Research

Selection Signature Mapping Reveals Genomic Regions Shaped by Artificial Selection in Jaunpuri Goats

A. K. Srivastav[1], Jaswant Singh[1], M. K. Verma[1]*, Pooja Chhabra[2], Ritika Gera[2], A. K. Verma[1], K. D. Singh[1], S. S. Kashyap[1], D. P. Srivastav[1], Anmol[1] and Reena Arora[2]

[1]*Acharya Narendra Deva University of Agriculture and Technology, Kumarganj, Ayodhya - 224229, Uttar Pradesh, India*

[2]*ICAR-National Bureau of Animal Genetic Resources, Karnal - 132001, Haryana, India*

ABSTRACT

Jaunpuri is an important indigenous dairy goat breed of northern India, but the genomic basis of its economically important traits remains poorly understood. This study investigated genome-wide selection signatures using four complementary approaches namely nucleotide diversity (π), Tajima's *D*, composite likelihood ratio (CLR) and integrated haplotype score (iHS). The π analysis identified 948 candidate genes enriched for transcriptional regulation and chromatin organization. Tajima's *D* detected 848 candidate genes with similar functional enrichment, whereas the CLR approach identified 6,410 selective sweep regions enriched for axon guidance and cell-adhesion pathways. The iHS analysis highlighted candidate regions containing *LDB2* and *DCHS2*, which overlap reported milk production QTLs, together with haemoglobin-related genes. Fourteen candidate genes consistently identified by π, Tajima's D, iHS and CLR analyses, represent the most robust candidate loci associated with the genetic architecture of the Jaunpuri goat. These genes are known to be involved in mammary gland development, lactation and reproductive biology. Their detection may reflect historical or ongoing selection for enhanced milk production and reproductive performance in Jaunpuri goats.

Keywords: Goat, Selection signatures, Whole genome sequencing

***Corresponding author:** manojvet550@gmail.com

Received: July 22, 2026
Accepted: July 27, 2026

INTRODUCTION

In many developing nations, goats are known to be the best livestock species for improving the economic health of marginal and landless farmers (Verma, 2018). The amount to which this priceless genetic resource contributes to the livelihood of the impoverished is not well appreciated and its significance is undervalued. Goats play an important role in the animal agriculture economies of various nations for generating income, storing capital, creating jobs and providing food for the home. Due to their smaller size, they are simpler to handle, take up less room and can be managed by women and children. Goat milk is an excellent choice for people who are allergic to or intolerant to cow milk because it is rich in nutrients and often simpler to digest (Sen *et al*., 2004). However, because goat milk yield is less than cow milk, it is more expensive. Therefore, to produce goats with better milk yields, it is essential to understand the genetics and determine the genes linked to milk production. The distinctive indigenous Jaunpuri goat, is mostly found in the eastern part of Uttar Pradesh, India. The breed is valued for milk as well as meat. The medium-sized Jaunpuri goat is distinguished by its robust body, drooping ears and tendency to give birth to twins (Patel, 2017). To comprehend the unique genetic characteristics of Jaunpuri goat and discover the genetic effects of the selection forces they have encountered, it is essential to identify the markers of selection (Verma, 2014). Since the genome of Jaunpuri goat has not been extensively studied, this study was undertaken to identify signatures of selection using whole genome resequencing data. The results will enable a thorough understanding of their genetic makeup and identify candidate genes underlying the genomically selected regions. These findings will serve as a starting point for more research on the genetic underpinnings of significant economic characteristics for future improvement of Jaunpuri goats.

MATERIALS AND METHODS

Sample collection and genomic DNA isolation

Blood samples were obtained from 10 unrelated Jaunpuri goats originating from their native breeding tract in eastern Uttar Pradesh, India. The animal handling and sample collection were performed in accordance with approved ethical guidelines. Genomic DNA was isolated from whole blood using the conventional phenol-chloroform extraction protocol.

The concentration and purity of the extracted DNA were determined spectrophotometrically, while DNA integrity was verified by agarose gel electrophoresis before library preparation.

Whole-genome resequencing and library preparation

Whole-genome resequencing was carried out to generate high-quality genomic data for the identification of genome-wide sequence variants. Approximately 1 µg of genomic DNA from each individual was randomly fragmented to produce inserts of approximately 300-400 bp. Fragmented DNA was subjected to end repair, addition of adenine overhangs and ligation of indexed sequencing adapters. Adapter-ligated fragments were subsequently amplified by PCR to enrich the sequencing libraries. Library concentration was measured using a Qubit Fluorometer (Invitrogen, Life Technologies, Grand Island, NY, USA) and fragment size distribution was evaluated using the High Sensitivity DNA Kit on an Agilent TapeStation system. Libraries with satisfactory quality were pooled at equimolar concentrations and used for cluster generation. Paired-end sequencing (2 × 150 bp) was then performed on the Illumina NovaSeq 6000 system.

Bioinformatics analysis

FastQC v.0.11.9 was used to perform a per-base sequence quality check on the raw fastq files (Andrews 2010). Adapter sequences and bases with a Phred score of less than 20 were cut using Cutadapt (Mohideen *et al*., 2020). The Burrows-Wheeler Aligner tool (Li and Durbin, 2010) used the bwa mem command to map the trimmed sequences to the goat reference genome, GCF_001704415.2_ARS1.2_genomic.fna. Following mapping, the resultant Sam data were converted into BAM files. These BAM files were then sorted and compiled using Samtools (Li *et al*., 2009). BCFTOOLS was used to create a vcf file (Li, 2011). Variant quality control was carried out using vcftools (Danecek *et al*., 2011).

Detection of selection signatures

According to Nei and Li (1979), nucleotide diversity is the average number of nucleotide changes per site between two DNA sequences taken from a population; larger values signify greater diversity. The nucleotide diversity was estimated using the program vcftools (Danecek *et al*., 2011) with window size of 50 kb that contained at least fifteen SNVs. The 2% threshold was used to evaluate statistical significance. vcftools software was used to calculate Tajima's D statistics within the population at window sizes of 50 kb. Only windows with fifteen or more markers were taken into account. After applying a 4% threshold, the lowest D value found was −2.67499. Tajima's D statistic combines two measures of genetic diversity: the number of segregating sites (S) and the average number of pairwise nucleotide variations between sequences in a sample (Tajima, 1989). Tajima's D test is used to determine if a DNA sequence is evolving randomly or under a non-random process, such as selection, population contraction, population expansion or introgression. According to the conventional neutral evolutionary paradigm, Tajima's D statistic value should be zero. Deviations from zero, however, could be a sign of selection or non-neutral forces. Positive Tajima's D values indicate balanced selection, which represents an abundance of intermediate allele frequencies, while negative values imply the fixation of alleles or rare alleles (Tajima, 1989). Selscan was used for Integrated Haplotype Score (iHS) analysis (Szpiech and Hernandez, 2014). In order to account for the reliance of iHS on allele frequency, the resulting raw iHS values were adjusted by normalizing scores across allele frequency using the norm tool. Candidate genomic areas under recent positive selection were identified using the normalized absolute iHS values and SNPs with high iHS values were regarded as possible selection signals. SNPs with iHS > 3 were selected as significant and kept for further analysis. SweeD (Chu *et al*., 2023) was used to perform Composite Likelihood Ratio (CLR) analysis in order to determine whether genomic areas were impacted by recent positive selection. To estimate CLR values throughout the genome, the filtered VCF file was used as input. At predetermined grid locations, SweeD compares the likelihood of a neutral model with that of a selective sweep model to scan the genome. Under recent positive selection, genomic areas with the highest CLR values were regarded as candidate regions. The top 1% of the genome-wide CLR distribution contained SNPs (or genomic windows) that were considered significant and chosen for further analysis. To display the chromosomal distribution of selection statistics, Manhattan plots were constructed with the qqman function in R. To find high-confidence genes under selection, all the three methods were evaluated for overlap. Genes identified by atleast 2 or more methods were considered promising candidates and visually represented using Venny2.1.0 (Oliveros, 2015). DAVID bioinformatics resources (Sherman *et al*., 2022) were used to perform gene ontology and define the biological functions of candidate genes.

RESULTS AND DISCUSSION

Whole-genome sequencing of 10 samples of Jaunpuri goat produced high quality sequencing data with consistently good read mapping rates to the reference genomes. Jaunpuri goat samples yielded clean reads ranging from 468 to 789 million, with more than 99%

reads aligned to the reference genome (Table 1). After variant calling and quality filtering, 30,380,017 high-confidence single nucleotide polymorphisms (SNPs) were discovered in Jaunpuri goats. The high mapping percentages and number of discovered SNPs imply that the sequencing data was of good quality and providing a strong dataset for downstream analysis. The chromosome wise distribution of SNPs in Jaunpuri goat is shown in Fig 1.

Table 1: Whole genome sequencing read summary of Jaunpuri goats

Sample	Total reads	Mapped reads	Mapping percent	GC content (%)
Jaunpuri goat 1	310462204	309897075	99.82%	44.72
Jaunpuri goat 2	315964542	315445902	99.84%	42.96
Jaunpuri goat 3	321489756	321059265	99.87%	42.21
Jaunpuri goat 4	363138501	362637618	99.86%	43.47
Jaunpuri goat 5	408606279	408041021	99.86%	44.72
Jaunpuri goat 6	246775902	246139631	99.74%	42.96
Jaunpuri goat 7	271812117	271034834	99.71%	43.97
Jaunpuri goat 8	324433366	323891859	99.83%	42.21
Jaunpuri goat 9	234256924	233568538	99.71%	43.47
Jaunpuri goat 10	394464516	393775148	99.83%	43.97

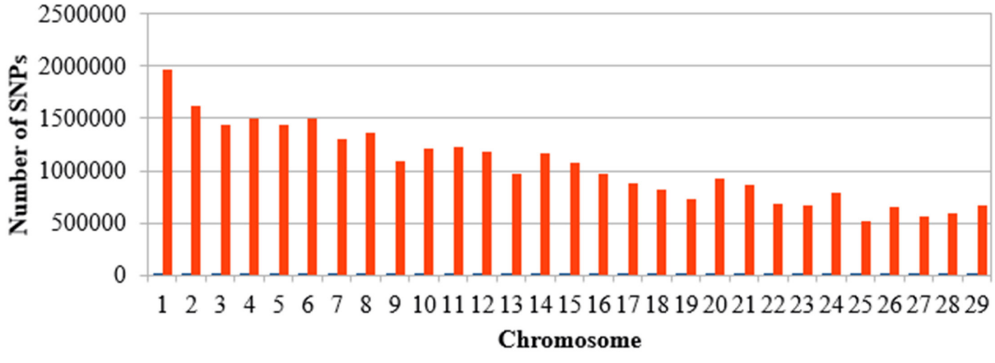

Fig. 1: *Distribution of SNPs across the goat chromosomes.*

Genome-wide selection signature analysis in Jaunpuri goats, employing four complementary approaches (nucleotide diversity (π), Tajima's D, composite likelihood ratio (CLR), and integrated haplotype score (iHS), identified thousands of candidate genomic regions under putative recent positive selection, several of which overlapped with goat QTL database entries for economically important traits. The π-based scan (mean $\pi = 1.43 \times 10^{-3}$) flagged 949 low-diversity windows harboring 948 candidate genes including *HOXA5–HOXA10, HOXD8–HOXD13, TBX15, LCORL, FOXO3, JAK1, NR5A1* and *ZEB2*, enriched for transcriptional regulation and chromatin organization. These were clustered on chromosomes 2, 3, 6, 7, 14, 17, 19 and 22, suggesting these loci may have been targets of selection for lactation. Tajima's D analysis (mean = −1.80) similarly identified 848 candidate genes enriched for chromatin- and DNA-binding functions, while the CLR scan resolved 6,410 sweep regions (6,409 candidate genes) concentrated on chromosomes 6, 15 and 3, enriched for axon guidance and cell-adhesion processes. The iHS analysis identified 949 candidate regions, most notably *LDB2* (Chr6) and *DCHS2* (Chr17), both mapping to milk production QTLs, alongside enrichment for oxygen-transport and haemoglobin-related genes (e.g., *HBBC, BPGM*). These convergent signatures across independent statistical approaches indicate that selection in the Jaunpuri breed has acted on multiple genomic regions associated with transcriptional regulation, lactation and other production-relevant traits.

Genome-wide Manhattan plots from four intra-population selection signature analyses (nucleotide diversity (π), Tajima's D, composite likelihood ratio (CLR) and integrated haplotype score (iHS) showed the genomic distribution of potential selection signals across all autosomes (Fig. 2). Reduced genetic diversity and negative Tajima's D values, respectively, were highlighted by the π and Tajima's D plots, which are signs of recent positive selection. On the other hand,

genomic areas with high absolute iHS values and higher likelihood ratios were found using the CLR and iHS Manhattan plots, indicating broad haplotype homozygosity and selective sweeps. Candidate regions that were scattered across several chromosomes and exceeded the predetermined thresholds for each approach were highlighted, indicating that selection has affected numerous genetic loci.

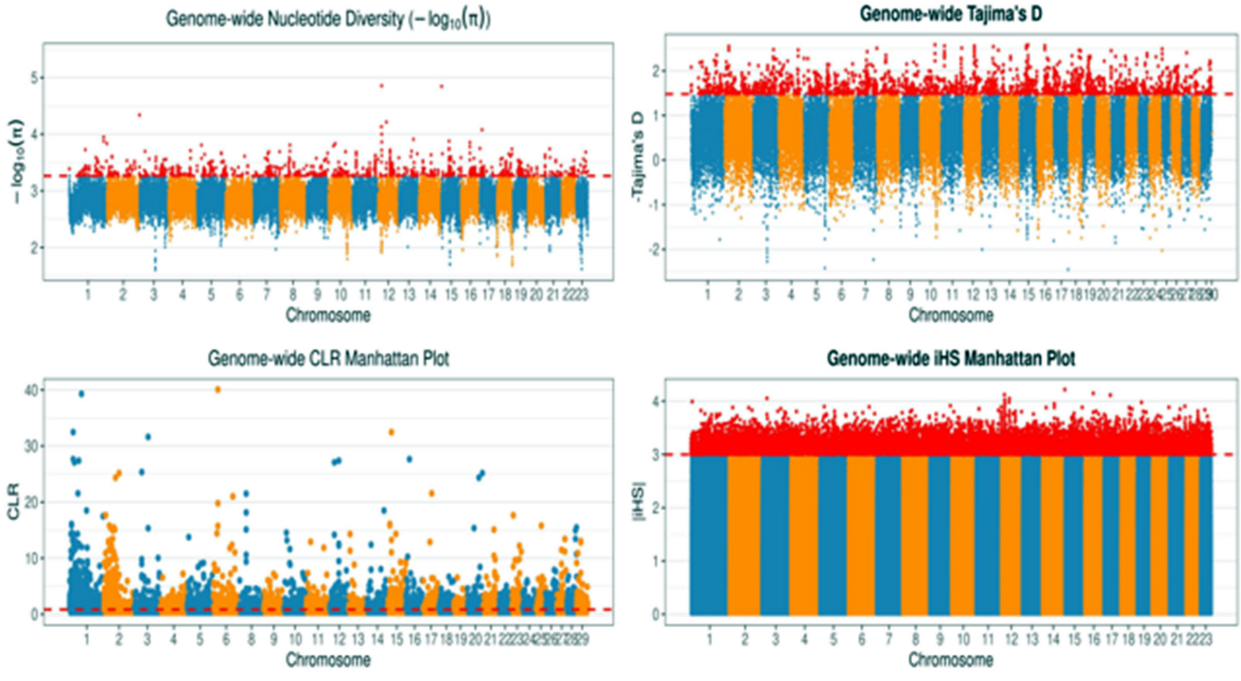

Fig. 2: *Genome-wide distribution of candidate selection regions identified by nucleotide diversity (π), Tajima's D, CLR and iHS.*

To identify strong positive selection signatures, overlapping genes across all the four methods were identified (Fig. 3). A total of 29 genes were found common in Tajima's D, iHS and CLR whereas 8 were common in π, Tajima's D and iHS, 20 were common in Tajima's D, iHS and CLR. Fourteen genes were observed to be common all the four intra population tests. Since they are corroborated by independent statistical methods, genes found by several testing offer more convincing proof of actual selection targets.

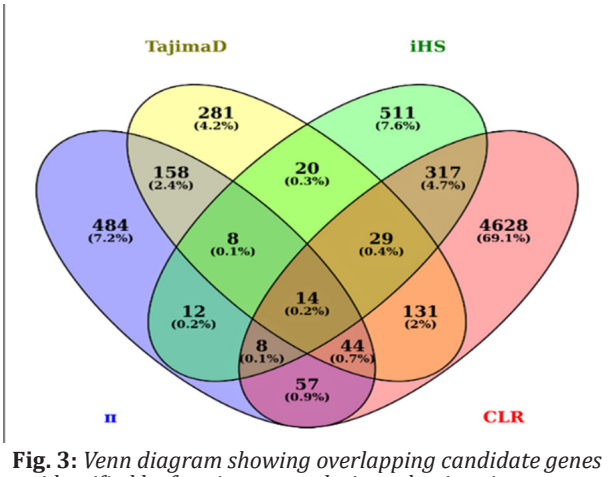

Fig. 3: *Venn diagram showing overlapping candidate genes identified by four intra-population selection signature analyses*

Comparison of the four selection signature approaches identified 14 candidate genes common to all methods, including *HDAC9, ATP8B4, DUS2, SUGCT, GNB1, ATXN7, ERC2, LOC106502391, LOC102181832, KMT2C, ZBTB20, NF1, BOP1* and *TTN*. These genes consistently identified by π, Tajima's D, iHS and CLR analyses, represent the most robust candidate loci associated with the genetic architecture of the Jaunpuri goat. Among the identified candidate genes, *ZBTB20* has been shown to play an essential role in mammary gland function. Targeted deletion of *ZBTB20* disrupts prolactin expression and lactotrope development, leading to impaired lactogenesis, failure of milk secretion, and neonatal mortality in knockout mice (Han *et al*., 2022). Given the central role of prolactin signaling in the initiation and maintenance of lactation, the identification of *ZBTB20* within our candidate loci supports its potential involvement in regulating milk production and/or lactation onset. This warrants further validation through expression profiling and association analyses in mammary tissue.

Similarly, *KMT2C*, a histone H3K4 methyltransferase, has established relevance to mammary gland biology. Using a mammary-specific conditional knockout model, Gala *et al*. (2018) demonstrated that *KMT2C* is indispensable for estrogen receptor-α (ERα)-mediated

pubertal mammary gland development and epithelial cell proliferation, in addition to its well-recognized role in breast tumorigenesis. *HDAC9*, a class II histone deacetylase, may also contribute to mammary function through epigenetic regulation (Lapierre *et al*., 2016). Histone deacetylase-mediated chromatin remodeling has been implicated in the transcriptional regulation of casein and other milk protein genes. However, a direct functional association between *HDAC9* and milk production traits has not yet been established and therefore should be interpreted with caution.

NF1, a well-characterized tumor suppressor that negatively regulates RAS signaling, has been associated with an increased risk of aggressive mammary tumorigenesis in both humans and rodent models (Dischinger *et al*., 2018), indicating an important role in mammary epithelial growth and homeostasis. This gene has also been identified as a positional candidate gene associated with litter size in a multi-breed sheep GWAS, particularly in the highly prolific Finnsheep breed, where it was implicated in folliculogenesis and luteinizing hormone (LH)-signaling pathways (Xu *et al*., 2018). Further, *BOP1* gene has been associated with milk fat percentage in cattle (Dong et al., 2026).

These findings indicate that several of the candidate genes identified in the present study are functionally linked to mammary gland development, lactation or reproductive biology. Their detection may reflect historical or ongoing selection for enhanced milk production and/or reproductive performance in Jaunpuri goats. Nevertheless, functional validation through gene expression analyses, fine-mapping and experimental studies in future may confirm their biological roles and causal relationships with economically important traits.

ACKNOWLEDGEMENTS

We are grateful to the Director, ICAR- National Bureau of Animal Genetic Resources (NBAGR), Karnal, and Indian Council of Agricultural Research (ICAR), New Delhi for providing the necessary facilities. The financial support by the ICAR-Consortium Research Platform- Genomics is gratefully acknowledged.

REFERENCES

Andrews S. 2014. FastQC: A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Chu J, Ma Y, Song H, Zhao Q, Wei X, Yan Y, Fan S, Zhou B, Li S, and Mou C. 2023. The genomic characteristics affect phenotypic diversity from the perspective of genetic improvement of economic traits. *iScience*, 26:107043. doi:10.1016/j.isci.2023.107043.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, and 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158. doi:10.1093/bioinformatics/btr330.

Dischinger PS, Tovar EA, Essenburg CJ, Madaj ZB, Gardner EE, Callaghan ME, Turner AN, Challa AK, Kempston T, Eagleson B, Kesterson RA, Bronson RT, Bowman MJ, Graveel CR and Steensma MR. 2018. NF1 deficiency correlates with estrogen receptor signaling and diminished survival in breast cancer. *Breast Cancer*, 4(1):29. doi:10.1038/s41523-018-0080-8.

Dong C, Ma P, Tang Y, Han H, Chen S, Yang J, Sun Z, Wang W, Yu Y and Li S. 2026. Genetic parameter estimation and fine-mapping of milk-production traits and somatic cell score in Chinese Simmental cattle. *Journal of Dairy Science*, 109(4):4195–4208. doi:10.3168/jds.2025-27039.

Gala K, Li Q, Sinha A, Razavi P, Dorso M, Sanchez-Vega F, Chung YR, Hendrickson R, Hsieh JJ, Berger M, Schultz N, Pastore A, Abdel-Wahab O, and Chandarlapaty S. 2018. KMT2C mediates the estrogen dependence of breast cancer through regulation of ERα enhancer function. *Oncogene*, 37(34):4692–4710. doi:10.1038/s41388-018-0273-5.

Han Q, Yan X, Ye Y, Han L, Ma X, Wang T, Cao D, and Zhang WJ. 2022. ZBTB20 regulates prolactin expression and lactotrope function in adult mice. *Endocrinology*, 163(12):bqac181. doi:10.1210/endocr/bqac181.

Lapierre M, Linares A, Dalvai M, Duraffourd C, Bonnet S, Boulahtouf A, Rodriguez C, Jalaguier S, Assou S, Orsetti B, Balaguer P, Maudelonde T, Blache P, Bystricky K, Boulle N, and Cavaillès V. 2016. Histone deacetylase 9 regulates breast cancer cell proliferation and the response to histone deacetylase inhibitors. *Oncotarget*, 7(15):19693–19708. doi:10.18632/oncotarget.7564.

Li H and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760. doi:10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993. doi:10.1093/bioinformatics/btr509.

Mohideen AMSH, Johansen SD, and Babiak I. 2020. High-throughput identification of adapters in single-read

sequencing data. *Biomolecules*, 10:878. doi:10.3390/biom10060878.

Nei M, and Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10):5269–5273. DOI: 10.1073/pnas.76.10.5269.

Oliveros JC. 2007–2015. Venny: An interactive tool for comparing lists with Venn's diagrams.

Patel AK, Singh VK, Singh J, Singh NS, and Verma MK. 2017. Phenotypic characterization of Jaunpuri goat of Uttar Pradesh. *Journal of Livestock Biodiversity*, 7(1).

Sen AR, Santra A, and Karim SA. 2004. Carcass yield, composition and meat quality attributes of sheep and goat under semi-arid conditions. *Meat Science*, 66(4):757–763. doi:10.1016/S0309-1740(03)00182-4.

Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, and Chang W. 2022. DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*, 50(W1):W216–W221. doi:10.1093/nar/gkac194.

Szpiech ZA and Hernandez RD. 2014. Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, 31(10):2824–2827. doi:10.1093/molbev/msu211.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595. doi:10.1093/genetics/123.3.585.

Verma NK. 2018. Goat genetic diversity of India—An overview. *Journal of Livestock Biodiversity*, 8(1) :1-10.

Xu SS, Gao L, Xie XL, Ren YL, Shen ZQ, Wang F, Shen M, Eyþórsdóttir E, Hallsson JH, Kiseleva T, Kantanen J, and Li MH. 2018. Genome-wide association analyses highlight the potential for different genetic mechanisms for litter size among sheep breeds. *Frontiers in Genetics*, 9:118. doi:10.3389/fgene.2018.00118