Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme environments

Ji Yang,[†,1] Wen-Rong Li,[†,2] Feng-Hua Lv,[†,1] San-Gang He,[†,2] Shi-Lin Tian,[†,3] Wei-Feng Peng,[1,4] Ya-Wei Sun,[2,5] Yong-Xin Zhao,[1,4] Xiao-Long Tu,[3] Min Zhang,[1,6] Xing-Long Xie,[1,4] Yu-Tao Wang,[7] Jin-Quan Li,[8] Yong-Gang Liu,[9] Zhi-Qiang Shen,[10] Feng Wang,[11] Guang-Jian Liu,[3] Hong-Feng Lu,[3] Juha Kantanen,[12,13] Jian-Lin Han,[14,15] Meng-Hua Li,*[,1] and Ming-Jun Liu*[,2]

Supplementary Information

Table of Contents

Supplementary Results and Discussion

Supplementary Figures and Figure Legends

Supplementary Materials and Methods

Ethics statement

The methods were carried out in accordance with the approved guidelines of the Good Experimental Practices adopted by the Institute of Zoology, Chinese Academy of Sciences (CAS). All experimental procedures and animal collection were conducted under a permit (No. IOZ13015) approved by the Committee for Animal Experiments of the Institute of Zoology, Chinese Academy of Sciences, and conformed to the China Wildlife Protection Law (CWPL).

Sample information

A total of 80 animals, representing 21 Chinese native sheep (*Ovis aries*) breeds (77 ewes) and three wild species (one animal from each species) of the Subfamily *Caprinae*, viz *Ovis aries musimon*, *Ovis ammon polii* and *Capra ibex*, were included in the whole-genome sequencing analysis. The wild species were used as outgroups in the sheep phylogenetic tree and materials for comparison of genetic diversity, SNP sharing and demographic changes (i.e., PSMC plots) with native sheep. Among the native sheep, several breeds have inhabited extreme environments for thousands of years (Du 2011). Specifically, 16 Tibetan sheep from areas of Tibet on the Qinghai-Tibetan Plateau (four animals from each of the four locations: Nagqu County [ZNQ], Qamdo County [ZCD], Shigatse [ZRK] and Nyingchi [ZLZ]) are from a plateau environment (> 4,000 m, except for ZLZ, at *c*. 2,900 m), and nine animals from breeds in the Taklimakan Desert region (four animals from Lop sheep [LOP] and

five animals from Baerchuke sheep [BRK]) are from a desert environment (average annual precipitation < 10 mm; fig. 1A and B, supplementary table S1, Supplementary Material online). Furthermore, five animals representing Hu sheep (HUS) and five animals representing Wadi sheep (WDS) from non-extreme environments in Eastern China (altitude < 100 m, average annual precipitation > 600 mm; fig. 1A and B, supplementary table S1, Supplementary Material online) were sequenced for genomic comparison. In addition, 20 animals from breeds on the Yunnan-Kweichow Plateau (altitude = 1,700 – 3,300 m; five animals from each of the four breeds: Tengchong sheep [TCS], Diqing sheep [DQS], Shiping Gray sheep [SPS] and Weining sheep [WNS]), four animals of Ganzi sheep [GZS] from the southeastern margin of the Qinghai-Tibetan Plateau, five animals of Minxian Black Fur sheep (MXS) and four animals of Guide Black Fur sheep (GDS) from Northwestern China, five animals representing sheep breeds from Xinjiang (one animal from each of the five breeds: Altay sheep [ALS], Bayinbuluke sheep [BYK], Kazakh sheep [KAZ], Hetian sheep [HTS] and Tashkurgan sheep [TSK]) and four animals representing sheep breeds from Inner Mongolia (one animal from each of the four breeds: Ujimqin sheep [WZS], Wuranke sheep [WRS], Sunite sheep [SNS] and Hulun Buir sheep [HLS]; fig. 1A, supplementary table S1, Supplementary Material online) were sequenced to cover all of the main biogeographic groups of Chinese native sheep (Zhong et al. 2010; Lv et al. 2015).

At a larger geographic scale, on the basis of the altitudinal differences among the

sampling locations, the geographic origins of the 77 native sheep can be roughly categorized as high-altitude regions (> 1,500 m; defined as one type of extreme environment; including ZNQ, ZCD, ZRK, ZLZ, GDS, GZS, MXS, TCS, WNS, SPS, DQS, TSK and HTS) and low-altitude regions (< 1,300 m; defined as the counterpart; comprising HUS, WDS, LOP, BRK, ALS, BYK, KAZ, WZS, WRS, SNS and HLS; fig. 1A). Likewise, the geographic origins of the breeds can be divided into arid and humid zones according to the 400 mm average annual precipitation line (Piao et al. 2010) in China. The arid zones (average annual precipitation < 400 mm, representing arid and semi-arid regions; Piao et al. 2010; defined as one type of extreme environment) provide homes for LOP, BRK, ALS, BYK, KAZ, TSK, HTS, ZNQ, ZRK, GDS, WZS, WRS, SNS and HLS, and the humid zones (average annual precipitation > 400 mm, representing humid and semi-humid regions; Piao et al. 2010; defined as the counterpart) accommodate HUS, WDS, ZCD, ZLZ, GZS, MXS, TCS, WNS, SPS and DQS (fig. 1B).

Genome sequencing

The overall analysis pipeline is detailed in supplementary fig. S14, Supplementary Material online. In summary, we sequenced the whole genomes of 75 native sheep and three wild animals at an average depth of ~5×, using a sequencing strategy similar to that applied in the 1000 Genomes Project (see http://www.1000genomes.org; The 1000 Genomes Project Consortium 2010). In addition, we sequenced two animals (one Tibetan sheep, ZNQ24, and one Lop sheep, LOP41) at a high depth of ~42× to

facilitate SNP calling and identify breed-specific structural variants in breeds from extreme environments. This strategy enabled us to obtain more information regarding genetic variability, and we were also able to obtain reliable genotype calls through population-based SNP calling.

Genomic DNA was extracted from ear tissue using the standard phenol-chloroform protocol (Sambrook and Russell 2000). High-quality DNA for genome sequencing was processed to construct short-insert (500 bp) DNA libraries according to the manufacturer's specifications (Illumina, San Diego, CA). To generate 500-bp mate-paired libraries, we used the Covaris Ultrasonic Processor (Covaris, Woburn, MA) to cut genomic DNA into 500-bp fragments randomly, followed by the process of end repairing, adding A to the tails, purification and PCR amplification. The Qubit v.2.0 kit (Life Technologies, Gaithersburg, MD) was used to analyze the quality of the constructed libraries. After diluting each library to 1 ng/μl, an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA) was used to check the insert size of the libraries, and real time q-PCR was then performed to detect the effective concentration of the libraries. The qualified libraries with appropriate insert size (500 bp) and concentration (> 2 nM) were sequenced using the Illumina HiSeq 2000 platform (Illumina, San Diego, CA), and 100-bp paired-end reads were generated and managed using Illumina HiSeq Control Software (HCS) v3.3 (Illumina, San Diego, CA). Overall, we produced approximately 13.56 billion raw reads (totaling 1356.75 Gb (gigabases) of raw data) from the 80 samples (supplementary table S30,

Supplementary Material online). The information of raw reads and high-quality data for each sample are summarized in supplementary table S30, Supplementary Material online.

Owing to base-calling duplicates and adapter contamination in each lane, the Illumina Pipeline may generate low-quality reads (Li et al. 2014). To obtain reliable reads and avoid reads with artificial biases that might affect downstream mapping and other analyses, we implemented quality control procedures to exclude the following types of reads:

i) Unidentified nucleotides (N-content) ≥ 10%;

ii) More than 50% of the read bases with a Phred quality score (Q-score) less than 5;

iii) More than 10 nucleotides overlapped with the adapter sequence (≤ 10% mismatches allowed);

iv) Duplicate reads which were generated by PCR amplification during DNA library construction (i.e., read 1 and read 2 of two paired-end reads that were completely identical). PCR duplicates were removed by using the command 'rmdup' in SAMtools v0.1.19 (Li et al. 2009).

Read alignment

After raw reads filtering, the high-quality reads from each sample were aligned to the sheep reference genome assembly Oar_v3.1.75 using the Burrows-Wheeler Aligner (BWA) tool (Li and Durbin 2009). First, the reference genome sequence was indexed

with the command ‘index’. The command ‘mem -t 10 -k 32’ was then used to identify the suffix array (SA) coordinates of good hits for each individual read. Additionally, the SA coordinates were converted into the best alignments in BAM format using SAMtools v0.1.19 (Li et al. 2009). Other parameters in the BWA tool were set to default values.

To obtain high-quality alignments for SNP calling and subsequent analyses, we conducted the steps below:

i) Filter the alignment read with mismatches ≥ 5 and mapping quality = 0;

ii) Remove putative duplicated alignments. When two or multiple read-pairs were aligned to identical positions on the reference genome, we retained only the pair with the highest mapping quality score;

iii) Extract properly paired mapped reads with the command ‘samtools view -f 0x2’.

Variant calling

The highest accuracy alignments for each animal were processed using a Bayesian approach as implemented in SAMtools v0.1.19 (Li et al. 2009). The command ‘mpileup’ was used to perform SNPs and indels (length < 100 bp) calling, with the parameters ‘-C -D -S -m 2 -F 0.002 -d 1000’. The raw SNPs were filtered in the downstream analysis by requiring a minimum coverage depth of 5 and a maximum of 100, a minimum RMS (root mean square) mapping quality score of 20, and no gap present within a 3-bp window. We then merged the SNPs from 77 native sheep into a

population SNP-matrix and subsequently identified population SNPs with a strict criteria. We filtered the SNPs with MAF (minor allele frequency) < 0.05 and missing genotype > 10% of animals in sheep population. The high-quality SNPs identified in native sheep were categorized according to their genomic locations, including exons, introns, UTRs and intergenic regions, and the SNPs located in exons were further divided into synonymous and nonsynonymous SNPs. The high-quality SNPs obtained here were subsequently regarded as 'called high-quality SNPs' and used in the SNP summary, regions of homozygosity, linkage disequilibrium, selective sweep, gene ontology and target gene analyses. In the two samples sequenced at a high read depth (ZNQ24 and LOP41, each at ~42×), we detected structural variations (SVs, 100 bp–chromosome level; Alkan et al. 2011), including large fragment insertions (INS) and deletions (DEL), inversions (INV), intra-chromosomal translocations (ITX) and inter-chromosomal translocations (CTX), using the package BreakDancer v1.1 (Chen et al. 2009) with the default parameters (e.g., -y 40, SV score greater than 40; -m 1000000000, maximum SV size of 1000000000), although the -q option (i.e., the mapping quality threshold) was set to 20 and the -r option was set to 2 (i.e., SVs were supported by at least two paired-end reads). We also identified copy number variations (CNVs, 200 bp–5 Mb; Freeman et al. 2006) using the software CNVnator v0.3 (Abyzov et al. 2011) with 100-bp bins and standard parameters. All of the genomic variations were annotated using ANNOVAR software (Wang et al. 2010).

Validation of the called SNPs

To examine the reliability of the called SNPs, we first compared the SNPs identified here with those from Build 143 of the sheep dbSNP database in the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/SNP, last accessed November 12, 2015). We then compared the genotypes of the called SNPs with those on the Ovine 50K BeadChip array (Illumina, San Diego, CA) for 33 native sheep with available chip data. The chip-based SNPs were filtered using the same criteria as those adopted in this study, and the genomic locations of the SNPs were adjusted according to the sheep reference genome assembly Oar_v3.1.75.

Analysis of regions of homozygosity

Based on the 21.26 million called high-quality SNPs in the native sheep, we measured the regions of homozygosity (ROH) for each breed/population using the 'runs of homozygosity' function in the program PLINK v1.07 (Purcell et al. 2007), with the following command '--homozyg-window-kb 5000 --sheep --homozyg-window-snp 50 --homozyg-window-het 1 --homozyg-snp 10 --homozyg-kb 100 --homozyg-density 10 --homozyg-gap 100'.

Inbreeding and linkage disequilibrium analyses

Because LD patterns can be affected by inbreeding, we estimated the value of identity-by-state (IBS) between all samples to identify the genetic relatedness among sheep individuals, using the 21.26 million called high-quality SNPs and the command '--file filename --cluster --matrix' in the program PLINK v1.07 (Purcell et al. 2007).

Based on the non-inbred animals (IBS < 0.9), we then compared the patterns of linkage disequilibrium (LD) among different sheep breeds/populations. The squared correlation coefficient (r^2; Hill and Robertson 1968) between pairwise SNPs was calculated to estimate the decay of LD using the software Haploview v4.2 (Barrett et al. 2005). The parameters in the program were set as: '-n -dprime -minGeno 0 -missingCutoff 1 -minMAF 0.01'. The average r^2 value was measured in a 500-kb window size. We found differences in the rate of decay and the level of LD value, which reflected the variations in population demographic history and effective population size (N_e) among the breeds/populations.

Population genetics analysis

To avoid statistical bias from low-coverage data, our population genetics inference was based on genotype likelihoods (GL) which can take genotype uncertainty into account. Following reads mapping, high-quality alignments (BAM files) were input to the program ANGSD v.0.902 (Analysis of Next Generation Sequencing Data; Korneliussen et al. 2014). An empirical Bayesian approach was then used to compute GL, major/minor states and SNP calls (Kim et al. 2011), with the SAMtools model (Li 2011) and a SNP *p*-value threshold of 0.01. The ANGSD command was './angsd -bam bam.list -GL 1 -doMaf 1 -doGlf 4 -SNP_pval 0.01 -doMajorMinor 1 -nInd 80 -doGeno 5 -dopost 1 -out sheep'. This procedure produced 32.06 million GL-associated high-quality SNPs (coverage depth ≥ 5 and ≤ 100, RMS mapping quality ≥ 20, MAF ≥ 0.05 and the missing ratio of samples within population < 10%)

for the subsequent analyses of genetic diversity, population structure and demographic history.

Based on the high-quality SNPs identified in the 77 native sheep from ANGSD, we calculated the genomic variation within each breed (or each population within Tibetan sheep) and within each group of breeds using pairwise nucleotide diversity (θ_π, Tajima 1983). The statistic was calculated by a sliding-window approach, with a window width of 100-kb and a stepwise distance of 50-kb. We used the *p*-distance, which represents the proportion of different amino acid sites between two sequences, to infer the phylogeny. The *p*-distance was calculated for pairwise comparisons across all 80 samples. The distance matrix was subsequently used to construct a phylogenetic tree for the 77 native sheep with the neighbor-joining (NJ) method (Saitou and Nei 1987) as implemented in the software TreeBeST v1.9.2 (http://treesoft.sourceforge.net/treebest.shtml, last accessed August 6, 2013), and the tree was rooted with either one wild species (*O. a. musimon*) or three wild species (*O. a. musimon*, *O. a. polii* and *C. ibex*). We also conducted an individual-scale principal component analysis (PCA) for the 77 naïve sheep with the smartpca program in the package EIGENSOFT v5.0 (Patterson et al. 2006).

To further investigate the genetic relationships among sheep breeds, we performed a genetic structure analysis based on the GL-associated high-quality SNPs. Highly linked SNPs ($r^2 > 0.5$) were excluded from the analysis using the command

‘‘indep-pairwise 50 5 0.5’’ in PLINK v1.07. The list of pruned GL-associated SNPs was used to infer the genetic structure of Chinese native sheep using the program FRAPPE v1.170 (Tang et al. 2005), which employs the maximum likelihood and expectation-maximization algorithm (Tang et al. 2005) to estimate ancestry proportions for each individual. The number of assumed genetic clusters (*K*) ranged from 2 to 9, with 10,000 iterations for each run. The graph outputs of the population structure analysis were displayed by DISTRUCT v1.1 (Rosenberg 2004). The population genetic differentiation between the three identified groups of breeds (Qinghai-Tibetan, Yunnan-Kweichow and Northern and Eastern Chinese breeds; see Results) was measured by pairwise F_{ST} value (Weir and Cockerham 1984).

Demographic history and population admixture analyses

We used the pairwise sequentially Markovian coalescent (PSMC; Li and Durbin 2011) method to estimate the changes in effective population size (N_e) of sheep and three wild species over the last one million years. This method uses a hidden Markov model to reconstruct the history of N_e based on the SNP distribution in an individual diploid genome (Li and Durbin 2011). Two representative samples with a high sequencing depth (ZNQ24 and LOP41; ~42×) and the other samples with relatively low coverage (*circa* 5×) were analyzed separately. First, the SNP genotype of each animal was called using the package SAMtools v0.1.19 (Li et al. 2009), with the command ‘mpileup’ and the parameter ‘-C 50 -D -S -m 2 -F 0.002’. Next, the program ‘fq2psmcfa’ in SAMtools was employed to transform the consensus sequences into a

FASTA-like format in which the i-th character in the output sequence indicates whether there is at least one heterozygote in the bin [100i, 100i+100]. The parameters were set as follows: -N30 -t15 -r5 -p '4+25*2+4+6'. Time was measured in units of $2N_0$ generations, and the N_e at time t was scaled to N_0. Because of the absence of reported estimates of the mutation rate per nucleotide in sheep, we used the average mutation rate in human (μ) of 2.5×10^{-8} per base per generation. The generation time (g) of the sheep was set to two years, which approximates the mean time required for male and female sheep to reach sexual maturity (de Magalhães and Costa 2009). To assess the consistency of the estimates, we performed 100 bootstrap replications for each sample. In addition, we retrieved atmospheric surface air temperature (°C) and global relative sea level data for the past one million years from the NCDC (http://www.ncdc.noaa.gov, last accessed October 27, 2013) and included them in the PSMC output graphs.

As the PSMC approach does not have sufficient power to reconstruct demographic events within 10,000 years due to the limited recombination events in this short time period (Li and Durbin 2011), we also used the diffusion approximation for demographic inference (*∂a∂i*; Gutenkunst et al. 2009) approach to infer the recent demographic history (e.g., < 4,000 years) of the three identified sheep groups (Qinghai-Tibetan, Yunnan-Kweichow and Northern and Eastern Chinese breeds; see Results). This method employs the site frequency spectrum (SFS) of SNP data for populations (Gutenkunst et al. 2009) rather than recombination events within the

individual genome as in the PSMC approach (Li and Durbin 2011). Briefly, *∂a∂i* computed a SFS for each of the tested demographic scenarios and then sought the maximum similarity between the expected SFS of one tested scenario and the observed SFS over the parameter values space. The software employed a composite-likelihood ratio test to evaluate model fitting and to optimize model selection. The model with the highest likelihood value was identified as the optimal one.

The SFS of all 80 samples and the 77 native sheep were estimated using a two-step procedure implemented in ANGSD (Nielsen et al. 2012). First, sample allele-frequency likelihood files (.saf) were generated using the option '-doSaf 2', with the ancestral state assigned according to the sheep reference genome assembly Oar_v3.1.75. The command was './angsd -bam bam.list -GL 1 -dosaf 2 -out sheep -anc Ovis_aries.Oar_v3.1.dna.toplevel.fa -doMaf 1 –doMajorMinor'. Next, the allele-frequency likelihood files were optimized with the realSFS program to estimate the SFS. Genotypes were called using the full set of GL data. Then the posterior probabilities of the genotypes were computed at each site for each animal using the sample allele frequency as a prior. The command was './realSFS sheep.saf.idx > sheep.sfs'. To improve the accuracy of the genotypes and infer missing genotypes, we used the program BEAGLE (Browning and Browning 2007) to infer the haplotypes of the samples. After investigating the empirical distributions of the MAF, the haplotypes were inferred for all genotype sites with MAF > 0.01. Only sites showing

a correlation value between the imputed and observed data (r^2) greater than 0.9 were retained. To determine the potential bias of imputation filtering, the SFS obtained before and after filtering were compared.

We tested five possible divergence models (supplementary table S18, Supplementary Material online):

(i) Model 1: The ancestor simultaneously evolved into Northern and Eastern Chinese sheep breeds, Yunnan-Kweichow sheep breeds and Qinghai-Tibetan sheep breeds;

(ii) Model 2: The ancestor first evolved into Yunnan-Kweichow sheep breeds and non-Yunnan-Kweichow sheep breeds, and non-Yunnan-Kweichow sheep breeds were then split into Northern and Eastern Chinese sheep breeds and Qinghai-Tibetan sheep breeds;

(iii) Model 3: The ancestor first evolved into Northern and Eastern Chinese sheep breeds and non-Northern and Eastern Chinese sheep breeds, and non-Northern and Eastern Chinese sheep breeds were then split into Yunnan-Kweichow sheep breeds and Qinghai-Tibetan sheep breeds;

(iv) Model 4: The ancestor first evolved into Qinghai-Tibetan sheep breeds and non-Qinghai-Tibetan sheep breeds, and non-Qinghai-Tibetan sheep breeds were then split into Northern and Eastern Chinese sheep breeds and Yunnan-Kweichow sheep breeds;

(v) Model 5: The ancestor first evolved into Northern and Eastern Chinese sheep

breeds and Yunnan-Kweichow sheep breeds, and these two groups of breeds were then admixed to form Qinghai-Tibetan sheep breeds.

To avoid the influence of SNPs under selection on demographic inferences, we used only SNPs within the intergenic regions of all autosomes. As suggested in the *∂a∂i* approach (Gutenkunst et al. 2009), we started with simple models and increased the model complexity gradually by adding more parameters. We used a strategy in which a newly added parameter was discarded if it did not bring a marked improvement to the likelihood. Gene flow was modeled as discrete migration events at a certain time after population divergence. The unfolded frequency spectrum was used to avoid biases, because there was no trinucleotide substitution used for statistical correction. After model selection, scaled parameters for the best-supported model were transformed into the real values using the same μ and g as described above for the PSMC analysis.

A population-level admixture analysis was carried out in the TreeMix v.1.12 program (Pickrell and Pritchard 2012). This program uses genome-wide SNP sites to infer the Maximum Likelihood (ML) tree for the 21 native sheep breeds (77 animals) and an outgroup (*O. a. musimon*), with the command '-i input -bootstrap -k 10000 -root outgroup -o output'. The covariance matrix of the ML tree was estimated to reflect the correspondence between the ML tree and the SNP data and to identify pairs of populations that showed poor fits in the ML tree (Pickrell and Pritchard 2012). When

the ML tree did not fully describe the data, an admixture analysis was performed by allowing migration events, with the poor-fit populations regarded as candidates for adding potential migration edges (Pickrell and Pritchard 2012). Here, from one to 16 migration events were gradually added to the ML tree of the 21 native breeds until 98% of the variance between the breeds could be explained by the model. The command was '-i input -bootstrap -k 10000 -m migration events -o output'.

Genome-wide selective sweep test

We defined four pairs of groups of populations (i.e., four extreme-control group pairs; supplementary table S2, Supplementary Material online) for the selection tests:

(i) The Tibetan sheep from the plateau environment (encoded as the Tibetan group, including the Nagqu County [ZNQ], Qamdo County [ZCD] and Shigatse [ZRK] populations; 12 animals) *vs*. the Control group from East China (including Hu sheep (HUS) and Wadi sheep (WDS); 10 animals);

(ii) Sheep breeds from the Taklimakan Desert region from the desert environment (encoded as the Taklimakan Desert group, including Lop sheep [LOP] and Baerchuke sheep [BRK]; 9 animals) *vs*. the Control group from East China (10 animals);

(iii) Sheep breeds from low-altitude areas (encoded as the Low-altitude group, including HUS, WDS, LOP, BRK, ALS, BYK, KAZ, WZS, WRS, SNS and HLS; 26 animals) *vs*. Sheep breeds from high-altitude areas (encoded as the High-altitude group, including ZNQ, ZCD, ZRK, ZLZ, GDS, GZS, MXS, TCS,

WNS, SPS, DQS, TSK and HTS; 51 animals);

(iv) Sheep breeds from the humid zone (encoded as the Humid group, including HUS, WDS, ZCD, ZLZ, GZS, MXS, TCS, WNS, SPS and DQS; 47 animals) *vs*. Sheep breeds from the arid zone (encoded as the Arid group, including LOP, BRK, ALS, BYK, KAZ, TSK, HTS, ZNQ, ZRK, GDS, WZS, WRS, SNS and HLS; 30 animals).

The Nyingchi (ZLZ, 2900 m) population was excluded from the extreme-control selection test because the elevation of its distribution region is much lower than that of other populations of Tibetan sheep (4100– 4700 m; supplementary table S1, Supplementary Material online).

We calculated the genome-wide distribution of F_{ST} values (Weir and Cockerham 1984) and θ_π ratios (Nei and Li 1979; i.e., $\theta_{\pi\text{-Control}}/\theta_{\pi\text{-Tibetan}}$, $\theta_{\pi\text{-Control}}/\theta_{\pi\text{-Taklimakan Desert}}$, $\theta_{\pi\text{-Low-altitude}}/\theta_{\pi\text{-High-altitude}}$ and $\theta_{\pi\text{-Humid}}/\theta_{\pi\text{-Arid}}$) for the four defined group pairs using a sliding-window approach (100-kb windows with 50-kb increments). The F_{ST} values were Z-transformed as follows: $Z(F_{ST}) = (F_{ST} - \mu F_{ST}) / \sigma F_{ST}$, in which μF_{ST} is the mean F_{ST} and σF_{ST} is the standard deviation of F_{ST}. The θ_π ratios were $\log_2$-transformed. Subsequently, we estimated and ranked the empirical percentiles of $Z(F_{ST})$ and $\log_2(\theta_\pi$ ratio) in each window. The windows that simultaneously showed significantly high $Z(F_{ST})$ (the top 5% level for empirical percentile, $Z(F_{ST}) > 1.834$, 1.859, 1.731 and 1.823 for the Control group/Tibetan group, Control group/Taklimakan Desert group, Low-altitude group/High-altitude group and Humid

group/Arid group, respectively) and $\log_2(\theta_\pi$ ratio) (the top 5% level for empirical percentile, $\log_2(\theta_\pi$ ratio) > 0.352, 0.277, 0.247 and 0.198 for the Control group/Tibetan group, Control group/Taklimakan Desert group, Low-altitude group/High-altitude group and Humid group/Arid group, respectively) values were considered as candidate outliers in strong selective sweeps. All outlier regions were assigned to corresponding SNPs and annotated genes. Furthermore, we estimated the cross-population extended haplotype homozygosity (XP-EHH; Sabeti et al. 2007) statistic for the Tibetan group and the Taklimakan Desert group, using the Control group as a reference. The genetic map was assumed to be 1 cM/Mb for the sheep genome (Kijas et al. 2012). Also, we used the program LFMM (Frichot et al. 2013) to calculate the correlations between the genetic variants of native sheep and climate variables (i.e., altitude and precipitation). The z scores for all variants were calculated using a burn-in of 100, 1,000 sweeps, and $K = 3$ latent factors on the basis of the results from the population structure analysis. The threshold for identifying candidate genes in the XP-EHH and LFMM analyses was set to the top 5% and top 1% percentile outliers, respectively.

Candidate gene analysis

Using the 77 native sheep, we tested whether the selective signals detected from the extremely contrasted group pairs (e.g., Control group *vs*. Tibetan group, 22 animals; Control group *vs*. Taklimakan Desert group, 19 animals) were robust in a larger panel of samples (i.e., 77 animals). Firstly, NJ trees were constructed by TreeBeST v1.9.2

with the SNPs located in the candidate genes of Tibetan sheep and breeds from the Taklimakan Desert region. Also, we performed an analysis of the large-effect SNPs (i.e., nonsynonymous SNPs in candidate genes) for the candidate genes of Tibetan sheep and breeds from the Taklimakan Desert region. When the candidate genes were specific to Tibetan sheep, absolute values of the frequency differences for the large-effect SNPs were calculated between pairwise groups of Qinghai-Tibetan breeds, Yunnan-Kweichow breeds and Northern and Eastern Chinese breeds (fig. 1E). For the candidate genes in breeds from the Taklimakan Desert region, the absolute values of the frequency differences for the large-effect SNPs were estimated between pairwise groups of Qinghai-Tibetan breeds, Yunnan-Kweichow breeds, breeds from the Taklimakan Desert region (i.e., LOP and BRK) and Northern and Eastern Chinese breeds (except for breeds from the Taklimakan Desert region; fig. 1E). The statistical significance of the frequency difference values between the pairwise-group comparisons was evaluated by the Kruskal-Wallis nonparameter test (Kruskal and Wallis 1952). In addition, we compared the $Z(F_{ST})$ and $\log_2(\theta_\pi$ ratio) values for the selective genomic regions with those at the whole-genome scale for Tibetan sheep, breeds from the Taklimakan Desert region, breeds from high-altitude areas and breeds from the arid zone. For the altitude-associated and arid-associated genes identified in extreme environments, we further compared the overlap between the candidate genes identified here and the predefined gene panels (i.e., the previously published candidate genes of other mammalian species under similar extreme environments) with the overlap expected by chance (Hancock et al. 2008; Lv et al. 2014). The

predefined high-altitude gene panel comprised human (Simonson et al. 2010; Yi et al. 2010), dog (Gou et al. 2014; Li et al. 2014), wolf (Zhang et al. 2014), yak (Qiu et al. 2012), pig (Li et al. 2013) and Tibetan antelope (Ge et al. 2013), whereas the predefined arid gene panel included the Bactrian camel (Wu et al. 2014).

We used the sheep reference genome assembly Oar_v3.1.75 to identify the coordinates of nucleotide sequences, and used the most up-to-date database of sheep gene annotation in Ensembl (Cunningham et al. 2015; http://www.ensembl.org, last accessed October 2, 2015) to assess gene function. We performed Gene Ontology (GO) and functional pathway analyses of the candidate genes (i.e., all of the annotated genes in the outlier windows with the top 5% Z(F_{ST}) and $\log_2$(θ_π ratio) values) using the PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System version 10 (Mi et al. 2013; Mi et al. 2016). Specifically, the genes were submitted to PANTHER for functional enrichment analysis of GO biological processes (GO-BP), molecular function (GO-MF) and Cellular Component (GO-CC) terminologies (The Gene Ontology Consortium 2013) and pathways (Pathway; Mi and Thomas 2009). In the PANTHER tests, all known human genes were used as the background, and the Binomial distribution test (Mi et al. 2013) was used to assess the statistical significance (*P*-values) of the tests. Only significantly ($P < 0.05$) over-represented GO terms were considered to be biologically meaningful. We also positioned the candidate genes on known KEGG pathways (http://www.kegg.jp, last accessed October 16, 2015) and Reactome pathways (http://www.reactome.org, last

accessed October 16, 2015). The functions of the candidate genes were consulted based on the annotations in the NCBI (http://www.ncbi.nlm.nih.gov, last accessed October 2, 2015) and Ensembl (http://www.ensembl.org, last accessed October 2, 2015) databases. The expression levels of the candidate genes from the proposed network of pathways (see Results) in different sheep organs were tested through four functional genomics experiments in the EBI Gene Expression Atlas database (Petryszak et al. 2014).

Target gene analysis

To further dissect the genetic mechanisms that enable sheep to adapt to the plateau and desert environments, we focused on two target genes that exhibited high F_{ST} ranking and were located in the proposed network of pathways (*SOCS2* in Tibetan sheep; *GPX3* in breeds from the Taklimakan Desert region; see Results). The Z(F_{ST}) and $\log_2(\theta_\pi$ ratio) values of the two target genes were compared with adjacent regions. As Tajima's *D* (Tajima 1989) value is an informative indicator of selective signature (e.g., a negative value indicates a recent selective sweep and a positive value shows balancing selection), we used Tajima's *D* to test the target genes. Tajima's *D* was computed as $D = \frac{d}{\sqrt{\hat{V}(d)}}$, in which d is the difference between the mean number of pairwise differences (π) and the number of segregating sites (S), and $\hat{V}(d)$ is the square root of the standard deviation of d (Tajima 1989). We expected smaller (i.e., more negative) values of Tajima's *D* to be observed in the target genes under extreme environments than under contrasting environments. The inter-species NJ tree was

constructed for the *SOCS2* gene using TreeBeST v1.9.2 (http://treesoft.sourceforge.net/treebest.shtml, last accessed August 6, 2013) based on the orthologous sequences from 12 other vertebrates retrieved from Ensembl (Cunningham et al. 2015; http://www.ensembl.org, last accessed October 2, 2015). The allele frequencies of the SNPs in the *GPX3* gene were summarized across the native sheep breeds. Moreover, the expression levels of the two target genes in different sheep and human organs were tested through four and seven functional genomics experiments in the EBI Gene Expression Atlas database (Petryszak et al. 2014), respectively.

Supplementary Results and Discussion

Genome sequencing and mapping

By sequencing the whole genomes of 77 Chinese native sheep and three wild species, we generated a total of 1356.75 Gb of raw sequences and obtained 1267.51 Gb of high-quality sequences, with 95.16% and 87.29% of the data showing Phred quality scores higher than 20 and 30, respectively (supplementary table S30 and fig. S15, Supplementary Material online). Based on the high-quality sequence data, the genomic GC content was on average 44.51% for the 77 native sheep, 44.20% for *O. a. musimon*, 42.45% for *O. a. polii* and 42.67% for *C. ibex* (supplementary table S30, Supplementary Material online). Such GC ratios are typical and comparable to those observed in the genomes of cattle, yak (Qiu et al. 2012), Tibetan antelope, cow, horse, human and mouse (Ge et al. 2013), indicating that our sequencing data were not

influenced by GC bias. The mapping rates of our samples to the Oar_v3.1.75 reference genome assembly ranged from 83.90% to 99.85%, and the mean mapping coverage was 96.01% at ≥ 1× and 60.45% at ≥ 4× for all 80 samples (supplementary table S30, Supplementary Material online). After mapping, we obtained 1149.70 Gb of aligned high-quality data and a 438× effective sequence coverage. For individual samples, an average of 123.04 Gb of high-quality data, a 41.77× effective sequencing depth, 98.99% coverage at≥ 1 × and 98.52% coverage at ≥ 4× were generated for the two high-coverage sequencing samples (ZNQ24 and LOP42); an average of 13.10 Gb (11.48 – 17.52 Gb) of high-quality data, at 4.54× (3.84× – 6.10×) effective sequencing depth, 95.93% (91.93% – 97.76%) coverage at≥ 1 × and 59.48% (46.83% – 77.61%) coverage at ≥ 4× were obtained for the other 78 individual samples (supplementary table S30, Supplementary Material online). With regard to the genomic distribution of high-quality SNPs in the 77 native sheep, the number of SNPs located on the 26 autosomes ranged from 0.34 to 2.29 million, and 0.62 million SNPs were located on the X chromosome. For the individual genome of native sheep (i.e., average over 77 individuals), there were 0.0678 to 0.398 million SNPs located on the autosomes and 0.103 million located on the X chromosome (supplementary table S8, Supplementary Material online).

Principal component analysis (PCA) and site frequency spectrum (SFS)

In the individual-based PCA, the first component explicitly distinguished Yunnan-Kweichow breeds from the others (fig. 1F, supplementary fig. S4,

Supplementary Material online). Furthermore, the second and third eigenvectors further split Tibetan sheep from Northern and Eastern Chinese breeds, with GZS, MXS and GDS being situated between the two groups (fig. 1F, supplementary fig. S4, Supplementary Material online).

A comparison of the SFS distribution among all 80 samples (supplementary fig. S16A, Supplementary Material online) and the 77 native sheep (supplementary fig. S16B, Supplementary Material online) indicated that substantially more SNPs were shared among the native sheep than between the native sheep and the three wild species. After filtering, a large number of low-frequency SNPs were eliminated, which increased the proportion of SNPs shared among all animals (supplementary fig. S16A and B, Supplementary Material online).

Functions of important candidate genes in plateau environment

Regarding the pathway candidate genes involved in plateau adaptations in Tibetan sheep (fig. 7A), *NOX4* (NADPH Oxidase 4), a source of reactive oxygen species (ROS), could function as an oxygen sensor that regulates *HIF1a* activity and oxygen-dependent VEGF gene expression in human cells (Lassègue and Clempus 2003; Meng et al. 2012). *PDK1* (Pyruvate Dehydrogenase Kinase, Isozyme 1), which is induced by *HIF1a*, plays critical roles in conserving mitochondrial function under hypoxic conditions by diverting metabolic intermediates from the tricarboxylic acid cycle (TCA) to glycolysis and preventing toxic ROS production (Kim et al. 2006; Jian

et al. 2010). A deficiency of *PDK1* in cardiac muscle could result in heart failure and increased sensitivity to hypoxia (Mora et al. 2003). *NCOA3* (Nuclear Receptor Coactivator 3) affects the expression of the *EPO* (Erythropoietin) gene which is known as the primary hormone in regulation of erythrocyte differentiation and circulating erythrocyte mass in response to hypoxia (Stopka et al. 1998; Wang et al. 2010). Knocking down *NCOA3* in human Hep3B cells decreases hypoxia-induced *EPO* transcriptional activity (Wang et al. 2010). *LONP1* (Lon Peptidase 1, Mitochondrial) can be induced by hypoxia and ROS, and provides protection for cells against oxidative stress (Pinti et al. 2015). *PDGFD* (Platelet Derived Growth Factor D) has a role in the regulation of blood vessel maturation during angiogenesis (Uutela et al. 2004). *RRAS* (Related RAS Viral Oncogene Homolog) regulates the integrity and functionality of tumor blood vessels under pathological hypoxic conditions (Sawada et al. 2012). *NF1* (Neurofibromin 1) is associated with typical hypervascular tumors which show high VEGF expression levels, and silencing this gene induces upregulation of VEGF expression in murine cells (Kawachi et al. 2013), indicating a potential role of *NF1* in preventing tumors or hypoxic diseases related to VEGF. *HAND2* (Heat And Neural Crest Derivatives Expressed 2) is essential for cardiac morphogenesis, and is required for vascular development and the regulation of angiogenesis (Villanueva et al. 2002). *PRKG1* (Protein Kinase, CGMP-Dependent, Type 1) acts as key mediator of NO and cGMP in the VSMC pathway, and phosphorylates proteins that regulate smooth muscle contraction and cardiac function (Tang et al. 2003). *NOSIP* (Nitric Oxide Synthase Interacting Protein) modulates the

activity and localization of nitric oxide synthase (*NOS1* and *NOS3*) and thus NO production (Dreyer et al. 2004; Schleicher et al. 2005). NO dilates blood vessels and consequently increases blood flow and delivers more oxygen to tissues (Ge et al. 2013).

For the genes governing the body morphology of Tibetan sheep, *FSTL1* (Follistatin-Like 1), which encodes a secreted protein of the BMP inhibitor and functions through interactions with the BMP signaling pathway, has been implicated to control the development of different organs like skeletal, lung and ureter in zebrafish and mouse (Geng et al. 2011; Sylva et al. 2011; Sylva et al. 2013). *EXT2* (Exostosin Glycosyltransferase 2) encodes an enzyme that synthesizes heparan sulfate which is involved in limb and brain development in mouse (Inatani and Yamaguchi 2003; Norton et al. 2005). *ALX4* (ALX Homeobox 4) plays an essential role in regulating skull, limb, skin and embryonic development in vertebrates (Panman et al. 2005; Boras-Granic et al. 2006). Dysfunction of *ALX4* substantially disrupts craniofacial and epidermal development in human (Kayserili et al. 2009). *SOX6* (Sex Determining Region Y-Box 6) encodes a protein that is required for normal development of the nervous system, chondrogenesis, and skeletal muscle (Hagiwara 2011). Disruption of *SOX6* is associated with delayed development and dysmorphic features in human (Ebrahimi-Fakhari et al. 2015). *BMPR2* (Bone Morphogenetic Protein Receptor, Type II) is involved in endochondral bone formation, ovary development and embryogenesis (Zhao et al. 2002; Rossi et al. 2016). Mutations in

BMPR2 have been associated with typical hypoxia-induced diseases, including primary pulmonary hypertension and pulmonary venoocclusive disease (Long et al. 2015).

Functions of the pathway genes in desert environment

Regarding the pathway candidate genes involved in desert environment adaptations (fig. 7B), *ANXA6* (Annexin A6), *CALM2* (Calmodulin 2) and *CACNA2D1* (Calcium Channel, Voltage-Dependent, Alpha 2/Delta Subunit 1) are all involved in the regulation of cellular Ca^{2+}, which are probably functionally related to water-salt metabolism. *ANXA6* encodes a protein that belongs to a family of calcium-dependent membrane proteins. This gene may regulate the release of Ca^{2+} from intracellular stores and is involved in the GO term of ion transmembrane transport (GO:0034220). *CALM2* (Calmodulin 2) controls a large number of enzymes and ion channels by Ca^{2+}, and *CACNA2D1* regulates the calcium current density and calcium channels (Schleithoff et al. 1999). *PTGS2* (Prostaglandin-Endoperoxide Synthase 2), also known as *COX2* (Cyclooxygenase 2), encodes a rate-limiting enzyme in the conversion of arachidonic acid into prostaglandins. The prostaglandins regulate blood flow and water-salt absorption (Gatalica et al. 2008), and interact with prostanoid receptors which exert important regulatory effects on renal function (Breyer and Breyer 2000). *CPA3* (Carboxypeptidase A3) encodes a secretory granule metalloexopeptidase. The GO annotations related to this gene include regulation of angiotensin level in blood (GO:0002002) and angiotensin maturation (GO:0002003),

both of which are potentially involved in renal vasodilation and natriuresis (Nehme et al. 2015). *ECE1* (Endothelin Converting Enzyme 1) encodes proteins associated with the proteolytic processing of endothelin precursors to biologically active peptides that mediate vasoconstriction (Maquire et al. 1997). *KCNJ5* (Potassium Channel, Inwardly Rectifying Subfamily J, Member 5) is functionally annotated to GO terms related to potassium ion transport process (GO:0006813, GO:0010107, GO:0034765 and GO:0071805). *SLC4A4* (Solute Carrier Family 4, Member 4) is involved in regulating bicarbonate secretion and absorption (Aalkjaer et al. 2004). The GO annotations for this gene are mainly associated with sodium ion (GO:0006814) and bicarbonate (GO:0015701) transport.

Positively selected genes related to high-altitude and arid environment adaptations

The functions of the following candidate genes are discussed based on the annotations in the NCBI (http://www.ncbi.nlm.nih.gov, last accessed October 2, 2015) and Ensembl (http://www.ensembl.org, last accessed October 2, 2015) databases. Regarding high-altitude environment adaptation, *STK17A*, which exhibited the second highest Z(Fst) value, is a regulator of cellular reactive oxygen species (ROS) — an important active molecule in the HIF-1 pathway. An important paralog of *STK17A*, the *MYLK4* gene, is functionally involved in the vascular smooth muscle contraction (VSMC) pathway, which plays an essential role in regulating the diameter of blood vessels and the speed of blood flow. *EIF2AK3* has functions associated with

integrated stress responses such as the response to oxidative stress. An important paralog of *TMTC2*, the *TMTC3* gene, is a member of the functionally enriched GO terms related to lung (GO:0030324) and lung alveolus (GO:0048286) development. *PTPN9* plays a key role in haematopoiesis and modulates the *EGFR* gene located in the HIF-1 pathway. *MREG* has GO annotations in pigmentation (GO:0043473) and melanocyte differentiation (GO:0030318), which are possibly associated with response to UV. Regarding arid environment adaptation, *RAB11FIP2* is involved in the regulation of water balance through the renal aquaporins pathway (Reactome), and its associated GO terms are related to renal water homeostasis (GO:0003091) and water transport (0006833). *CPVL* is located in the renin-angiotensin system pathway (KEGG), and the GO annotation of *MFSD6* includes transmembrane transport (GO:0055085).

Author Contributions

M.-H.L. conceived and managed the project. M.-H.L., M.-J.L. and W.-R.L. designed the project; M.-H.L., F.-H.L., S.-G.H., W.-F.P., Y.-X.Z., M.Z., X.-L.X., J.-Q.L., Y.-G.L., Z.-Q.S., F.W., Y.-W.S., Y.-T.W., J.K. and J.-L.H. prepared the samples; F.-H.L., J.Y. and S.-L.T. analyzed the data with contributions from X.-L.T., G.-J.L. and H.-F.L.; J.Y., F.-H.L., M.-H.L., M.-J.L. and W.-R.L. interpreted the data; J.Y. wrote the manuscript with contributions from M.-H.L, F.-H.L and J.-L. H. All authors verified the final version of the manuscript.

Supplementary References

Aalkjaer C, Frische S, Leipziger J, Nielsen S, Praetorius J. 2004. Sodium coupled bicarbonate transporters in the kidney, an update. *Acta Physiol Scand*. 181:505–512.

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 21:974–984.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet*. 12:363–376.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265.

Boras-Granic K, Grosschedl R, Hamel PA. 2006. Genetic interaction between *Lef1* and *Alx4* is required for early embryonic development. *Int J Dev Biol*. 50:601–610.

Breyer MD, Breyer RM. 2000. Prostaglandin receptors: their role in regulating renal function. *Curr Opin Nephrol Hypertens*. 9:23–29.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 81:1084–1097.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*.

6:677–681.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2015. Ensembl 2015. *Nucleic Acids Res*. 43:D662–D669.

de Magalhães JP, Costa J. 2009. A database of vertebrate longevity records and their relation to other life-history traits. *J Evol Biol*. 22:1770–1774.

Dreyer J, Schleicher M, Tappe A, Schilling K, Kuner T, Kusumawidijaja G, Müller-Esterl W, Oess S, Kuner R. 2004. Nitric oxide synthase (*NOS*)-interacting protein interacts with neuronal *NOS* and regulates its distribution and activity. *J Neurosci*. 24:10454–10465.

Du L-X. 2011. Animal Genetic Resources in China. Beijing: China Agriculture Press.

Ebrahimi-Fakhari D, Maas B, Haneke C, Niehues T, Hinderhofer K, Assmann BE, Runz H. 2015. Disruption of *SOX6* is associated with a rapid-onset dopa-responsive movement disorder, delayed development, and dysmorphic features. *Pediatr Neurol*. 52:115–118.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME. 2006. Copy number variation: new insights in genome diversity. *Genome Res*. 16:949–961.

Frichot E, Schoville SD, Bouchard G, François O. 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol*. 30:1687–1699.

Gatalica Z, Lilleberg SL, Koul MS, Vanecek T, Hes O, Wang B, Michal M. 2008.

COX-2 gene polymorphisms and protein expression in renomedullary interstitial cell tumors. *Hum Pathol*. 39:1495–1504.

Ge R-L, Cai Q, Shen Y-Y, San A, Ma L, Zhang Y, Yi X, Chen Y, Yang L, Huang Y, et al. 2013. Draft genome sequence of the Tibetan antelope. *Nat Commun*. 4:1858.

Geng Y, Dong Y, Yu M, Zhang L, Yan X, Sun J, Qiao L, Geng H, Nakajima M, Furuichi T, et al. 2011. Follistatin-like 1 (*Fstl1*) is a bone morphogenetic protein (*BMP*) 4 signaling antagonist in controlling mouse lung development. *Proc Natl Acad Sci USA*. 108:7058–7063.

Gou X, Wang Z, Li N, Qiu F, Xu Z, Yan D, Yang S, Jia J, Kong X, Wei Z, et al. 2014. Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome Res*. 24:1308–1315.

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

Hagiwara N. 2011. *Sox6*, jack of all trades: a versatile regulatory protein in vertebrate development. *Dev Dyn*. 240:1311–1321.

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet*. 4:e32.

Hill WG, Robertson A. 1968. The effects of inbreeding at loci with heterozygote advantage. *Genetics* 60:615–628.

Inatani M, Yamaguchi Y. 2003. Gene expression of *EXT1* and *EXT2* during mouse

brain development. *Dev Brain Res*. 141:129–136.

Jian B, Wang D, Chen D, Voss J, Chaudry I, Raju R. 2010. Hypoxia-induced alteration of mitochondrial genes in cardiomyocytes - role of *Bnip3* and *Pdk1*. *Shock* 34:169–175.

Kawachi Y, Maruyama H, Ishitsuka Y, Fujisawa Y, Furuta J, Nakamura Y, Ichikawa E, Furumura M, Otsuka F. 2013. *NF1* gene silencing induces upregulation of vascular endothelial growth factor expression in both Schwann and non-Schwann cells. *Exp Dermatol.* 22:262–265.

Kayserili H, Uz E, Niessen C, Vargel I, Alanay Y, Tuncbilek G, Yigit G, Uyguner O, Candan S, Okur H, et al. 2009. *ALX4* dysfunction disrupts craniofacial and epidermal development. *Hum Mol Genet.* 18:4357–4366.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Neto LRP, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K, et al. 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol*. 10:e1001258.

Kim JW, Tchernyshyov I, Semenza GL, Dang CV. 2006. *HIF-1*-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. *Cell Metab.* 3:177–185.

Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, Grarup N, Jiang T, Andersen G, Witte D, et al. 2011. Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* 12:231.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15:356.

Kruskal WH, Wallis WA. 1952. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc*. 47:583–621.

Lassègue B, Clempus RE. 2003. Vascular *NAD(P)H* oxidases: specific features, expression, and regulation. *Am J Physiol Regul Integr Comp Physiol*. 285:R277–R297.

Li C, Zhang Y, Li J, Kong L, Hu H, Pan H, Xu L, Deng Y, Li Q, Jin L, et al. 2014. Two Antarctic penguin genomes reveal insights into their evolutionary history and molecular changes related to the Antarctic environment. *GigaScience* 3:27.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475:493–496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.

Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, Wang T, Yeung CK, Chen L, Ma J, et al. 2013. Genomic analyses identify distinct patterns of selection in domesticated

pigs and Tibetan wild boars. *Nat Genet*. 45:1431–1438.

Li Y, Wu D-D, Boyko AR, Wang G-D, Wu S-F, Irwin DM, Zhang Y-P. 2014. Population variation revealed high-altitude adaptation of Tibetan mastiffs. *Mol Biol Evol*. 31:1200–1205.

Long L, Ormiston ML, Yang X, Southwood M, Gräf S, Machado RD, Mueller M, Kinzel B, Yung LM, Wilkinson JM, et al. 2015. Selective enhancement of endothelial *BMPR-II* with *BMP9* reverses pulmonary arterial hypertension. *Nat Med*. 21:777–785.

Lv F-H, Agha S, Kantanen J, Colli L, Stucki S, Kijas JW, Joost S, Li M-H, Ajmone Marsan P. 2014. Adaptations to climate-mediated selective pressures in sheep. *Mol Biol Evol*. 31:3324–3343.

Lv F-H, Peng W-F, Yang J, Zhao Y-X, Li W-R, Liu M-J, Ma Y-H, Zhao Q-J, Yang G-L, Wang F, et al. 2015. Mitogenomic meta-analysis identifies two phases of migration in the history of eastern Eurasian sheep. *Mol Biol Evol*. 32:2515–2533.

Maquire JJ, Johnson CM, Mockridge JW, Davenport AP. 1997. Endothelin converting enzyme (*ECE*) activity in human vascular smooth muscle. *Br J Pharmacol*. 122:1647–1654.

Meng D, Mei A, Liu J, Kang X, Shi X, Qian R, Chen S. 2012. *NADPH* oxidase 4 mediates insulin-stimulated *HIF-1a* and VEGF expression, and angiogenesis in vitro. *PLoS ONE* 7:e48393.

Mi H, Muruganujan A, Casagrande JT, Thomas PD. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc*. 8:1551–1566.

Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. 2016. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res*. 44:D336–D342.

Mi H, Thomas P. 2009. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol*. 563:123–140.

Mora A, Davies AM, Bertrand L, Sharif I, Budas GR, Jovanović S, Mouton V, Kahn CR, Lucocg JM, Gray GA, et al. 2003. Deficiency of *PDK1* in cardiac muscle results in heart failure and increased sensitivity to hypoxia. *EMBO J*. 22:4666–4676.

Nehme A, Zibara K, Cerutti C, Bricca G. 2015. Gene expression analysis and bioinformatics revealed potential transcription factors associated with rennin-angiotensin-aldosterone system in atheroma. *J Hypertens*. 33:e115.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA*. 76:5269–5273.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. 2012. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE* 7:e37558.

Norton WHJ, Ledin J, Grandel H, Neumann CJ. 2005. HSPG synthesis by zebrafish *Ext2* and *Extl3* is required for *Fgf10* signalling during limb development. *Development* 132:4963–4973.

Panman L, Drenth T, Tewelscher P, Zuniga A, Zeller R. 2005. Genetic interaction of *Gli3* and *Alx4* during limb development. *Int J Dev Biol*. 49:443–448.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2:e190.

Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, et al. 2014. Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res*. 42:D926–D932.

Piao S, Ciais P, Huang Y, Shen Z, Peng S, Li J, Zhou L, Liu H, Ma Y, Ding Y, et al. 2010. The impacts of climate change on water resources and agriculture in China. *Nature* 467:43–51.

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

Pinti M, Gibellini L, Liu Y, Xu S, Lu B, Cossarizza A. 2015. Mitochondrial Lon protease at the crossroads of oxidative stress, aging and cancer. *Cell Mol Life Sci*. 72:4807–4824.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: A tool set for Whole-Genome association and population-based linkage analyses. *Am J Hum Genet*. 81:559–575.

Qiu Q, Zhang G, Ma T, Qian W, Wang J, Ye Z, Cao C, Hu Q, Kim J, Larkin DM, et al. 2012. The yak genome and adaptation to life at high altitude. *Nat Genet*. 44:946–949.

Rosenberg NA. 2004. DISTRUCT: a program for the graphical display of population

structure. *Mol Ecol Notes*. 4:137–138.

Rossi RO, Costa JJ, Silva AW, Saraiva MV, Van den Hurk R, Silva JR. 2016. The bone morphogenetic protein system and the regulation of ovarian follicle development in mammals. *Zygote* 24:1–17.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918.

Saitou N, Nei M. 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 4:406–425.

Sambrook J, Russell D. 2000. Molecular Cloning: A Laboratory Manual, 3rd edn. Cold Spring Harbor Laboratory, New York: Cold Spring Harbor Laboratory Press.

Sawada J, Urakami T, Li F, Urakami A, Zhu W, Fukuda M, Li DY, Ruoslahti E, Komatsu M. 2012. Small *GTPase R-Ras* regulates integrity and functionality of tumor blood vessels. *Cancer Cell* 22:235–249.

Schleicher M, Brundin F, Gross S, Müller-Esterl W, Oess S. 2005. Cell cycle-regulated inactivation of endothelial NO synthase through *NOSIP*-dependent targeting to the cytoskeleton. *Mol Cell Biol*. 25:8251–8258.

Schleithoff L, Mehrke G, Reutlinger B, Lehmann-Horn F. 1999. Genomic structure and functional expression of a human alpha(2)/delta calcium channel subunit gene (*CACNA2*). *Genomics* 61:201–209.

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo FR, Xing J, Jorde LB, et al. 2010. Genetic evidence for high-altitude adaptation in

Tibet. *Science* 329:72–75.

Stopka T, Zivny JH, Stopkova P, Prchal JF, Prchal JT. 1998. Human hematopoietic progenitors express erythropoietin. *Blood* 91:3766–3772.

Sylva M, Li VS, Buffing AA, van Es JH, van den Born M, van der Velden S, Gunst Q, Koolstra JH, Moorman AF, Clevers H, et al. 2011. The *BMP* antagonist follistatin-like 1 is required for skeletal and lung organogenesis. *PLoS ONE* 6:e22616.

Sylva M, Moorman AFM, van den Hoff MJB. 2013. Follistatin-like 1 in vertebrate development. *Birth Defects Res C Embryo Today*. 99:61–69.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol*. 28:289–301.

Tang M, Wang GR, Lu P, Karas RH, Aronovitz M, Heximer SP, Kaltenbronn KM, Blumer KJ, Siderovski DP, Zhu Y, et al. 2003. Regulator of G-protein signaling-2 mediates vascular smooth muscle relaxation and blood pressure. *Nat Med*. 9:1506–1512.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population scale sequencing. *Nature* 467:1061–1073.

The Gene Ontology Consortium. 2013. Gene Ontology annotations and resources.

Nucleic Acids Res. 41:D530–D535.

Uutela M, Wirzenius M, Paavonen K, Rajantie I, He Y, Karpanen T, Lohela M, Wiig H, Salven P, Pajusola K, et al. 2004. *PDGF-D* induces macrophage recruitment, increased interstitial pressure, and blood vessel maturation during angiogenesis. *Blood* 104:3198–3204.

Villanueva MP, Aiyer AR, Muller S, Pletcher MT, Liu X, Emanuel B, Srivastava D, Reeves RH. 2002. Genetic and comparative mapping of genes dysregulated in mouse hearts lacking the *Hand2* transcription factor gene. *Genomics* 80:593–600.

Wang F, Zhang R, Wu X, Hankinson O. 2010. Roles of coactivators in hypoxia induction of the erythropoietin gene. *PLoS ONE* 5:e10002.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 38:e164.

Weir BS, Cockerham CC. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution* 38:1358–1370.

Wu H, Guang X, Al-Fageeh MB, Cao J, Pan S, Zhou H, Zhang L, Abutarboush MH, Xing Y, Xie Z, et al. 2014. Camelid genomes reveal evolution and adaptation to desert environments. *Nat Commun*. 5:5188.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75–78.

Zhang W, Fan Z, Han E, Hou R, Zhang L, Galaverni M, Huang J, Liu H, Silva P, Li P, et al. 2014. Hypoxia adaptations in the grey wolf (*Canis lupus chanco*) from

Qinghai-Tibet Plateau. *PLoS Genet*. 10:e1004466.

Zhao M, Harris SE, Horn D, Geng Z, Nishimura R, Mundy GR, Chen D. 2002. Bone morphogenetic protein receptor signaling is necessary for normal murine postnatal bone formation. *J Cell Biol*. 157:1049–1060.

Zhong T, Han JL, Guo J, Zhao QJ, Fu BL, He XH, Jeon JT, Guan WJ, Ma YH. 2010. Genetic diversity of Chinese indigenous sheep breeds inferred from microsatellite markers. *Small Ruminant Res*. 90:88–94.

Supplementary Figures and Figure Legends

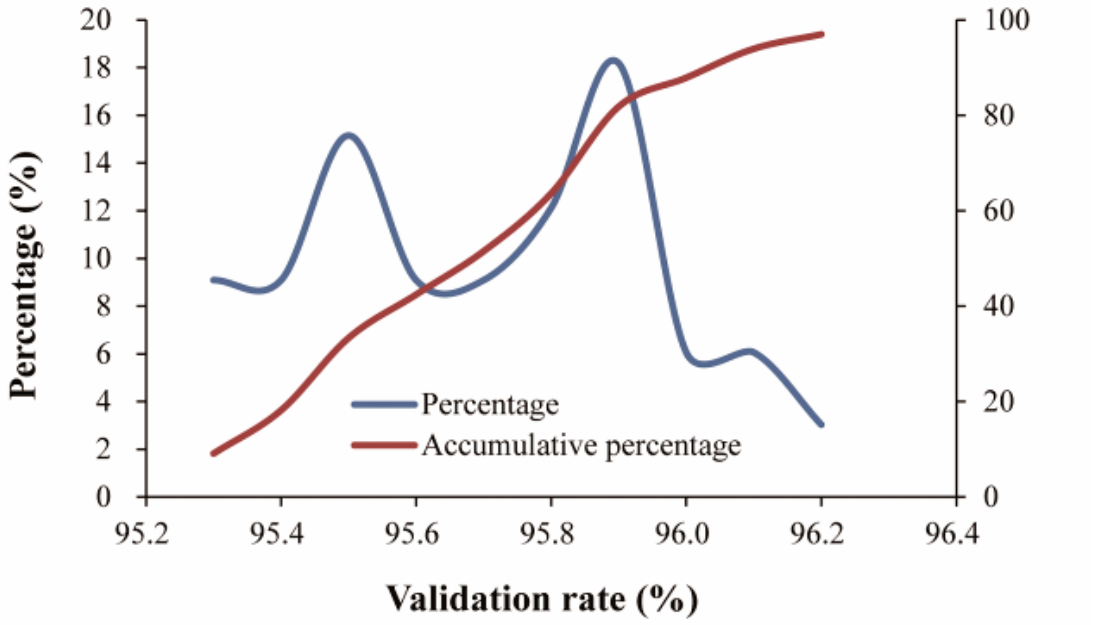

FIG. S1. SNP validation using a 50K chip array. The consistency between the sequence-based SNPs and the Illumina ovine 50K SNPs was tested in 33 native sheep samples. The sequencing-based SNPs with base pairs matching the ovine 50K chip SNPs were defined as validated SNPs. The distribution percentage (blue) and accumulative percentage (red) of validated SNPs for the 33 samples were recorded and plotted.

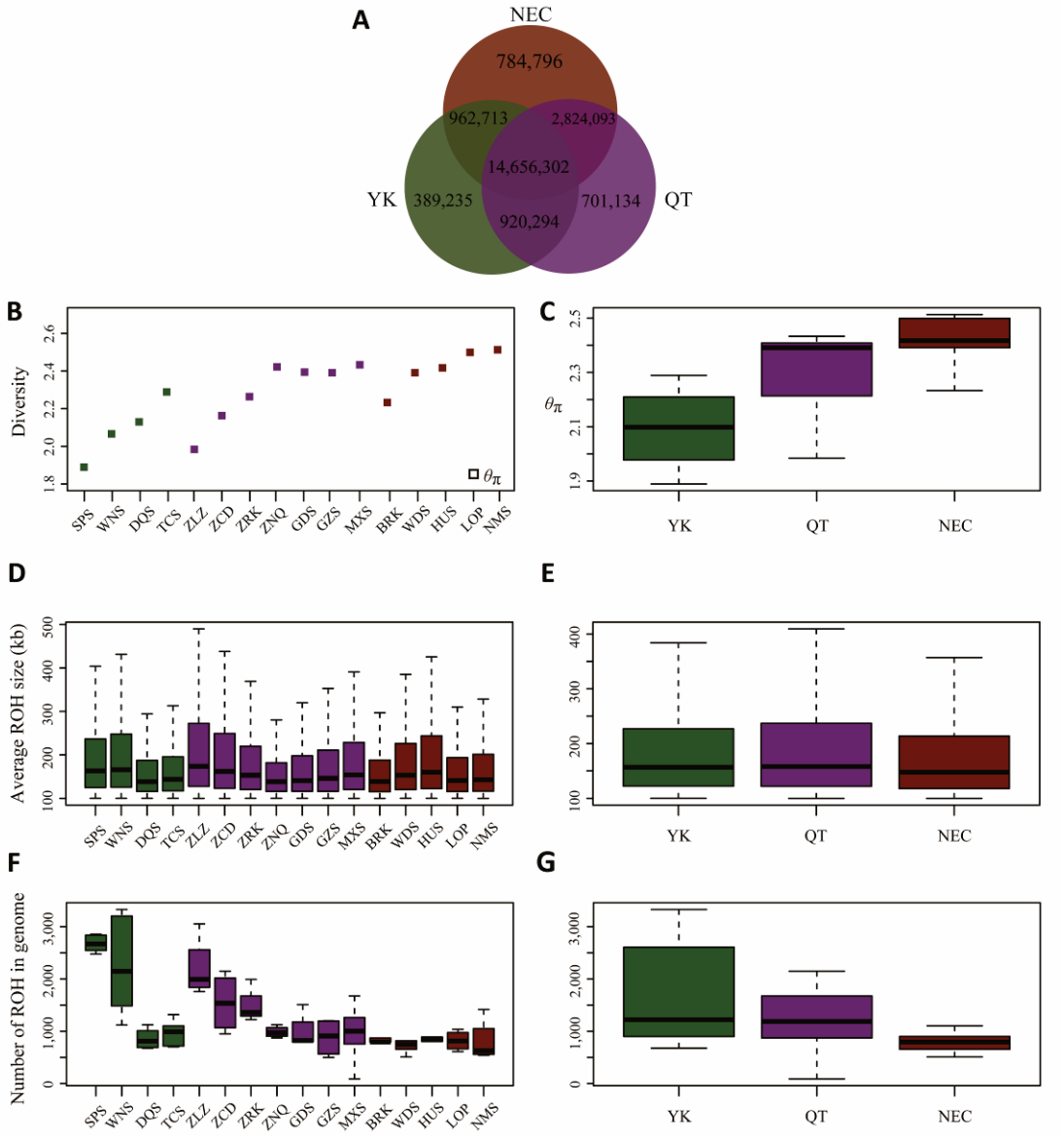

FIG. S2. Summary statistics for genomic variation. (A) A Venn diagram of the shared SNPs among the three groups of Chinese native sheep breeds (Qinghai-Tibetan breeds, n = 29; Yunnan-Kweichow breeds, n = 20; and Northern and Eastern Chinese breeds, n = 28). (B) Average within-breed pairwise nucleotide distance (θ_π) of the mutation rate per base pair. (C) Average within-group pairwise nucleotide distance (θ_π) values. (D) Average size of regions of homozygosity (ROH) in the genomes of Chinese native sheep breeds. (E) Average ROH size in the genomes of three Chinese sheep groups. (F) Average ROH number in the genomes

of Chinese native sheep breeds. (G) Average ROH number in the genomes of three Chinese sheep groups. NEC, Northern and Eastern Chinese breeds; QT, Qinghai-Tibetan breeds; YK, Yunnan-Kweichow breeds. See supplementary table S1, Supplementary Material online for the abbreviations of the breeds.

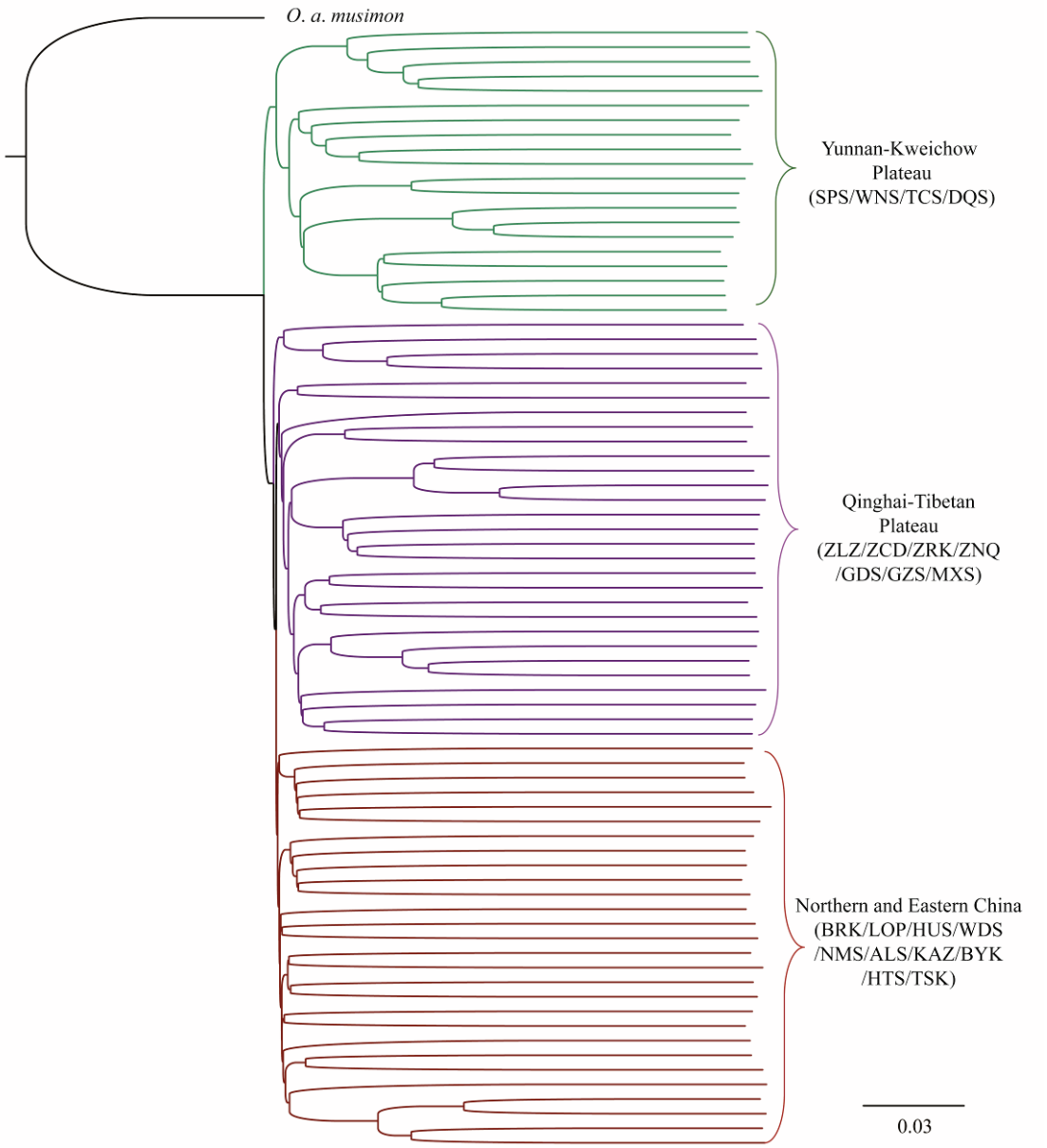

FIG. S3. Phylogenetic tree. Rooted NJ tree constructed from the *p*-distance matrix of SNPs among *O. aries* using *O. a. musimon* as an outgroup. Groups of breeds with different geographic origins are shown in different colors. NMS represents sheep breeds from Inner Mongolia and includes Ujimqin sheep (WZS), Wuranke sheep (WRS), Sunite sheep (SNS) and Hulun Buir sheep (HLS). For the abbreviations of the other sheep breeds, see supplementary table S1, Supplementary Material online.

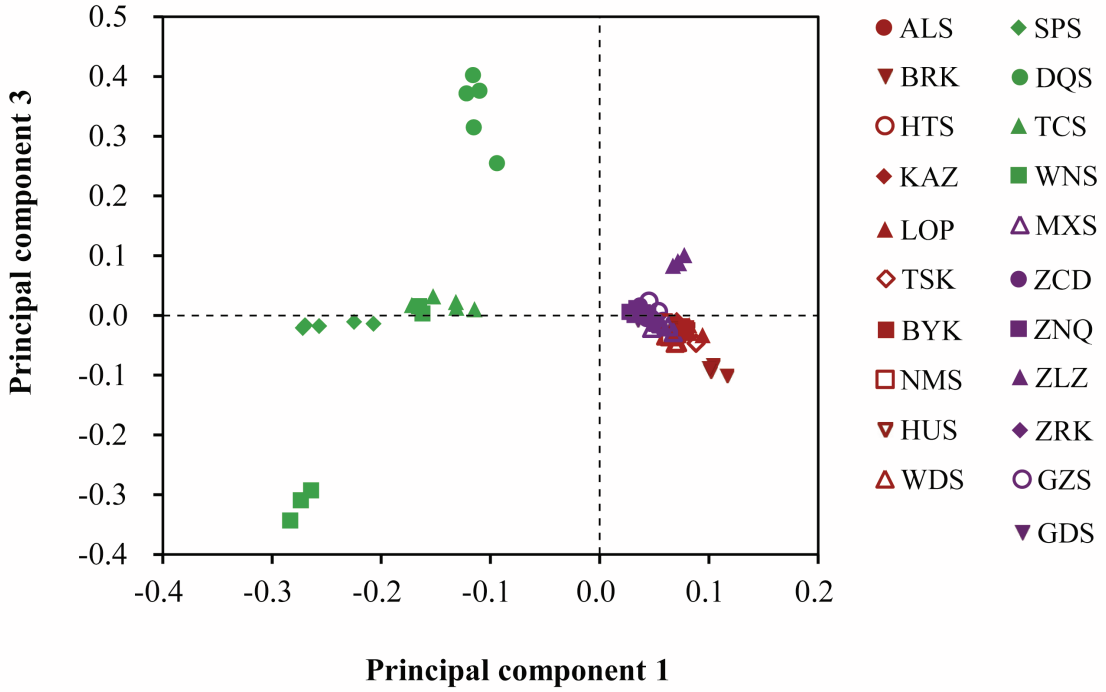

FIG. S4. PCA analysis. Principal components 1 and 3 from 21 native sheep breeds. NMS represents sheep breeds from Inner Mongolia and includes Ujimqin sheep (WZS), Wuranke sheep (WRS), Sunite sheep (SNS) and Hulun Buir sheep (HLS). For the abbreviations of the other sheep breeds, see supplementary table S1, Supplementary Material online.

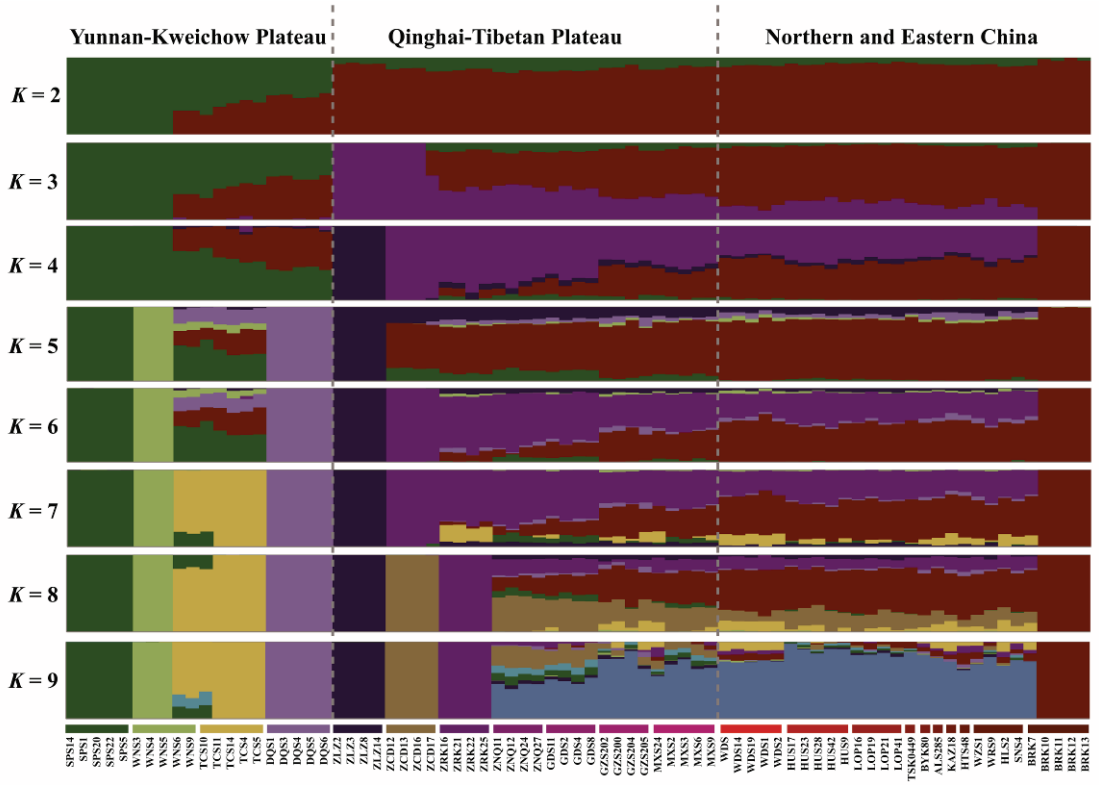

FIG. S5. Population genetic structure of 21 native sheep breeds inferred with FRAPPE v1.1. The length of each colored segment represents the proportion of the individual's genome inferred from K = 2–9 ancestral populations. For the abbreviations of the individuals, see supplementary table S1, Supplementary Material online.

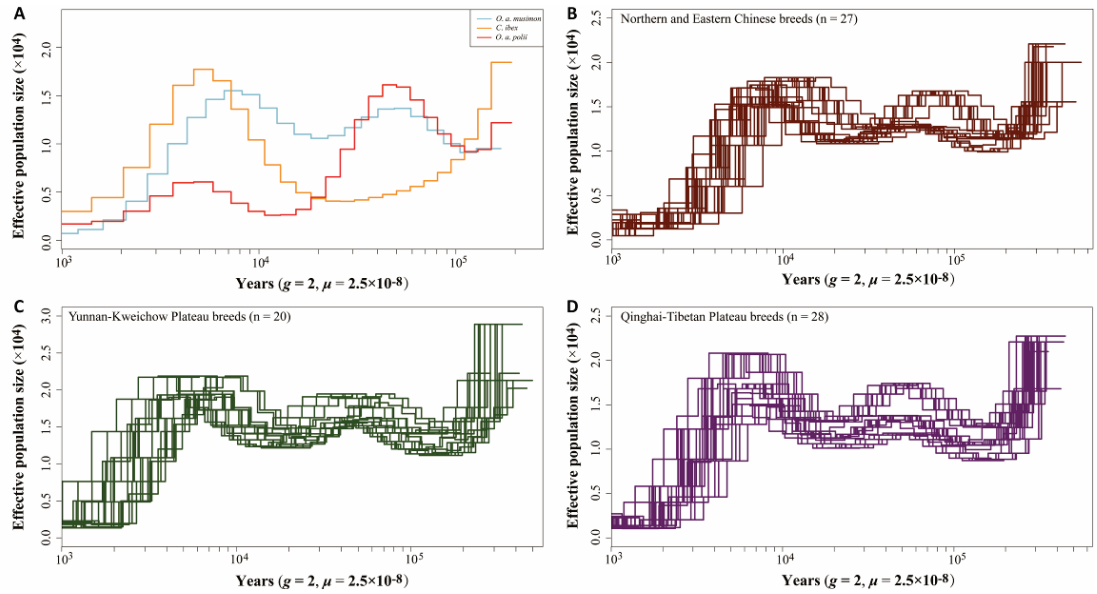

FIG. S6. Pairwise sequential Markovian coalescent (PSMC) analysis results for the native sheep sequenced at low coverage (~5×) exhibit inferred variations in N_e over the last 10^6 years. (A) PSMC results showing the demographic history of *O. a. musimon*, *O. a. polii* and *C. ibex*. (B) PSMC results showing the demographic history of Northern and Eastern Chinese sheep breeds (n = 27 animals). (C) PSMC results showing the demographic history of Yunnan-Kweichow sheep breeds (n = 20 animals). (D) PSMC results showing the demographic history of Qinghai-Tibetan sheep breeds (n = 28 animals).

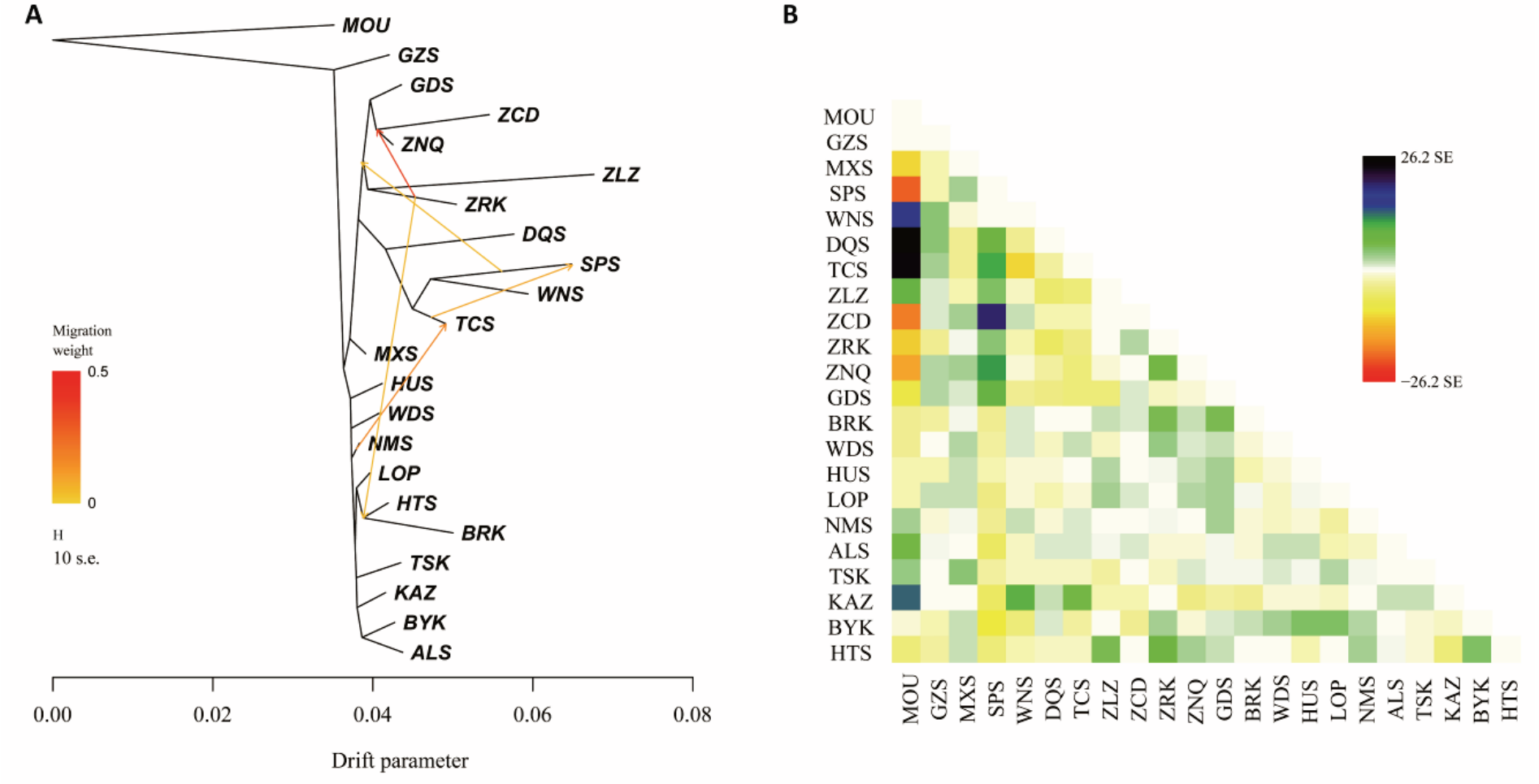

FIG. S7. Phylogenetic network of the inferred relationships among the 21 native breeds and an outgroup (*O. a. musimon*) with five migration

events. (A) Maximum likelihood (ML) tree of the 21 native breeds and an outgroup (*O. a. musimon*) with five migration events allowed**.** (B) Residual matrix fit from the ML tree. Arrows indicate migration events, and the spectrum of heat colors indicates the migration weights of the migration events. NMS represents sheep breeds from Inner Mongolia and includes Ujimqin sheep (WZS), Wuranke sheep (WRS), Sunite sheep (SNS) and Hulun Buir sheep (HLS). For the abbreviations of the other sheep breeds, see supplementary table S1, Supplementary Material online.

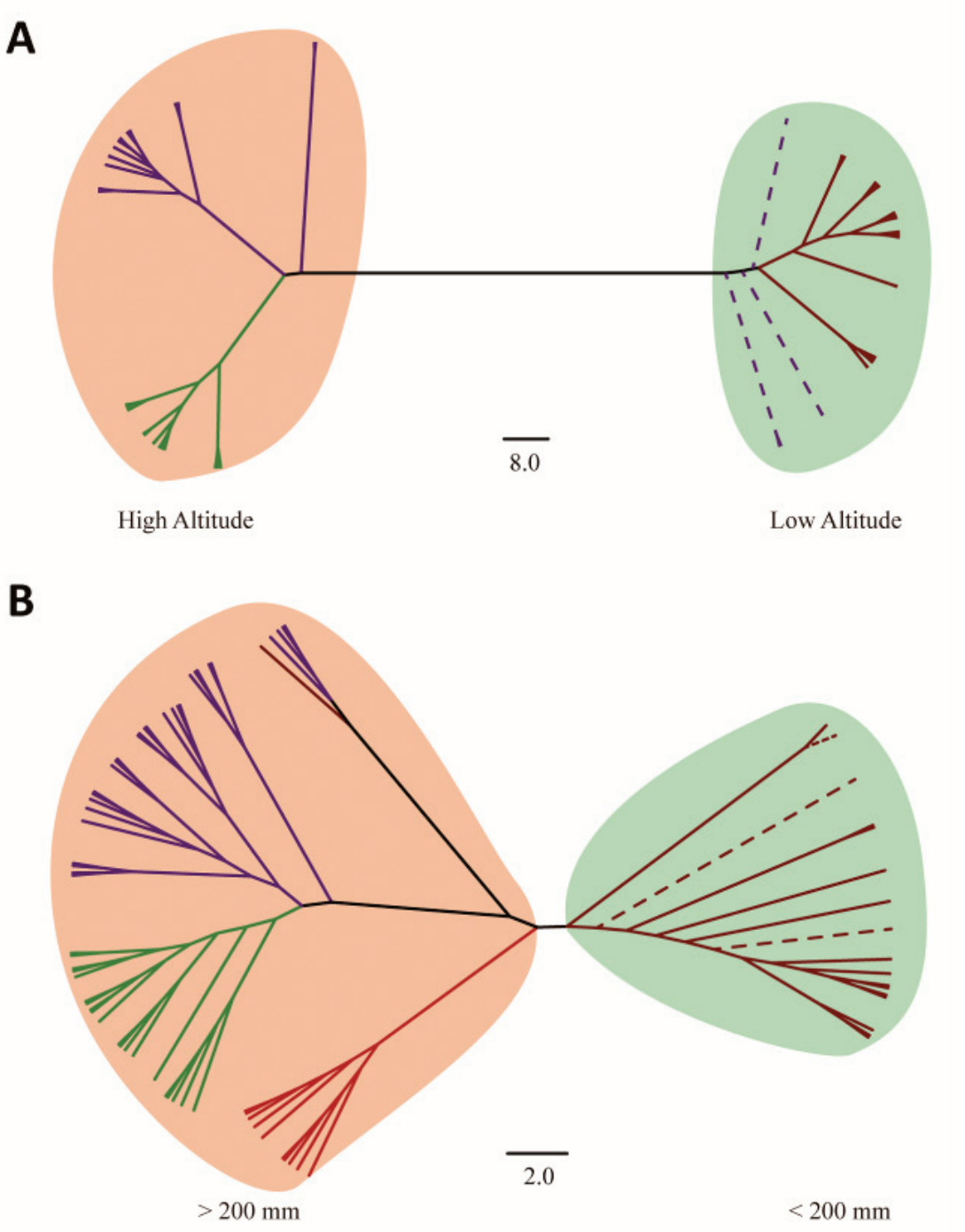

FIG. S8. Phylogenetic tree reconstructed from the SNPs located in candidate genes. (A) Neighbor-joining (NJ) tree reconstructed from the intragenic SNPs under selection in Tibetan sheep. (B) NJ tree reconstructed from the intragenic SNPs under selection in the sheep breeds from the Taklimakan Desert region. The samples

collected from the three geographic regions are indicated by colored lines (purple: Qinghai-Tibetan Plateau, China; green: Yunnan-Kweichow Plateau, China; red: Northern and Eastern China). The dotted lines represent outlier individuals sampled from high-altitude regions or the regions with average annual precipitation > 200 mm but were unexpectedly clustered into low-altitude regions or the regions with average annual precipitation < 200 mm.

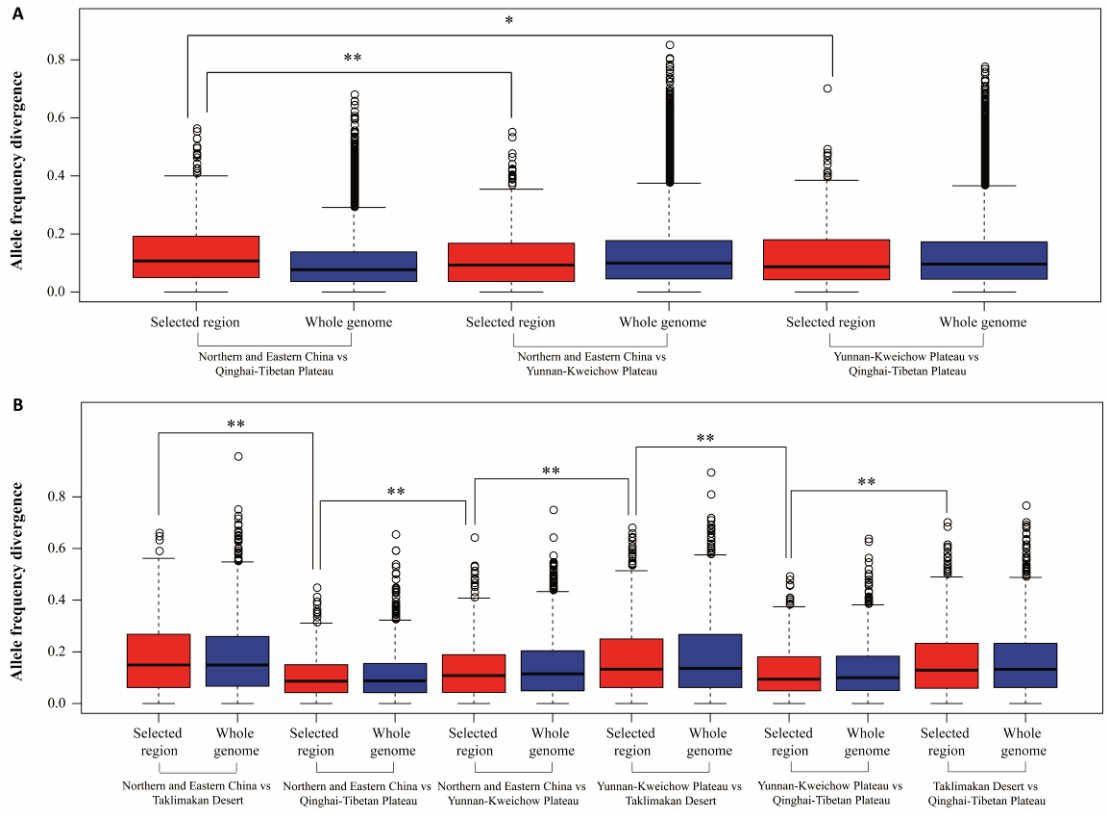

FIG. S9. Allele frequency differences of the large-effect SNPs (i.e., nonsynonymous SNPs) and genome-wide SNPs between the main groups of sheep breeds. (A) Allele frequency differences in the nonsynonymous mutations within the SNPs located in the candidate genes from the Tibetan sheep and allele frequency differences in genome-wide SNPs. (B) Allele frequency differences in the nonsynonymous mutations within the SNPs located in the candidate genes from sheep breeds from the Taklimakan Desert region and allele frequency differences in genome-wide SNPs. * and ** denote that differences between the two mean values are significant at the 0.05 and 0.01 levels (i.e., $P < 0.05$ and $P < 0.01$), respectively.

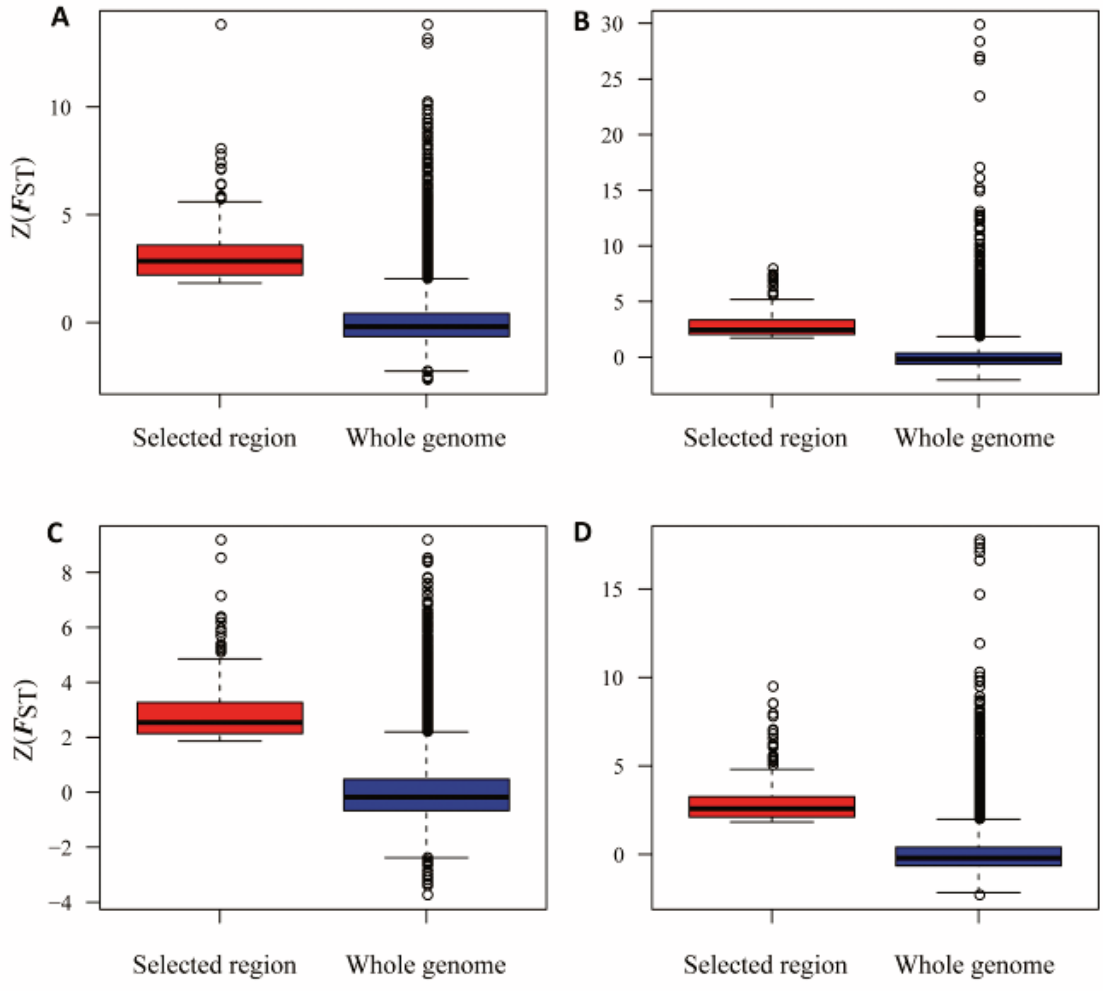

FIG. S10. Comparison between $Z(F_{ST})$ values for genomic regions that have undergone selective sweeps and $Z(F_{ST})$ values at the whole-genome scale. (A) Tibetan sheep from the plateau environment. (B) Sheep from high-altitude regions. (C) Sheep from the desert environment of the Taklimakan Desert region. (D) Sheep from arid zones. The upper, middle (within the box) and boundary lines of the boxes represent 25%, 50% (median value) and 75% of the $Z(F_{ST})$ values, respectively.

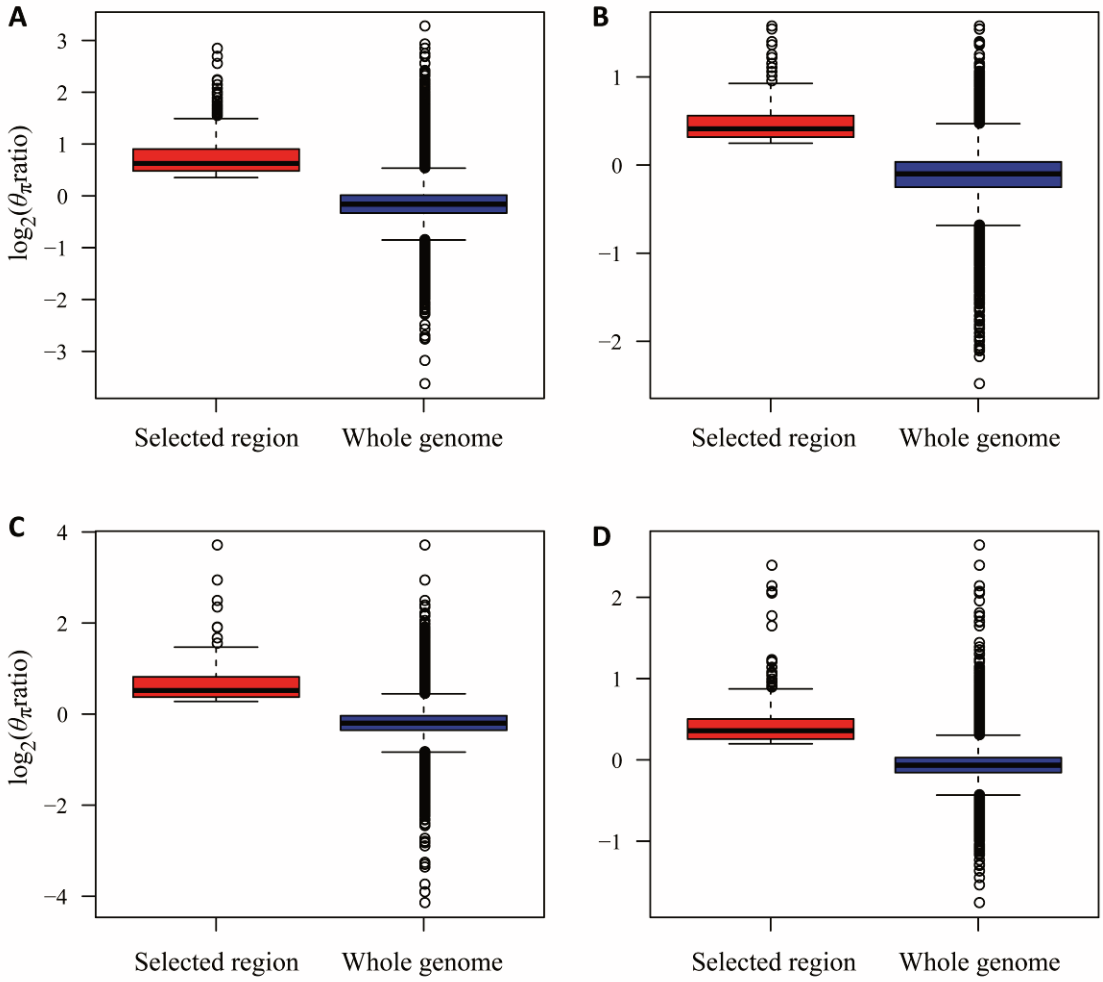

FIG. S11. Comparison between θ_π ratio values for genomic regions that have undergone selective sweeps and θ_π ratio values at the whole-genome scale. (A) Tibetan sheep from the plateau environment. (B) Sheep from high-altitude regions. (C) Sheep from the desert environment of the Taklimakan Desert region. (D) Sheep from arid zones. The upper, middle (within the box) and boundary lines of the boxes represent 25%, 50% (median value) and 75% of the θ_π ratio values, respectively.

FIG. S12. Male (left) and female (right) sheep from the Qinghai-Tibetan Plateau, Yunnan-Kweichow Plateau and Taklimakan Desert. The photos were obtained from Du (2011).

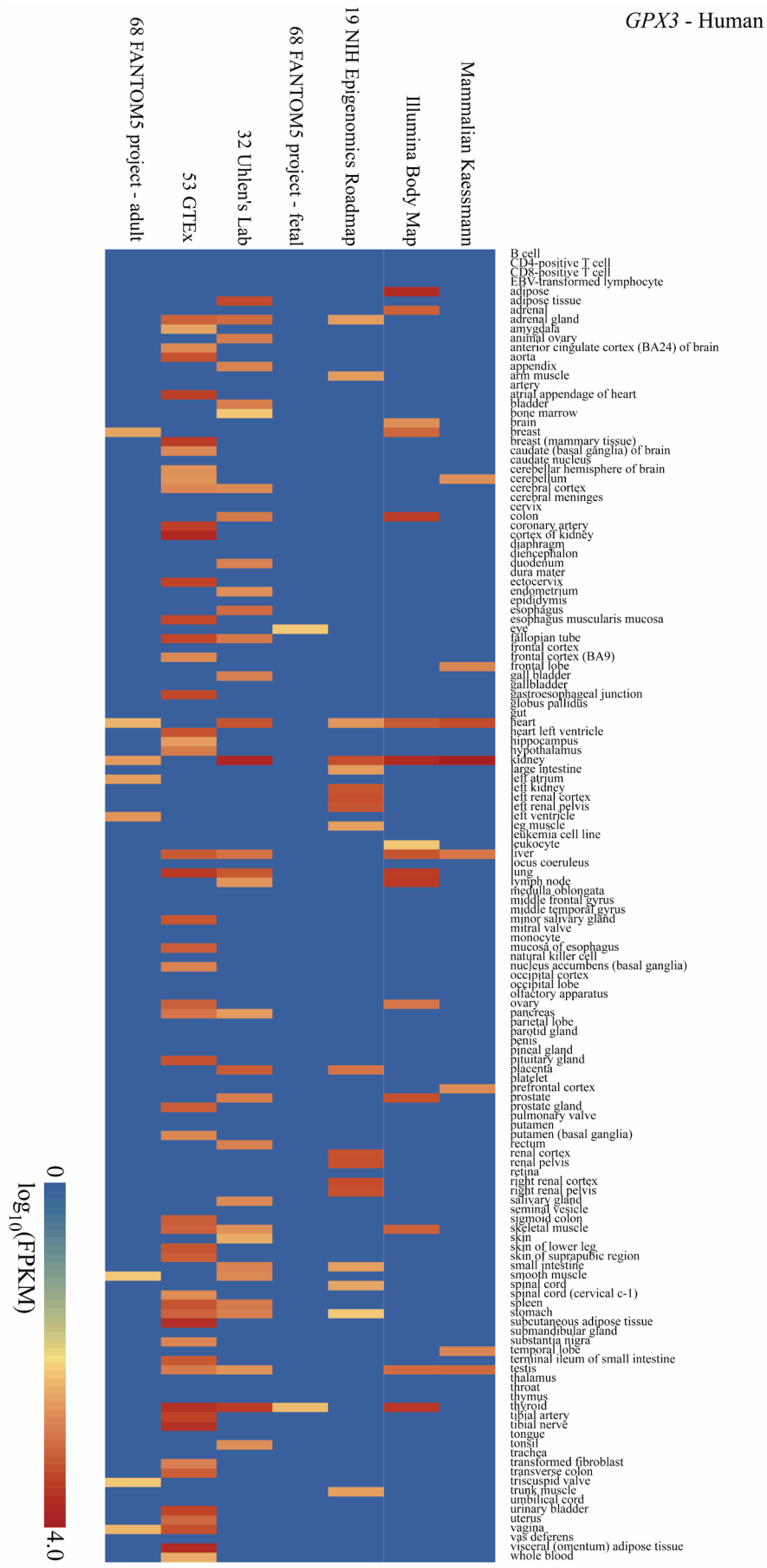
GPX3 - Human
Mammalian Kaessmann
Illumina Body Map
19 NIH Epigenomics Roadmap
68 FANTOM5 project - fetal
32 Uhlen's Lab
53 GTEx
68 FANTOM5 project - adult
B cell
CD4-positive T cell
CD8-positive T cell
EBV-transformed lymphocyte
adipose
adipose tissue
adrenal
adrenal gland
amygdala
animal ovary
anterior cingulate cortex (BA24) of brain
aorta
appendix
arm muscle
artery
atrial appendage of heart
bladder
bone marrow
brain
breast
breast (mammary tissue)
caudate (basal ganglia) of brain
caudate nucleus
cerebellar hemisphere of brain
cerebellum
cerebral cortex
cerebral meninges
cervix
colon
coronary artery
cortex of kidney
diaphragm
diencephalon
duodenum
dura mater
ectocervix
endometrium
epididymis
esophagus
esophagus muscularis mucosa
eye
fallopian tube
frontal cortex
frontal cortex (BA9)
frontal lobe
gall bladder
gallbladder
gastroesophageal junction
globus pallidus
gut
heart
heart left ventricle
hippocampus
hypothalamus
kidney
large intestine
left atrium
left kidney
left renal cortex
left renal pelvis
left ventricle
leg muscle
leukemia cell line
leukocyte
liver
locus coeruleus
lung
lymph node
medulla oblongata
middle frontal gyrus
middle temporal gyrus
minor salivary gland
mitral valve
monocyte
mucosa of esophagus
natural killer cell
nucleus accumbens (basal ganglia)
occipital cortex
occipital lobe
olfactory apparatus
ovary
pancreas
parietal lobe
parotid gland
penis
pineal gland
pituitary gland
placenta
platelet
prefrontal cortex
prostate
prostate gland
pulmonary valve
putamen
putamen (basal ganglia)
rectum
renal cortex
renal pelvis
retina
right renal cortex
right renal pelvis
salivary gland
seminal vesicle
sigmoid colon
skeletal muscle
skin
skin of lower leg
skin of suprapubic region
small intestine
smooth muscle
spinal cord
spinal cord (cervical c-1)
spleen
stomach
subcutaneous adipose tissue
submandibular gland
substantia nigra
temporal lobe
terminal ileum of small intestine
testis
thalamus
throat
thymus
thyroid
tibial artery
tibial nerve
tongue
tonsil
trachea
transformed fibroblast
transverse colon
triscuspid valve
trunk muscle
umbilical cord
urinary bladder
uterus
vagina
vas deferens
visceral (omentum) adipose tissue
whole blood
0
4.0
log10(FPKM)

FIG. S13. The expression levels of *GPX3* in different human tissues. The results are based on seven variable experiments deposited in the EBI Gene Expression Atlas database. The FPKM (fragments per kilobase of transcript per million mapped reads) value is used to measure expression levels.

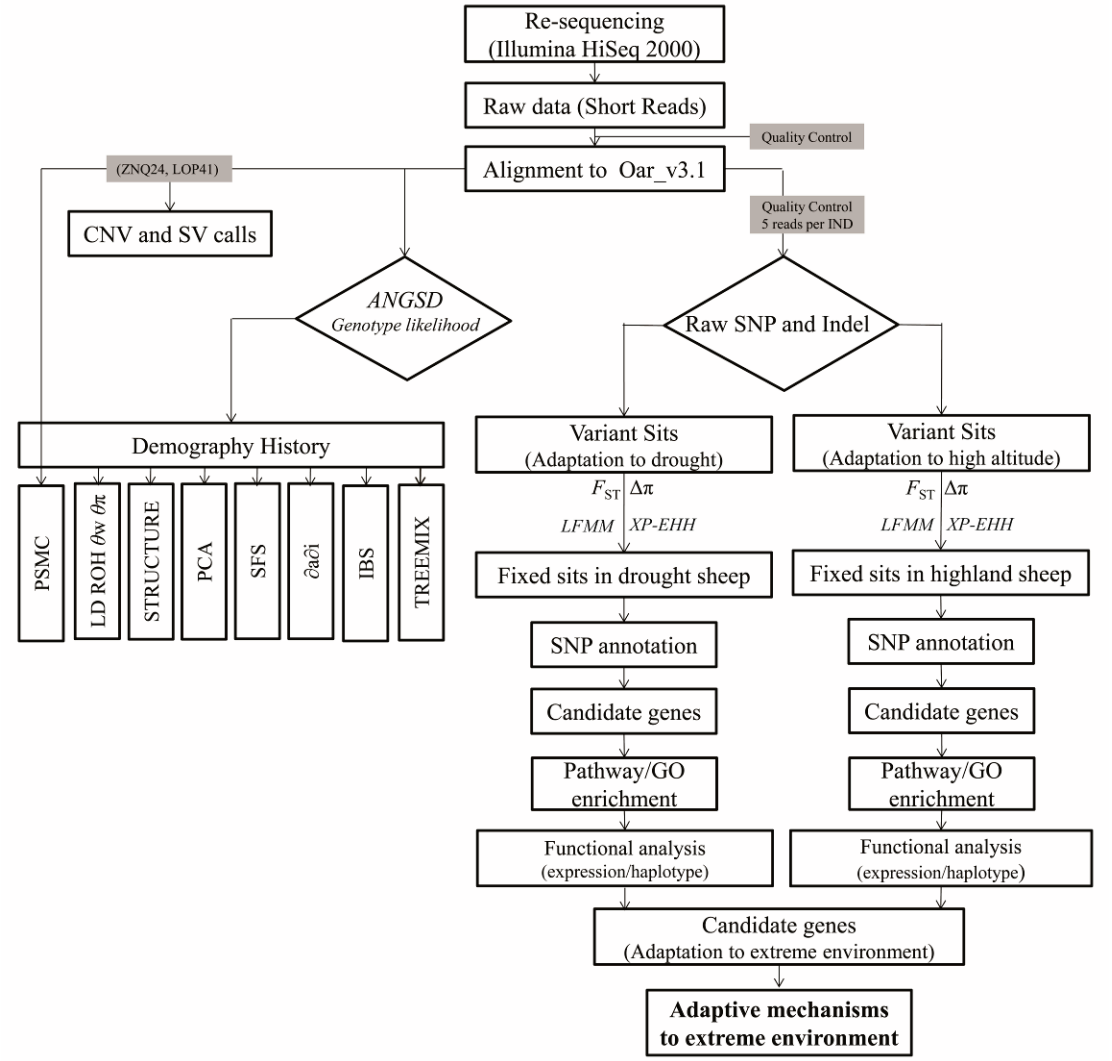

FIG. S14. Flow chart illustrating the overall pipeline of the genomic analyses conducted in this study.

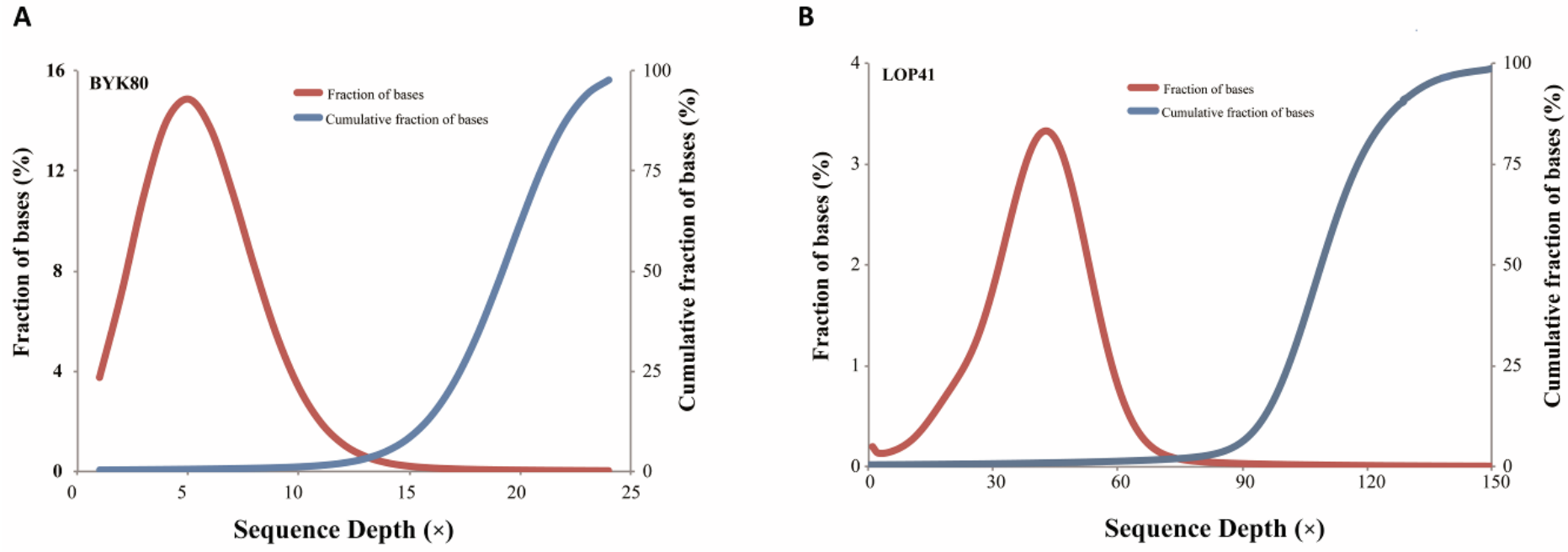

FIG. S15. Distribution of the sequencing depth and fraction of bases. (A) Depth and fraction distribution of BYK80, a representative sample with low-coverage (~5.74×) sequencing. (B) Depth and fraction distribution of LOP41, a representative sample with high-coverage (~41.95×) sequencing.

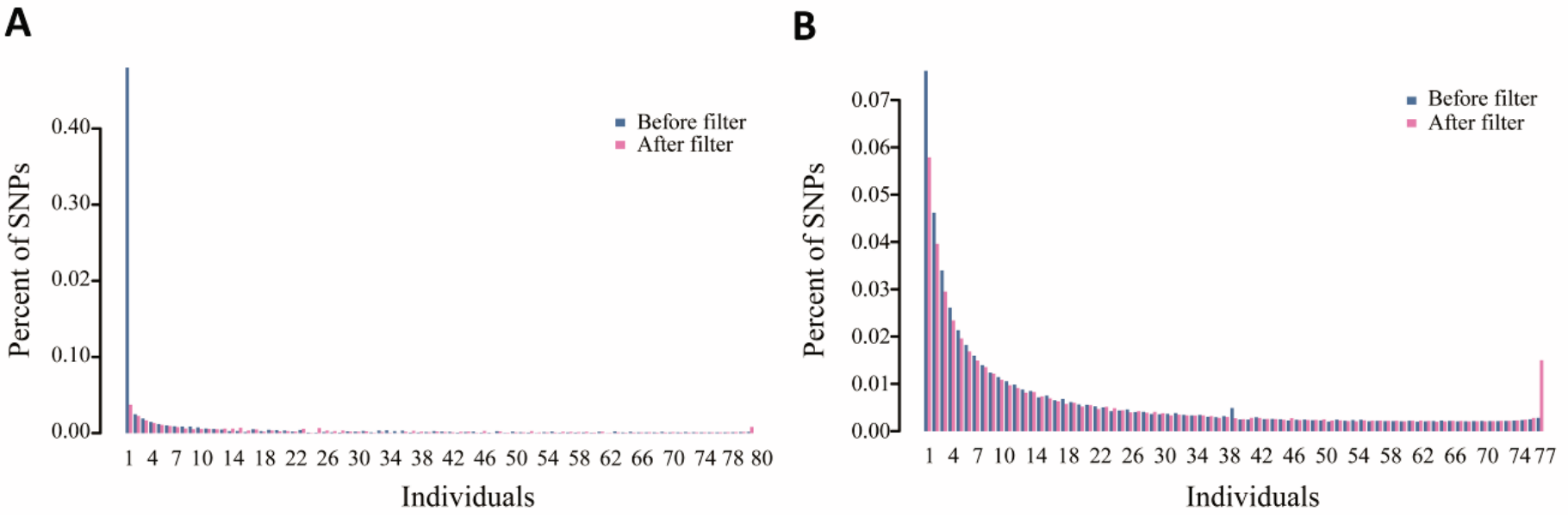

FIG. S16. Genome-wide site frequency spectrum (SFS) distribution. (A) Unfolded genome-wide SFS from all 80 samples. (B) Unfolded genome-wide SFS from the 77 native sheep. Different colors represent data before (blue) and after (red) imputation filtering of sites with correlations between the observed and imputed data smaller than 0.9.