demonstrate that this genomic resource will facilitate the breeding of cashmere goat and other *Capra* species in future.

Cashmere goat grows an outer coat of coarse hairs from its primary hair follicles and an inner coat of fine wool from its secondary hair follicles. This special fine wool fiber is known as cashmere wool or cashmere^{1,2}. It is finer and softer than sheep's wool, and contributes high economic values to the textile industry and impoverished remote areas^{3,4}. China is a major cashmere producer in the world, and has rich native cashmere goat genetic resources. In 2012, China supplied about 70% (18 thousand tons) of cashmere wool to the world market⁵. The Inner Mongolia (three subtypes: Alashan, Aerbasi, and Erlangshan⁶) and Liaoning cashmere goats are two native

¹College of Animal Science, Inner Mongolia Agricultural University, Hohhot, Inner Mongolia, 010018, China. ²Key Laboratory of Animal Genetics, Breeding and Reproduction - Inner Mongolia Autonomous Region, Inner Mongolia Agricultural University, Hohhot, Inner Mongolia, 010018, China. ³Key Laboratory of Mutton Sheep Genetics and Breeding, Ministry of Agriculture, Inner Mongolia Agricultural University, Hohhot, Inner Mongolia, 010018, China. ⁴Engineering Research Center for Goat Genetics and Breeding - Inner Mongolia Autonomous Region, Inner Mongolia Agricultural University, Hohhot, Inner Mongolia, 010018, China. ⁵State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, 650223, China. ⁶Center for Ecological and Environmental Sciences, Key Laboratory for Space Bioscience & Biotechnology, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, China. ⁷College of Animal Science and Technology, Jilin Agricultural University, Changchun, Jilin, 130118, China. ⁸Institute of Animal Husbandry, Academy of Agriculture and Stockbreeding Sciences, Hohhot, Inner Mongolia, 010030, China. ⁹The Key Laboratory for Farm Animal Genetic Resources and Utilization of Ministry of Agriculture of China, Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, 100193, China. ¹⁰Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, College of Animal Science and Technology, Sichuan Agricultural University, Chengdu, 611130, China. ¹¹College of Animal Science and Technology, Northwest A&F University, Yangling, 712100, China. ¹²College of Biological Big Data, Yunnan Agriculture University, Kunming, Yunnan, 650504, China. ¹³BGI-Shenzhen, Shenzhen, Guangdong, 518083, China. ¹⁴Yunnan Research Institute for Local Plateau Agriculture and Industry, Kunming, Yunnan, 650201, China. Xiaokai Li, Rui Su and Wenting Wan contributed equally to this work. Correspondence and requests for materials should be addressed to W.W. (email: wwang@mail.kiz.ac.cn) or Y.D. (email: loyalyang@163.com) or J.L. (email: lijinquan_nd@126.com)

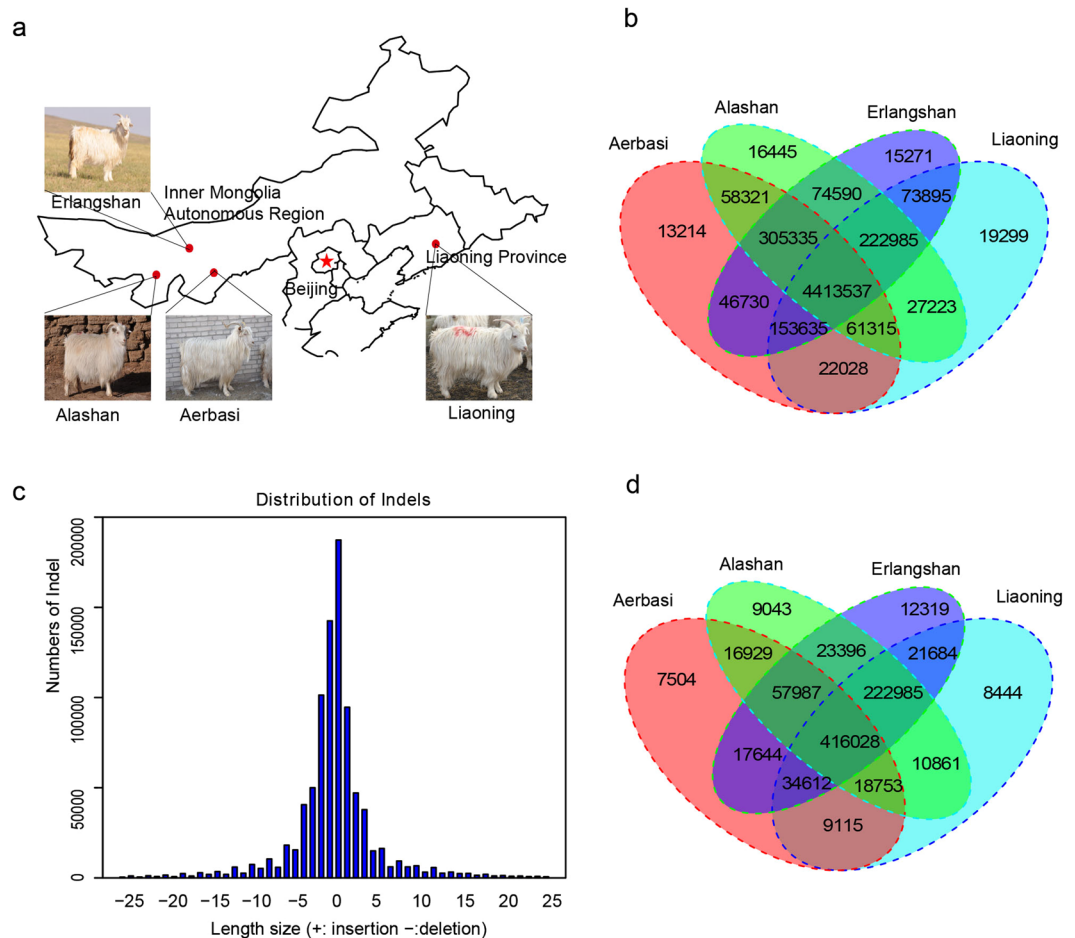


Figure 1. Summary of cashmere goats. **(a)** Geographic map indicating the distribution of the cashmere goats sampled in this study (Photographs were taken by Rui Su and Xiaokai Li). Each red dot represents the location of sampling. The map was generated using the ‘ggmap’ package in R (version 3.4.1) (<https://cran.r-project.org/>)^{60,61} and trimmed by Adobe Photoshop CS6 (<http://www.adobe.com/>). **(b)** Venn diagram of SNVs shows the overlap and population-specific identified SNVs among four cashmere goat populations. **(c)** Distribution of InDels. The length of each bar represents the number of InDels. **(d)** Venn diagram of InDels show the common and population-specific genetic variants among four cashmere goat populations.

breeds characterized by the thin cashmere fiber diameter and high yield of cashmere wool (Fig. 1a)⁷. For this reason, great research interest has been dedicated to finding new goat breed that produces finer and higher yield of cashmere wool^{8–13}.

Over the past decade, the next-generation sequencing technology has markedly facilitated the genetic studies of complex traits in domestic animal^{14–16}. This technology has been used to reveal natural and artificial selection footprint in many species, such as pig^{17,18}, sheep¹⁹, dog^{20,21}, and so on. With whole-genome resequencing of different sheep breeds, researchers have provided a comprehensive insight into the genetic basis of adaptive variation of sheep in different environment. For example, genes *OAR22_18929579-A*, *IFNGR2*, *MAPK4*, *NOX4*, *SLC2A4* and *PDK1* showed an apparent geographic pattern and significant correlations with climatic variation^{19,22}. Based on the draft goat genome assembly CHIR_1.0 and CHIR_2.0²³, several preliminary studies have attempted to explore the economic and adaptive traits in different goat breeds using the whole-genome resequencing strategy^{24–26}. Using parallel sequencing of pooled DNA from eight goat breeds, Wang *et al.* identified several genomic regions under strong selection that were associated with body size (e.g. *TBX15*, *DGCR8*, *CDC25A*, and *RDH16*), cashmere fiber (e.g. *LHX2*, *FGF9*, and *WNT2*), and coat color (e.g. *ASIP*, *KITLG*, and *HTT*)¹⁰. Guan *et al.* identified some candidate genes (e.g. *FGF5*) for improving fiber traits using the whole genome sequence of six cashmere goats and six non-cashmere goats²⁶. Despite these useful findings, the sample sizes of these studies were invariably small to elucidate the genetic basis of cashmere fiber trait. Furthermore, these studies did not include Inner Mongolia and Liaoning cashmere goats in their samples, which may miss important genetic information with regard to cashmere goat traits.

Here, we report the whole-genome resequencing of 70 cashmere goats from the Inner Mongolia and Liaoning regions. Analyses of the genetic variants identified population-specific molecular markers and candidate genomic regions under selective sweep that were related to cashmere traits. This genetic resource will not only help with future genome-wide association studies, but also increase the knowledge regarding the genetic architecture of quantitative traits.

Category		Number of InDels	Percent(%)
3'UTR		26325	0.48
5'UTR		5470	0.01
UTR5;UTR3		13	0
Downstream		32477	0.59
Exonic	nonsynonymous SNV	10606	0.52
	stop gain	81	
	stop loss	23	
	synonymous SNV	16895	
	unknown	1363	
Intergenic		3519939	63.72
Intronic		1877966	34.00
NcRNA_exonic		1250	0.02
NcRNA_intronic		2125	0.04
Splicing		52	0
Updstream		28293	0.51
Upstream/Dowstream		945	0

Table 1. Summary and annotation of SNPs in cashmere goat.

Results and Discussion

Whole-genome sequencing and genetic variant mapping. A total of 611.67 Gb paired-end DNA sequence data were obtained from 70 female cashmere goats on an Hiseq-2000 platform (Illumina, San Diego, CA, USA). About 534.66 Gb high-quality paired-end reads could be mapped to the latest goat reference genome assembly with a 2.61-fold average coverage (Supplementary Table 1). These data yielded 5,523,823 single-nucleotide polymorphisms (SNPs) and 710,600 short insertions and deletions (Indels) (MAF > 0.5; Fig. 1b and d; Supplementary Table 2). Compare to the dbSNP database (https://ftp.ncbi.nih.gov/snp/organisms/goat_9925/VCF/), about 4,819,577 (87.25) SNPs and 643,205 (90.52%) Indels were novel. The average transition-to-transversion (Ti/Tv) ratio was 2.36 for all cashmere goat samples, which indicated relatively low potential random sequencing errors. This number is comparable to previously reported Ti/Tv ratios for Moroccan goat (2.44) and Dazu black goat (2.33)^{25,27}, indicating high accuracy for the identified variants in this study (Supplementary Table 3). The density of SNPs along each chromosome (except X chromosomes) was proportional to the chromosome length (Supplementary Table 2). This result is consistent with the observation that lower proportion of mutant variants could be found on sex chromosomes in goat^{28,29}. Besides, the distribution of site depth of SNP was ranged from 23-fold to 7535-fold, with an average depth of 152.70-fold (Supplementary Fig. 1).

We examined the nucleotide diversity and ratio of heterozygous versus homozygous single nucleotide variations (SNVs) among Inner Mongolia and Liaoning cashmere goat. The higher average ratio of heterozygous versus homozygous SNVs was observed in Erlangshan population (Supplementary Table 3). The overall distributions of Inner Mongolia and Liaoning cashmere goat in terms of nucleotide diversity were similar, of which Alashan population showed lower nucleotide diversity (total average nucleotide diversity = 5.31×10^{-4}) than other populations (Supplementary Table 3 and Supplementary Fig. 2).

About 4,413,537 (79.90%) of identified SNPs were shared among all cashmere goat populations, indicating a high genetic similarity within cashmere goats. This is in line with the report that Inner Mongolia and Liaoning cashmere goats came from a recent common origin⁷. The number of population-specific SNPs was highest in the Liaoning population (19,299 or 0.35%), and was lowest in the Aerbasi population (13,214 or 0.24%) (Fig. 1b). The number of Indels shared among all four populations was 416,028 (58.55%). The numbers of breed-specific Indels ranged from 7504 (1.06%) in the Aerbasi population to 12,319 (1.73%) in the Erlangshan population (Fig. 1d). Compared to Liaoning cashmere goat (19299), Inner Mongolia cashmere goat (529906) have more specific SNPs that may be related to weaker intensive selection breeding.

Annotation of SNPs and Indels. The proportions of SNPs in the intergenic, intronic, and exonic regions of the genome were 73.72%, 34.00%, and 0.52%, respectively (Table 1). Among all identified SNPs, 28,968 SNPs could cause changes in the coding sequences of 9,621 genes, including 10,606 non-synonymous nucleotide substitutions, 81 stop-codon gain mutations, and 23 stop-codon loss mutations in the cashmere goat genomes (Supplementary Data 1). Enrichment analysis of these genes identified receptor activity related categories, such olfactory receptor activity (600 genes, $P = 1.01 \times 10^{-128}$), G-protein coupled receptor activity (754 genes, $P = 2.09 \times 10^{-94}$), transmembrane signaling receptor activity (871 genes, $P = 2.57 \times 10^{-75}$), transmembrane receptor activity (889 genes, $P = 2.17 \times 10^{-72}$), and signaling receptor activity (883 genes, $P = 1.92 \times 10^{-64}$) (Supplementary Data 2 and Supplementary Fig. 3). In addition, enrichment was found in the basic cellular functions, such as the binding of FAD, syntaxin, cytoskeletal protein, metal ion, actin and protein kinase activity.

The identified Indels were 1 bp to 25 bp in length (Fig. 1c), and the total number of deletions is a little larger than the total number of insertions. The frequency of Indels decreased as the sizes of the Indels increased. The proportions of Indels in the intergenic, intronic, and coding sequences of the genome were 63.42%, 34.29%, and 0.09%, respectively (Supplementary Table 4). About 284 Indels (131 deletions and 153 insertions) may result in

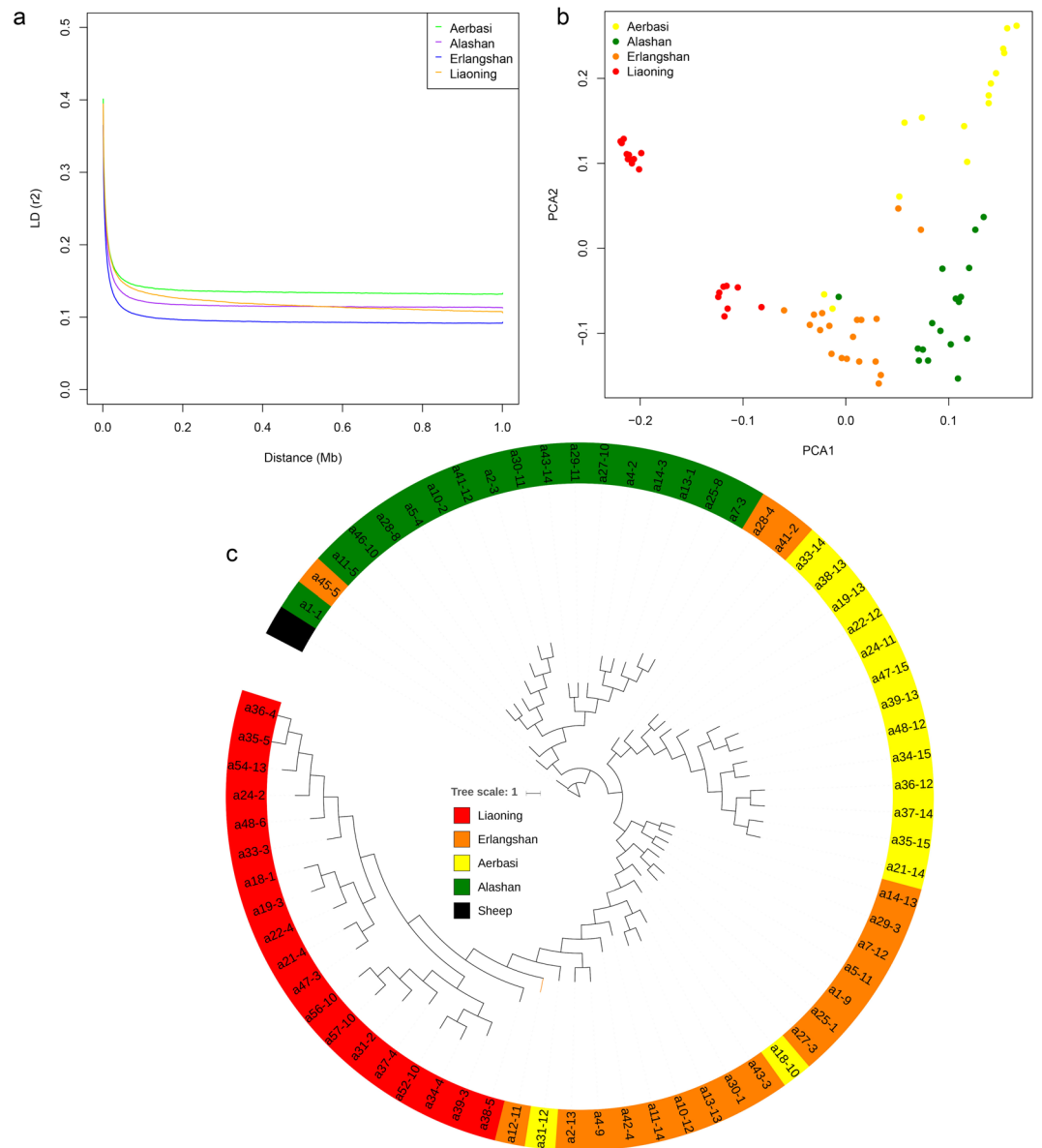


Figure 2. Population genetic relationship analysis. (a) LD patterns of cashmere goats (Liaoning and three subtype of Inner Mongolia cashmere goats). (b) PCA using thinned SNPs as markers. Each dots are index to samples, and each color represent on population. Most samples cluster together based on their geographic distribution. (c) Phylogenetic relationship of cashmere goats. The scale bar represents p distance.

frame-shift mutations in the coding sequences of 249 genes (Supplementary Data 3). GO annotation of these affected genes revealed enrichment in the biological process terms, such as keratinocyte differentiation (GO: 0030216), epidermal cell differentiation (GO:0009913), and skin development (GO:0043588) (Supplementary Data 4 and Supplementary Fig. 4).

Population structure analysis. The phylogenetic relationship of the 70 cashmere goats revealed genetically distinct clusters according to their geographic locations (Fig. 2c). This result was confirmed by the principle component analysis (PCA) using thinned genomic SNPs. The first eigenvector distinguished Liaoning cashmere goats from Inner Mongolia cashmere goat, and the second eigenvector distinguished Aerbasi, Alashan and Erlangshan populations (Fig. 2b). The genetic ancestry analysis with STRUCTURE showed that all cashmere goat samples had a mixed ancestry at $K = 4$ (Fig. 3). Population differentiation index (F_{st}) showed that the Aerbasi population had a higher genetic distance (0.11) from the Liaoning cashmere goats, which is consistent with the results of PCA and STRUCTURE (Supplementary Table 5). The linkage disequilibrium (LD) decay rates were similar between Liaoning and Aerbasi populations. The fastest LD decay was observed in the Erlangshan population (Fig. 2a).

Genome-wide selective sweep signals. In order to detect genome selection signals and SNPs related to cashmere fiber traits, we used 14 unpublished genomic data from non-cashmere goats (~12.50-fold average

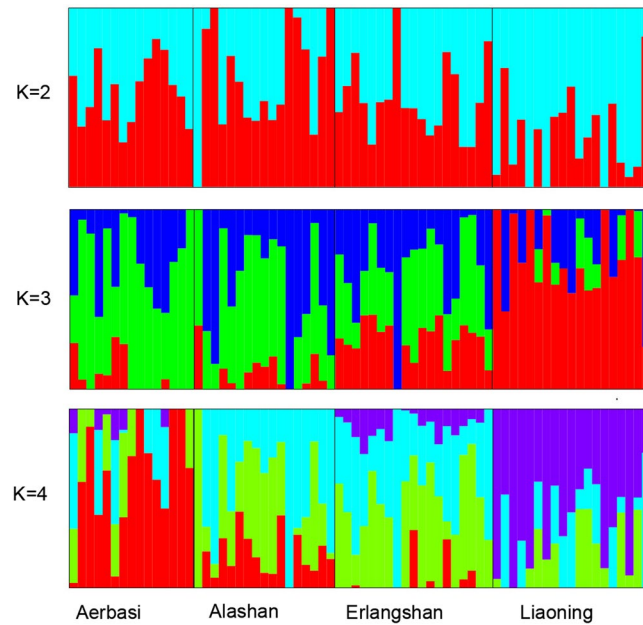


Figure 3. Population structure analysis of cashmere goats using STRUCTURE packages. Each sample is represented by a vertical bar. Enery color represents one ancestral population and the length of each colored segment in each vertical bar represents the proportion contributed by ancestral populations.

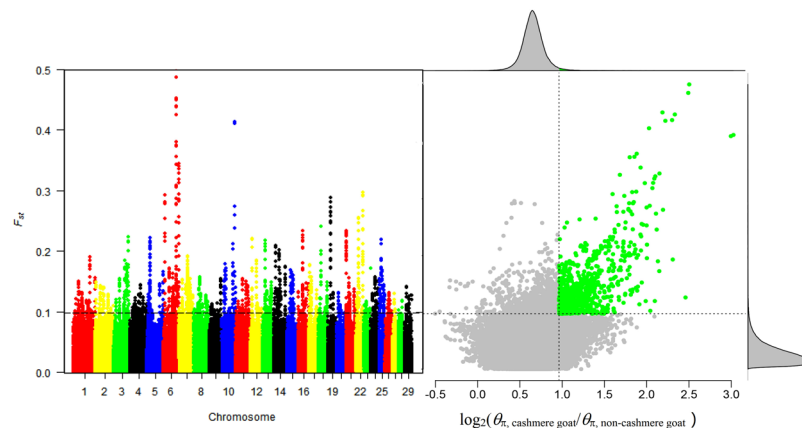


Figure 4. Identification of genomic regions with strong selective sweep signals in Cashmere goats. Distribution of $\log_2(\theta_\pi \text{ ratio}(\theta_{\pi, \text{cashmere goat}}/\theta_{\pi, \text{non-cashmere goat}}))$ and F_{ST} which are calculated in 100 kb windows sliding in 10 kb steps. Data points located to the right of the vertical lines (corresponding to 5% right tails of the empirical $\log_2(\theta_\pi \text{ ratio})$ distribution, where $\log_2(\theta_\pi \text{ ratio})$ is 1.26) and above the horizontal line (5% right tail of the empirical F_{ST} distribution, where F_{ST} is 0.10) were identified as selected regions for cashmere goat (blue points).

depth) courtesy of our collaborator (Supplementary Table 6). By using both θ_π cut-off ratio and high F_{ST} values, we identified a total of 135 genomic regions under selective sweep containing 650 candidate genes that were associated with cashmere goat traits (Fig. 4). Gene ontology analysis of these candidate genes revealed enrichments in 206 GO terms in the biological processes, 69 GO terms in the molecular functions, and 25 GO terms in the cellular components with a 5% FDR threshold for significance (Supplementary Data 5). KEGG enrichment analysis of these candidate genes identified 36 pathways with a 5% FDR threshold for significance (Supplementary Data 6 and Fig. 5). The variant location within selected genes was shown in Supplementary Data 7.

Candidates genes related to cashmere fiber traits. Candidate genes associated with cashmere fiber traits were identified in several genomic regions under selective sweep, including *ROCK1*, *FGF5*, *PRKCD*, *SGK3*, *IGFBP7*, and *OXTR*. *ROCK1* (Rho-associated protein kinase 1) plays an important role in regulation of keratinocyte proliferation and terminal differentiation in human^{30,31}. *FGF5* (fibroblast growth factor 5) regulates hair length in many species^{32–35}. A previous study showed that disruption of *FGF5* led to more secondary hair follicles and longer fibers in cashmere goat. In a mouse study, overexpression of *PRKCD* (protein kinase C delta) had an inhibitory effect on hair growth. The authors also proposed that *PRKCD* together with *PRKCA* (protein kinase C

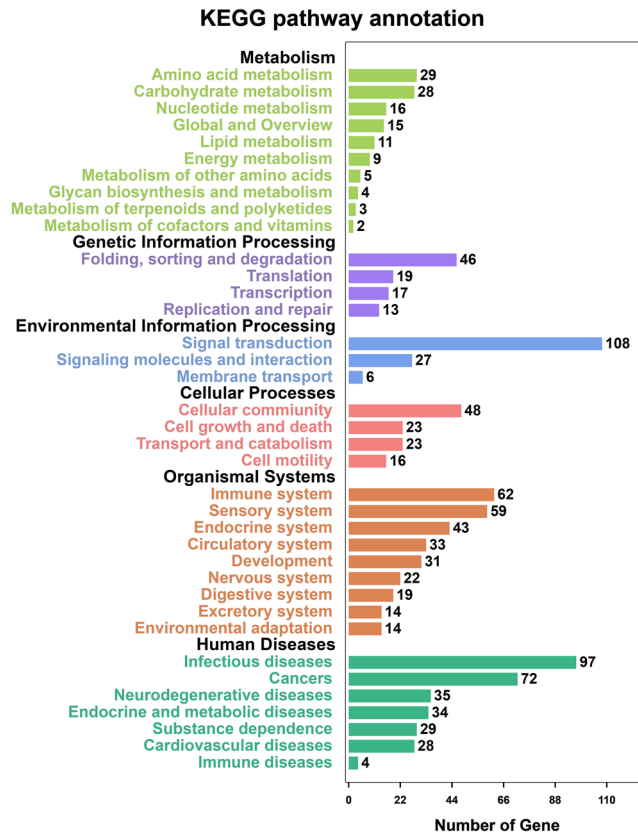


Figure 5. KEGG pathways enrichment analysis of candidate genes under selection within cashmere goats.

alpha) kept hair growth in balance³⁶. *SGK3* (aliases *SGK2*, serum/glucocorticoid-regulated kinase 3) belongs to the PI3K-Akt signaling pathway, and plays an important role in the development of postnatal hair follicle³⁷. Loss of *SGK3* led to reduced proliferation and increased apoptosis of hair follicles in mice^{38–40}. Mutations of *SGK3* were also responsible for the fuzzy hair phenotype in mice⁴¹. *OXT* (oxytocin receptor) is expressed in the primary human dermal fibroblasts and keratinocytes, and *OXT* decreases the proliferation of dermal fibroblasts and keratinocytes in a dose-dependent manner⁴². *IGFBP7* (insulin like growth factor binding protein 7) was found to be one of several keratinocyte-specific genes differentially expressed in keratinocyte⁴³.

Candidate genes related to the adaptation to a cold and dry environment. Cashmere goat usually live in a cold and dry environment. The fine cashmere fibers greatly help these animals to combat heat loss. This adaptive feature is also accompanied by other physiologic mechanisms that help maintain mineral and energy homeostases^{23,44}. For example, *ADCY4* (adenyl cyclase 4) was involved in the regulation of the oxytocin signaling pathway, insulin secretion, adrenergic signaling in cardiomyocytes, rap1 signaling pathway, and cGMP-PKG signaling pathway. Besides, Adenyl cyclase (AC)-stimulated cAMP is involved in cAMP-induced cell proliferation in cultured adrenal cells and a key mediator of Na and water transport^{45,46}. Four other genes *ROCK1* (Rho-associated protein kinase 1), *ACNA1C* (calcium voltage-gated channel subunit alpha1 C), *OXT* (oxytocin receptor) were also involved in the oxytocin signaling pathway. It was reported that they were all functionally related to the regulation of skin development, fat metabolism, and ion homeostasis. In addition, three candidate genes *CACNA2D1* (calcium channel, voltage-dependent, alpha2/delta subunit 1), *AGT* (angiotensinogen), and *PTGER2* (prostaglandin E receptor 2) were involved in the renin secretion pathway. *SLC24A4* (Sodium/Potassium/Calcium Exchanger 4) were located in the classical HIF-1 (hypoxia-induced factors) pathway, which plays a central role in regulating cellular responses to hypoxia¹⁹. *IGFBP7* and *FGF5* have high outlier value indicated under selection, then we further analyzed the allele frequency difference of each SNVs (Supplementary Data 8). It would be interesting to see how the genetic variations in these genes affect the phenotypes of cashmere goat in future studies.

Conclusion

The use of the latest high-quality reference goat genome assembly provided us more details of the genomic information compared to CHIR_1.0 and CHIR_2.0. To avoid false positives in identifying SNPs and Indels in our study, a series of filtering step were applied to remove low-quality SNPs. This procedure guaranteed high quality genetic variants for the downstream analyses. The large number of genetic variants identified in this study gives us a chance to further explore the genetic diversity and genetic basis of different phenotypes in goats. The population-specific molecular markers can be used to distinguish phenotypically similar animals with higher accuracy.

Our study provided comprehensive insights into the phylogenetic relationship between the two major Chinese cashmere goat breeds. We showed that the Erlangshan cashmere goats were closest related to Liaoning cashmere goats. This genetic information may be useful to explore the domestication and distribution of cashmere goat in Northern China. Our results also provided a large collection of candidate genes that may be targeted for trait improvement. As part of the Hapmap goat project, these cashmere goat genetic footprints and SNPs will serve as a useful tool for the breeding of *Capra* species.

Methods

Sample collection. The Inner Mongolia cashmere goat breed was sampled from three independent populations according to their geographical locations: Alashan league, Ordos city, and BayanNur city of Inner Mongolia Autonomous Region. The Liaoning cashmere goat was sampled from one independent population in Gai county of Liaoning province (Fig. 1 and Supplementary Table 1). All four goat populations are raised in local pastures and allowed to free range. With the assistance of local herdsmen, trained veterinarians randomly chose 15–19 three-year-old female goats from each population, and collected 5 ml whole blood from the left jugular vein of each animal into plastic collection tubes containing 4% (w/v) sodium citrate. The blood samples were snap frozen in liquid nitrogen, and stored at -80°C until delivered to Kunming Institute of Zoology on dry ice for further processing. All experimental procedures were approved by the Animal Care and Use Committee of the Inner Mongolia Agricultural University, and conducted in strict accordance with the animal welfare and ethics guidelines.

DNA isolation and sequencing. Genomic DNA was extracted from the whole blood samples with the XYGEN Blood and Tissue Extraction Kit (Corning, USA) according to the manufacturer's instructions. The extracted DNA were subjected to electrophoresis in 2% agarose gel and stained with ethidium bromide to assess overall quality. The DNA concentration was determined by Quant-iT™ PicoGreen® dsDNA Reagent and Kits (Thermo Fisher Scientific, USA) according to the manufacturer's instructions. Paired-end libraries with insert size of 300 bp from $\sim 2\ \mu\text{g}$ of sheared genomic DNA were constructed with the procedures of NEB DNA Library Prep Kit for Illumina (NEB, USA). These libraries were sequenced on an Illumina HiSeq. 2000 platform (Illumina; CA, USA) using a PE-101 module. In addition, to characterize the genetic variant relate to cashmere fiber based on selection signals and GWAS, we also downloaded the genome data of 18 individuals from, including.

Data filtering and clean reads generation. All raw data were first filtered and trimmed using NGSQCToolkit_v2.3.3 if any of the following criteria were met: (1) reads containing adapter and poly-N; (2) low quality reads with $>30\%$ bases having Phred quality ≤ 25 ; (3) the 5' and 3' ends 5 bp low quality base, which is generally considered high bias. This data filtering process resulted in a total of 534.7 Gb clean data from 70 cashmere goats (51 Inner Mongolia breed and 19 Liaoning breed), achieving an average 2.61-fold individual genomic coverage depth. At the population level, the coverage ranged from 41.02 to 50.88 fold for genetic variation detection and downstream analysis (Supplementary Table 1).

Variation discovery. The clean reads were aligned to the recently released version of the reference goat genome (ARS1)^{23,47} using Burrows-Wheeler Aligner v0.7.10-r789⁴⁸ with default settings. The reference genome sequence was indexed with bwa. The algorithm MEM was used to find the suffix array coordinates of good matches for each read⁴⁹. SAMtools⁵⁰ was used to convert file format from SAM to BAM and to filter the unmapped and non-unique reads. Picard (version 1.106, <http://broadinstitute.github.io/picard/>) was used to sort the BAM files, and remove potential PCR duplication if multiple read pairs had identical external coordinates. Read pairs with top mapping quality were retained. Local realignment around short insertions and deletions (Indels) was performed with duplication-removed reads using RealignerTargetCreator and IndelRealigner in the Genome Analysis Toolkit (GATK, version 3.3-0-g37228af)⁵¹. After local realignment, 'HaplotypeCaller' in GATK was used for generating a single call set in all individuals by joint calling. Single nucleotide polymorphisms (SNPs) and Indels were separated with the GATK tool 'selectVariants', and subjected to rigorous processing to exclude false positives. The SNP exclusion criteria⁵² were as follows: (1) hard filtration with parameter 'QD < 2.0 || ReadPosRankSum < -8.0 || FS > 10.0 || QUAL < 1349.1'; (2) "-max-missing 0.7 ||-maf 0.05". The Indel exclusion criteria were as follows: (1) hard filtration with parameter 'QUAL < 20.0 || QD < 2.0 || ReadPosRankSum < -8.0 || FS > 10.0 || QUAL < 1257.74; (2) "-maxIndelSize 25 ||-maf 0.05", only insertions and deletions shorter than or equal to 25 bp indels were taken into account. Except Venn diagram, only mapped autosomal SNPs and Indels were included in the downstream analyses.

Genomic annotation of SNPs and Indels. All filtered SNPs and Indels were annotated and categorized by packages Annovar with default settings⁵³. Venn diagrams representing SNVs were generated using the online method (<http://bioinformatics.psb.ugent.be/webtools/Venn/>). The transition-to-transversion (Ti/Tv) ratio based on all detected SNPs was calculated to evaluate potential sequencing errors, which is used as an indicator of potential sequencing errors⁵². The average ratios of homozygous versus heterozygous and nucleotide diversity are calculated for Inner Mongolia and Liaoning cashmere goat with VCFTools⁵⁴.

Population structure analysis. To explore their phylogenetic relationship, the whole-genome autosomal SNPs were extracted to construct the phylogenetic tree, and genotypes of sheep sequence were used to provide out-group information at corresponding positions. The neighbor-joining tree was constructed using the PHYLIP 3.696 software (<http://evolution.genetics.washington.edu/phylip.html>) based on distance matrix methods⁵⁵. iTOL (<http://itol.embl.de>) was used to illuminate and visualize the phylogenetic tree⁵⁶.

For both of principal component (PCA) and population structure analysis, a thinned SNPs dataset with a window of size 50 SNPs advanced by 5 SNPs at a time and an linkage-disequilibrium r^2 threshold of 0.5 were filtered

using PLINK (Version v1.90b3.38)⁵⁷ PCA was performed with the Genome-wide Complex Trait Analysis (GCTA, version: 1.25.3) software⁵⁸, and the first three eigenvectors (two eigenvectors for PCA analysis of Liaoning and Inner Mongolia) were plotted. Population structure was analyzed using the ADMIXTURE (Version: 1.3)⁵⁹ program which implement a block-relaxation algorithm. To explore the convergence of individuals, we predefined the number of genetic clusters K from 2–6 and ran with cross-validation error (CV) procedure. Default methods and settings were used in Admixture analysis. Population differentiation index (*Fst*) was measured by pairwise *Fst* values among pairwise populations⁵⁴.

Linkage disequilibrium (LD) was calculated using PLINK software⁵⁷. The pairwise *r*² values within and between different chromosomes were calculated. Regarding the LD for overall genome, the *r*² value was calculated for individual chromosomes using SNPs from the corresponding chromosome with parameter “-ld-window-r2 0-ld-window 99999-ld-window-kb 1000-r2”, and then the pairwise *r*² values were averaged across the whole genome. The LD for each group was calculated using SNP pairs only from the corresponding group.

Genome scanning for selective signals. We performed a genetic differentiation (*Fst*) and polymorphism levels ($\theta\pi$, pairwise nucleotide variation as a measure of variability) based cross approaches to investigate the selection signals across the whole genome. A 100 kb sliding window with 10 kb step approach was applied to quantify *Fst* and $\theta\pi$, and the cross top 10% of two values was selected as selective signals.

Functional enrichment analysis (GO and KEGG). GO and KEGG enrichment analysis was performed using the OmicShare tools, a free online platform for data analysis (www.omicshare.com/tools). Firstly, all candidate genes were mapped to GO terms in the Gene Ontology database (<http://www.geneontology.org/>), gene numbers were calculated for every term, significantly enriched GO terms in genes comparing to the genome background were defined by hypergeometric test. The calculated p-value was gone through FDR Correction, taking $FDR \leq 0.05$ as a threshold. GO terms meeting this condition were defined as significantly enriched GO terms in candidate genes. Secondly, all candidate genes were mapped to KO terms in the KEGG Pathway database (<http://www.genome.jp/kegg/ko.html>). KEGG pathway enrichment analysis identified significantly enriched metabolic pathways or signal transduction pathways in candidate genes comparing with the whole genome background. The calculating formula and significantly enriched standard is the same as that in GO analysis.

Availability and Requirements. The sequencing reads of each sequencing libraries have been deposited under NCBI with Project ID SRP082615.

References

- Ryder, M. L. Coat structure and seasonal shedding in goats. *Animal Production* **8**, 289–302 (1966).
- Nixon, A. J., Gurnseyb, M. P., Betteridgec, K., Mitchellc, R. J. & Welchc, R. A. S. Seasonal hair follicle activity and fibre growth in some New Zealand Cashmere-bearing goats (*Caprus hircus*). *Journal of Zoology* **224**, 589–598 (1991).
- Geng, R. Q. Species-specific PCR for the identification of goat cashmere and sheep wool. *Molecular & Cellular Probes* **29**, 39–42 (2014).
- Watkins, P. & Buxton, A. Luxury fibres: rare materials for higher added value. *Special Report - Economist Intelligence Unit (United Kingdom)*, no. 2633 (1992).
- Waldron, S., Brown, C. & Komarek, A. M. The Chinese Cashmere Industry: A Global Value Chain Analysis. *Social Science Electronic Publishing* **32**, 589–610 (2014).
- Resources, C. N. C. O. A. G. *Animal genetic resources in China: Sheep and goats*. (Chinese Agricultural Press, 2011).
- Li, C. Q. *et al.* Comparative Study on Skin and Hair Follicles Cycling between Inner Mongolia and Liaoning Cashmere Goats. *Acta Veterinaria Et Zootechnica Sinica* **36**, 674–679 (2005).
- Zhou, J. P. *et al.* A novel single-nucleotide polymorphism in the 5' upstream region of the prolactin receptor gene is associated with fiber traits in Liaoning cashmere goats. *Genetics & Molecular Research Gmr* **10**, 2511–2516 (2011).
- Shamsalddini, S., Mohammadabadi, M. R. & Esmailizadeh, A. K. Polymorphism of the prolactin gene and its effect on fiber traits in goat. *Russian Journal of Genetics* **52**, 405–408 (2016).
- Wang, X. *et al.* Disruption of FGF5 in Cashmere Goats Using CRISPR/Cas9 Results in More Secondary Hair Follicles and Longer Fibers. *Plos One* **11**, e0164640 (2016).
- Ye, G. *et al.* Comparative Transcriptome Analysis of Fetal Skin Reveals Key Genes Related to Hair Follicle Morphogenesis in Cashmere Goats. *PLoS One* **11**, e0151118 (2016).
- Geng, R., Chao, Y. & Chen, Y. Exploring Differentially Expressed Genes by RNA-Seq in Cashmere Goat (*Capra hircus*) Skin during Hair Follicle Development and Cycling. *PLoS One* **8**, e62704 (2013).
- Chunhui *et al.* Effects of melatonin implantation on cashmere yield, fibre characteristics, duration of cashmere growth as well as growth and reproductive performance of Inner Mongolian cashmere goats. *Journal of Animal Science and Biotechnology* **6**, 1–6 (2015).
- Day-Williams, A. G. & Zeggini, E. The effect of next-generation sequencing technology on complex trait research. *Eur J Clin Invest* **41**, 561–567 (2011).
- Rosenthal, E., Blue, E. & Jarvik, G. P. Next-generation gene discovery for variants of large impact on lipid traits. *Curr Opin Lipidol* **26**, 114–119 (2015).
- Groenen, M. A. A decade of pig genome sequencing: a window on pig domestication and evolution. *Genetics Selection Evolution* **48**, 23 (2016).
- Li, M. *et al.* Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. *Scientific Reports* **4** (2014).
- Ai, H. *et al.* Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics* **47**, 217 (2015).
- Ji, Y. *et al.* Whole-Genome Sequencing of Native Sheep Provides Insights into Rapid Adaptations to Extreme Environments. *Molecular Biology & Evolution* **33**, 2576 (2016).
- Gou, X. *et al.* Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome Research* **24**, 1308 (2014).
- Axelsson, E. *et al.* The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**, 360–364 (2013).
- Ly, F. H. *et al.* Adaptations to Climate-Mediated Selective Pressures in Sheep. *Molecular Biology & Evolution* **31**, 3324 (2014).
- Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nature Biotechnology* **31**, 135–141 (2013).

24. Wang, X. *et al.* Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits. *Sci Rep* **6** (2016).
25. Benjelloun, B. *et al.* Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (*Capra hircus*) using WGS data. *Frontiers in Genetics* **6**, 107 (2015).
26. Guan, D. *et al.* Scanning of selection signature provides a glimpse into important economic traits in goats (*Capra hircus*). *Scientific Reports* **6**, 36372 (2016).
27. Guan, D. *et al.* Scanning of selection signature provides a glimpse into important economic traits in goats (*Capra hircus*). *Scientific Reports* **6** (2016).
28. Nachman, M. & Crowell, S. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297 (2000).
29. Zhan, B. *et al.* Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. (*BMC Genomics*). **12**, 557, <https://doi.org/10.1186/1471-2164-12-557> (2011).
30. Lock, F. E. & Hotchin, N. A. Distinct roles for ROCK1 and ROCK2 in the regulation of keratinocyte differentiation. *PLoS One* **4**, 0008190 (2009).
31. Kalaji, R. *et al.* ROCK1 and ROCK2 regulate epithelial polarisation and geometric cell shape. *Biol Cell* **104**, 435–451 (2012).
32. Housley, D. J. & Venta, P. J. The long and the short of it: evidence that FGF5 is a major determinant of canine 'hair'-itability. *Animal Genetics* **37**, 309–315 (2006).
33. Cadieu, E. *et al.* Coat Variation in the Domestic Dog Is Governed by Variants in Three Genes. *Science* **326**, 150–153 (2009).
34. Legrand, R., Turet, L. & Abitbol, M. Two recessive mutations in FGF5 are associated with the long-hair phenotype in donkeys. *Genetics Selection Evolution* **46**, 1–7 (2014).
35. Hebert, J. M., Rosenquist, T., Gotz, J. & Martin, G. R. FGF5 as a regulator of the hair growth cycle: evidence from targeted and spontaneous mutations. *Cell* **78**, 1017–1025 (1994).
36. Li, L. F., Fiedler, V. C. & Kumar, R. The potential role of skin protein kinase C isoforms alpha and delta in mouse hair growth induced by diphenylprone-allergic contact dermatitis. *Journal of Dermatology* **26**, 98–105 (1999).
37. McCormick, J. A. *et al.* Targeted disruption of the protein kinase SGK3/CISK impairs postnatal hair follicle development. *Mol Biol Cell* **15**, 4278–4288 (2004).
38. Alonso, L. *et al.* Sgk3 links growth factor signaling to maintenance of progenitor cells in the hair follicle. *J Cell Biol* **170**, 559–570 (2005).
39. Okada, T. *et al.* The critical roles of serum/glucocorticoid-regulated kinase 3 (SGK3) in the hair follicle morphogenesis and homeostasis: the allelic difference provides novel insights into hair follicle biology. *Am J Pathol* **168**, 1119–1133 (2006).
40. Mauro, T. M. *et al.* Akt2 and SGK3 are both determinants of postnatal hair follicle development. *Faseb J* **23**, 3193–3202 (2009).
41. Campagna, D. R., Custodio, A. O., Antiochos, B. B., Cirlan, M. V. & Fleming, M. D. *Mutations in the serum/glucocorticoid regulated kinase 3 (Sgk3) are responsible for the mouse fuzzy (fz) hair phenotype.* (*J Invest Dermatol*. 2008 Mar;128(3):730–2. Epub 2007 Oct 4).
42. Deing, V. *et al.* Oxytocin modulates proliferation and stress responses of human skin cells: implications for atopic dermatitis. *Experimental Dermatology* **22**, 399 (2013).
43. Gazel, A. *et al.* Transcriptional Profiling of Epidermal Keratinocytes: Comparison of Genes Expressed in Skin, Cultured Keratinocytes, and Reconstituted Epidermis, Using Large DNA Microarrays. *Journal of Investigative Dermatology* **121**, 1459–1468 (2003).
44. Lv, F. H. *et al.* Adaptations to climate-mediated selective pressures in sheep. *Mol Biol Evol* **31**, 3324–3343 (2014).
45. Strait, K. A., Stricklett, P. M. & Kohan, D. E. Characterization of vasopressin-responsive collecting duct adenylyl cyclases in the mouse. *American Journal of Physiology* **298**, F859–867 (2010).
46. Al-Hakim, A., Rui, X., Tsao, J., Albert, P. R. & Schimmer, B. P. Forskolin-resistant Y1 adrenal cell mutants are deficient in adenylyl cyclase type 4. *Mol Cell Endocrinol* **214**, 155–165 (2004).
47. Du, X. *et al.* An update of the goat genome assembly using dense radiation hybrid maps allows detailed analysis of evolutionary rearrangements in Bovidae. *BMC Genomics* **15**, 1–16 (2014).
48. Wu, Y. P. *et al.* A fine map for maternal lineage analysis by mitochondrial hypervariable region in 12 Chinese goat breeds. *Animal Science Journal* **80**, 372–380 (2009).
49. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Quantitative Biology* **1303** (2013).
50. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352> (2009).
51. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303, <https://doi.org/10.1101/gr.107524.110> (2010).
52. Choi, J. W. *et al.* Whole-Genome Resequencing Analysis of Hanwoo and Yanbian Cattle to Identify Genome-Wide SNPs and Signatures of Selection. *Mol Cells* **38**, 466–473 (2015).
53. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
54. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
55. Plotree, D. & Plotgram, D. PHYLP-phylogeny inference package (version 3.2). *cladistics* **5**, 163–166 (1989).
56. Letunic, I. & Bork, P. Interactive tree of life (iTOL)v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44**, gkw290 (2016).
57. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).
58. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).
59. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–1664 (2009).
60. Team, R. D. C. R: A Language and Environment for Statistical Computing. *Computing* **1**, 12–21 (2013).
61. Kahle, D. & Wickham, H. ggmap: Spatial Visualization withggplot2. *R Journal* **5**, 144–161 (2016).

Acknowledgements

This work was financially supported by National Natural Science Foundation of China (31660639), National High Technology Research and Development Program of China(863plan) (2013AA102506), National Natural Science Foundation of China (31660640), and National Natural Science Foundation of China (31560619).

Author Contributions

W.W., Y.D., and J.L. designed the study. W.Z., H.J., Y.Z., R.W., Z.L., Z.W., B.L., Y.M. and Z.H. did samples collection. X.L., X.Q., Y.F., W.W., and Q.Z. did the sequencing experiment. R.S., R.D., T.Z., Y.J., and W.C. analyzed the data and wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-017-15516-0>.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017