BF-1628
[1-6]

SHORT COMMUNICATION Indian Journal of Animal Research

Domestication and Genetic Diversity Analysis of Several Sheep Breeds in China

Bujun Mei[1,2], Zhihua Wang[3], Yanping Liu[2], Rong Liu[4]

10.18805/IJAR.BF-1628

ABSTRACT

Sheep is the main livestock species in northern China. Under diverse climatic conditions, a gene pool of local sheep resources is now forming. There are many sheep breeds in northern China, including local breeds and hybrid breeds. However, little is known about the genetic background and genetic variation of local sheep populations in northern China. Therefore, our aims is to provide a characterization of the population structure and genetic relationship within sheep breeds in northern China, and to study its genetic relationship with other sheep breeds. We collected blood samples from 390 sheep for genetic diversity analysis using Reduced-Representation Genome Sequencing (RRGS). This study used bioinformatics methods to analyze linkage disequilibrium, population structure, evolutionary tree, selective sweep *etc*. The mixed model revealed in this study may be formed in the process of sheep breeding, partially overlapping with historical and archaeological findings. New information on the domestication and genetic variety of sheep is provided by this study, which will help sheep in the future be improved through genomics.

Key words: Domestication, Genetic Diversity, Sheep, Whole-genome resequencing.

This study uses whole genome sequencing (WGS) to re-sequence different individuals or tissues of sheep with known genomic information to discover differences between individuals. Through this method, genome-wide SNP and InDel molecular markers were developed, and SNP/InDel markers were used for population genetics research.

Animals and genome sequencing

Two sheep populations-NKY and THT - were analyzed in this study. A total of 5 mL jugular vein blood was collected from 390 individuals (185 NKY and 205 THT population) in 2022 at Bayannaoer Academy of Agriculture and Animal Husbandry Sciences, Inner Mongolia, China. After the sample ‘s genomic DNA was qualified, the DNA sequence was fragmented by ultrasound to form a random fragment.

Bioinformatic analysis workflow

After the Illumina NovaSeqTM sequencing data (Raw Data) is offline, the quality control of the offline data is performed, and the low-quality data is filtered to obtain high-quality data (Clean Data). Clean Data was aligned to the reference genome sequence using BWA software(Li and Durbin 2009)to obtain the sequence location attribution (*i.e*., BAM file). BAM files were corrected using GATK ‘s Best Practices process(McKenna et al. 2010), and SNP and Small InDel markers were detected. The functional annotation information of SNP and InDel was obtained by using SNPEff software (Cingolani et al. 2012) and gene prediction information of reference genome. Based on the obtained SNP and InDel molecular markers, genetic diversity, population structure, linkage disequilibrium and selective sweep were further studied (see Table 1 for bioinformatics analysis tools).

Linkage disequilibrium

In a population, the frequency of simultaneous inheritance of two genes at different loci is significantly higher than the expected random frequency, which is called linkage disequilibrium. The minimum genetic unit of species can be obtained by linkage disequilibrium analysis.

The linkage of SNP in all samples was analyzed by SNP combination on the same chromosome. Natural populations are represented by a linkage disequilibrium coefficient (r^2). The closer r^2 is to 1, the stronger the linkage. The distance between SNPs in the genome was fitted to r^2. Generally, the closer the distance between SNPs, the larger

---

[1]Laboratory of Small Ruminant Genetics, Breeding and Reproduction, College of Animal Science and Technology, Huazhong Agricultural University, Wuhan430070, People’s Republic of China.
[2]Department of Agriculture, Hetao College, Bayannur, 015000, People’s Republic of China.
[3]Department of Civil Engineering, Hetao College, Bayannur, 015000, People’s Republic of China.
[4]Inner Mongolia Autonomous Region Agriculture and Animal Husbandry Technology Extension Center, Huhhot, 010000, People’s Republic of China.

Corresponding Author: Bujun Mei, Department of Agriculture, Hetao College, Bayannur, 015000, People’s Republic of China. Email: meibujun@163.com

How to cite this article: Mei, B., Wang, Z., Liu, Y. and Liu, R. (2023). Domestication and Genetic Diversity Analysis of several Sheep Breeds in China. Indian Journal of Animal Research. doi:10.18805/IJAR.BF-1628.

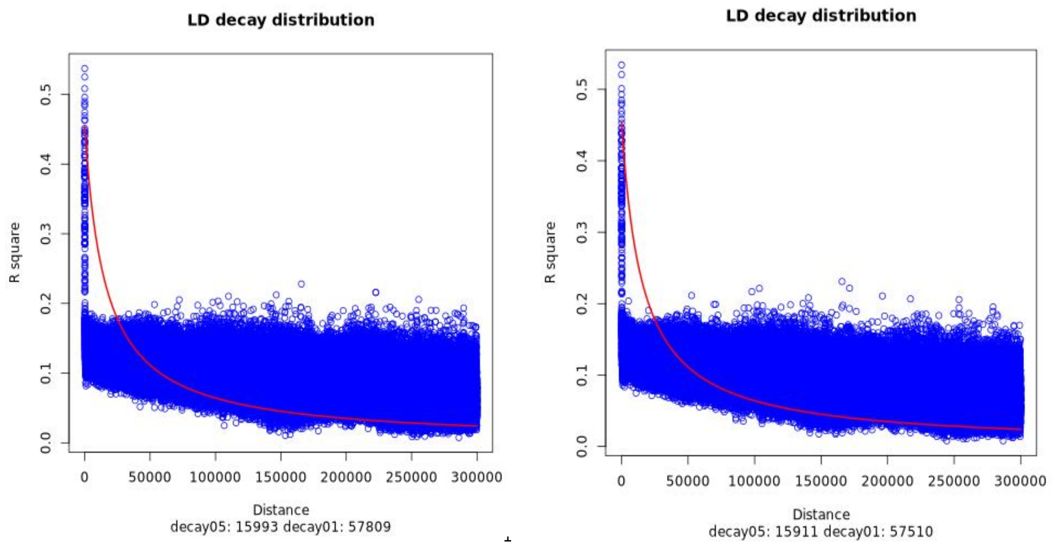
Submitted: 09-01-2023 **Accepted:** 29-04-2023 **Online:** 23-06-2023

the r^2, and the farther the distance between SNPs, the smaller the r^2. Generally, the distance between the corresponding sites when r^2 decays to half is used as the value of LD-decay. The longer the LD-decay, the slower the LD decay, the greater the probability of linkage between the SNP of the representative species. The shorter the LD-decay, the faster the LD decay, the smaller the probability of linkage between the SNP of the representative species. It is generally believed that the species with fast LD decay are relatively primitive. The LD-decay of each subpopulation is shown in Fig 1.

Genome-wide linkage disequilibrium (LD) research showed that the NKY and THT populations had nearly the same level of LD and decay, with a lower R^2 correlation coefficient becoming stable at a distance of about 100 kb.

Population structure

Principal component analysis (PCA) is a pure mathematical operation method, which can select a small number of important variables through linear transformation of multiple related variables. Based on SNP, principal component analysis (PCA) was performed by GCTA software to obtain the principal component clustering of 36 samples. Through PCA analysis, we can know which samples are relatively close and which samples are relatively distant, which can assist evolutionary analysis.

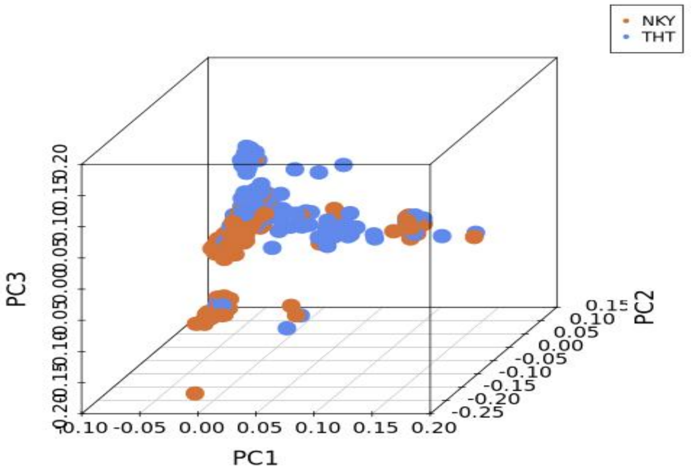


Left: NKY population. Right: THT population. The abscissa represents the SNP spacing and the ordinate represents the r^2 value. The corresponding SNP interval when r^2 decays to half is LD-decay.

Fig 1: LD of the two populations.

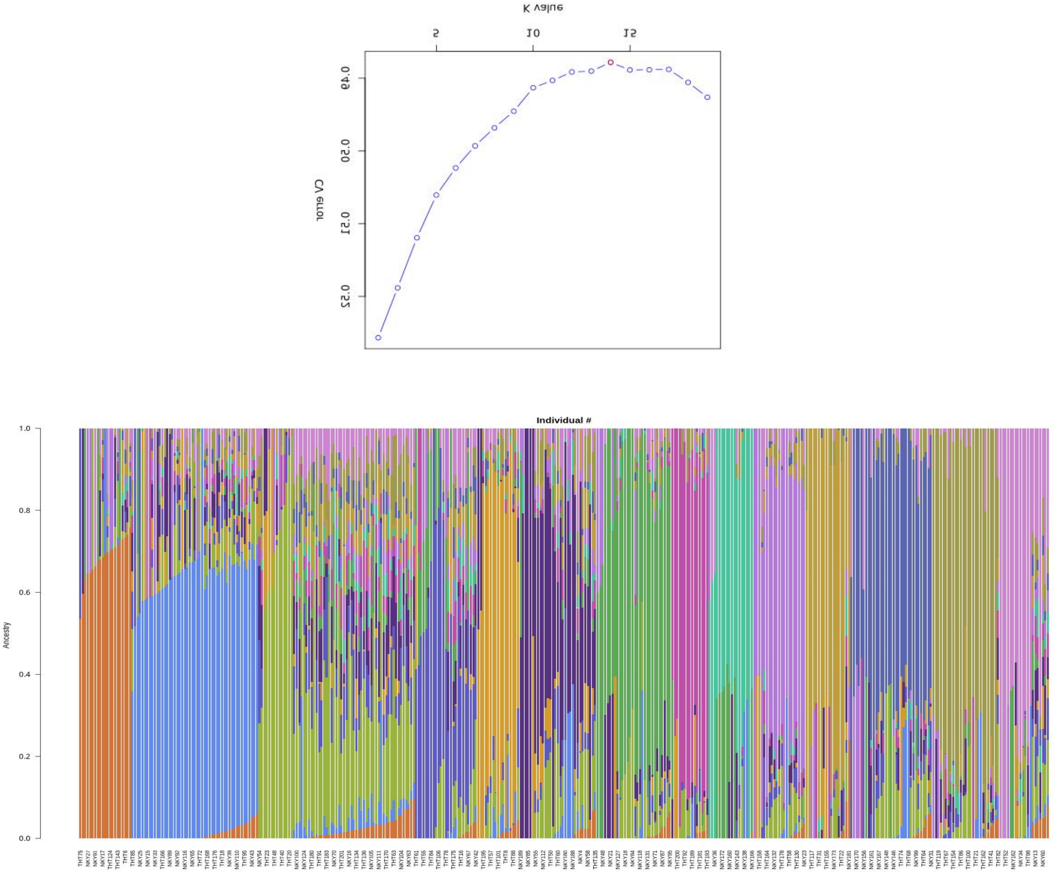
Table 1: Bioinformatics analysis tools.

Tool	Version	Purpose	Availability
Fastp	0.23.0	Raw data quality control	https://github.com/OpenGene/fastp
BWA	0.7.17-r1198-dirty	Mapping to reference gemome	http://bio-bwa.sourceforge.net/
GATK/SAMtools/sentieon	4.2.4.1/1.14/2021.12	SNP, InDel detection	https://software.broadinstitute.org/gatk/http://www.htslib.org/https://www.sentieon.com/products/
SnpEff	5.0e	Variation functional annotation	http://snpeff.sourceforge.net/
Blast	2.12.0+	Gene functions annotation	https://blast.ncbi.nlm.nih.gov/Blast.cgi
RAxML/Fast tree	v8.2/2.1.11	Evolutionary tree analysis	https://cme.h-its.org/exelixis/web/software/raxml/index.htmlhttp://www.microbesonline.org/fasttree/
Admixture	1.3.0	Population structure analysis	http://software.genetics.ucla.edu/admixture/
GCTA	1.93.2	PCA	http://cnsgenomics.com/software/gcta/#Overview
Vcftools	v0.1.16	selective sweep	https://vcftools.github.io/
PopLDdecay	v3.41	LD-decay	https://github.com/BGI-shenzhen/PopLDdecay



The samples were clustered into three dimensions by PCA analysis. PC1 represents the first principal component and PC2 represents the second principal component. PC3 represents the third principal component. A point represents a sample, a color represents a group. Generally, individuals with close genetic relationship will gather together.

Fig 2: PCA clustering diagram of samples.



Each color in the above figure represents a population and each row represents the case when all individuals correspond to a cluster value. For example, K = 14 is the population structure that represents all samples when divided into two genetic components; the following figure shows the CV error value for each K value. When K is 14, CV error is the smallest, which is the best K value.

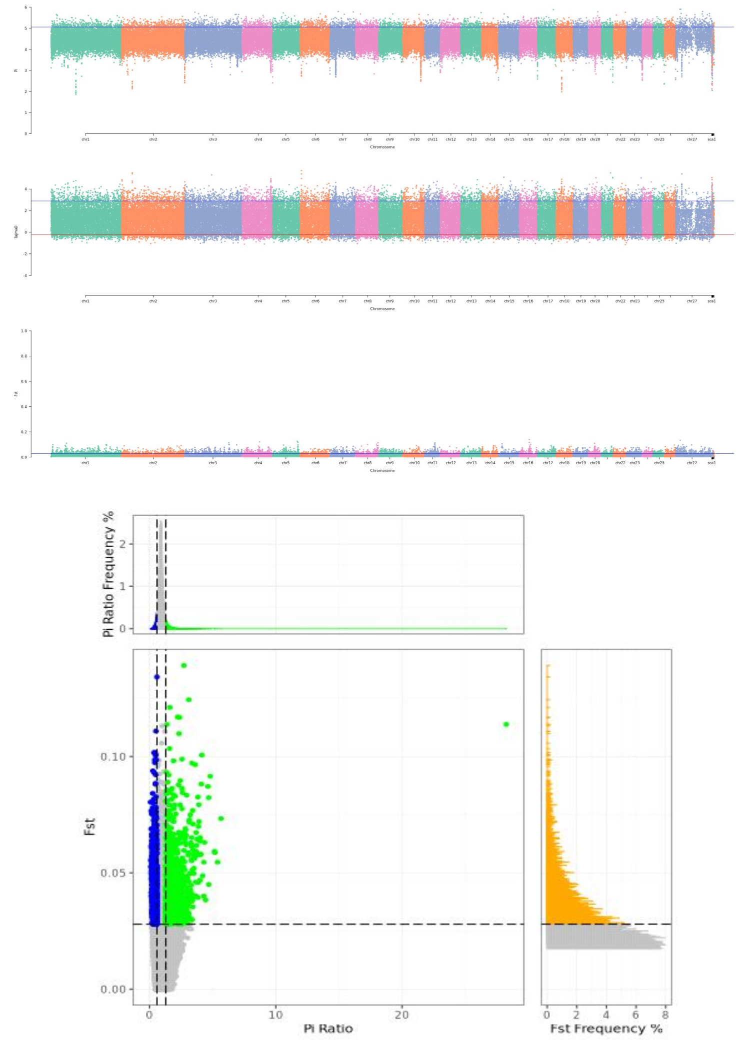
Fig 3: Population genetic structure based on whole genomes.

The results of the principal component analysis that was carried out on the genomic relatedness matrix between individuals are illustrated in Fig 2.

The population structure of the samples was analyzed. The number of clusters (K value) was assumed to be 1-20. The optimal number of clusters was determined to be 14 according to the K value corresponding to the lowest point of CV error (Cross validation error). It reflects that all our samples may come from 14 original ancestors. The population structure with K value of 14 is shown in Fig 3.

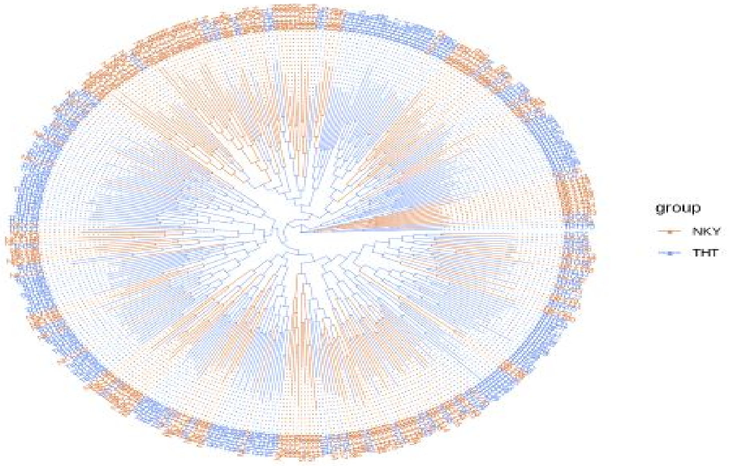
Through whole genome resequencing, it can be determined that there are regions with higher F_{ST} values and lower ð/è values on the genome. The q chromosome results are shown in Fig 4.

F_{ST} is an important index to measure the degree of genetic differentiation between populations. If an allele in a population experiences adaptive selection due to its high fitness to a specific environment, the increase of its frequency will increase the level of population differentiation, which is reflected in the higher F_{ST} value in F statistics. The



The abscissa represents the ratio of the two subpopulations $\theta\pi$ (θNKY/θTHT) and the ordinate represents F_{ST}. The top of the figure shows the ratio distribution of $\theta\pi$, the right side of the figure shows the F_{ST} distribution, the black curve shows the aggregation value of F_{ST} and $\theta\pi$ and the percentage of the number of SNPs less than or equal to the corresponding value in the total number of SNPs. The red region represents SNP loci with F_{ST} greater than or equal to 0.361, the blue region represents markers with F_{ST} greater than 0.361 and a ratio of $\theta\pi$ less than 0.57, these markers represent SNP loci of NKY population under positive selection, the green region represents markers with FST greater than 0.361 and a ratio of $\theta\pi$ greater than 1.10, these markers represent SNP loci of THT population under positive selection. The black and the dust area represent SNP loci that are not positively selected.

Fig 4: Selective sweep Analysis (Manhattan plot).



Each branch in the figure is a sample and the junction of two Adjacent branches becomes a node, indicating that the branch point of the existing population of the inferred ancestor at the base of the tree is called the root node and a single ancestor is defined as the evolutionary branch or single source population.

Fig 5: Phylogenetic tree.

π/θ represents the nucleotide polymorphism within the population, and the selective sweep will cause the selected genomic region to show that the polymorphism within the population is significantly lower than the average level of the genome, which corresponds to a lower π/θ value.

The evolutionary tree is used to represent the evolutionary relationship between species. According to the genetic relationship between various organisms, various organisms are placed on a branched tree-like figure to concisely represent the evolutionary history and genetic relationship of organisms. The phylogenetic tree is shown in Fig 5.

Before human-mediated specialization for wool and milk began some 4000-5000 years ago, sheep were initially raised for access to meat. One of the earliest morphological modifications that likely occurred alongside domestication and is now a trait shared by many modern breeds is the removal of horns, which has recently been demonstrated to be one of those regions of the genome that contain strong evidence for accelerated change in response to artificial selection(Mohamadipoor Saadatabadi *et al*. 2021). Additionally, sheep's other genomic areas that are being selected for comprise genes that regulate body size, reproduction and color. The separation of animals into breeds, followed by the discovery of better rams and their disproportionate genetic contribution through artificial insemination, has increased the rate of genetic gain for productive attributes during the previous few hundred years. Numerous genetic exchanges have taken place during the evolution of contemporary breeds, as evidenced by the high haplotype sharing and relatively short divergence times across breeds(Missohou *et al*. 2022). A population size of more than 300 has been maintained by around 75% of contemporary sheep breeds. There are between 850 and 1409 different breeds. But several sheep breeds, such as the Shetland, Soay, and Herdwick breeds, that were first chosen for their superior performance in a particular, sometimes remote, geographic location are now regarded as uncommon. It has become difficult to maintain genetic variety by preserving these historic breeds as generalist-type sheep have mostly taken over intensive sheep farming(Alberto *et al*. 2018).

CONCLUSION

Domestication is complex and entails multiple and parallel events and population admixtures. There are many breeds of sheep in China that have complex genetic exchanges. Native sheep breeds are abundant in China and some of them have unique characteristics including procreation, good lamb pelts, numerous horns, fat tails, and great stress tolerance. The focus of the research has been on all of these. However, there is still debate regarding the origins of sheep breeds and there have been few thorough investigations on the genetic diversity of local sheep in China. Their population dynamics and genetic variety were subsequently discovered, which was beneficial for our comprehension and preservation of these genetic resources. The current research, which for the first time examined the genome-wide diversity and population structure of sheep breeds with a Northern Chinese origin, advanced knowledge of the intricate genetic relationships and anticipated shared ancestry of the many sheep breeds in China.

ACKNOWLEDGEMENT

We are thankful to Bayannaoer Academy of Agriculture and Animal Husbandry for their assistance in the sampling.

Funding

This work was supported by Inner Mongolia Autonomous Region Science and Technology Plan Project(No. 2020GG0201), Natural Science Foundation of China (No. 31760660), Natural Science Foundation of Inner Mongolia Autonomous Region (No. 2019MS03092).

Competing interest

The authors declare that there are no competing interests associated with the manuscript.

Author contribution

BJ designed the study and analyzed the RNA-seq data and drafted the manuscript. ZH, YP and RL joined the research and improved the manuscript.

REFERENCES

Alberto, F.J., Boyer, F., Orozco-ter, W.P., Streeter, I., Servin, B., de Villemereuil, P., Benjelloun, B., Librado, P., Biscarini, L.F. *et al.* (2018). Convergent genomic signatures of domestication in sheep and goats. Nat Commun, 9: 813. doi: 10.1038/s41467-018-03206-y.

Alexander, D.H., Novembre, J. and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655-64.

Cingolani, E., Yee, K., Shehata, M., Chugh, S.S., Marban, E. and Cho, H.C. (2012). Biological pacemaker created by percutaneous gene delivery *via* venous catheters in a porcine model of complete heart block. Heart Rhythm. 9: 1310-8.

Kalds, P., Luo, Q., Sun, K., Zhou, S., Chen, Y. and Wang, X. (2021). Trends towards revealing the genetic architecture of sheep tail patterning: Promising genes and Investigatory Pathways. Anim Genet. 52: 799-812.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25: 1754-60.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K. A. *et al.* (2010). The genome analysis toolkit: A map reduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297-303.

Missohou, A., Kabore, B., Flori, L., Ayssiwede, S.B., Hornick, J.L., Raes, M. and Cabaraux, J.F. (2022). Analysis of the genetic diversity and population structure of four senegalese sheep breeds using medium-density single-nucleotide polymorphisms. Animals (Basel). 12(12), 1512; https://doi.org/10.3390/ani12121512.

Mohamadipoor, S.L., Mohammadabadi, M., Ghanatsaman, Z.A., Babenko, O., Stavetska, R., Kalashnik, O., Kucher, D. *et al.* (2021). Signature selection analysis reveals candidate genes associated with production traits in Iranian sheep breeds. BMC Vet Res. 17: 369.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22: 2688-90.

Taylor, W.T.T., Pruvost, M., Posth, C., Rendu, W., Krajcarz, M.T., Abdykanova, A., Brancaleoni, G., Spengler, R., Hermes, S.T. *et al.* (2021). Evidence for early dispersal of domestic sheep into Central Asia. Nat Hum Behav. 5: 1169-79.

Tsartsianidou, V., Kapsona, V.V., Sanchez-Molano, E., Basdagianni, Z., Carabano, M.J., Chatziplis, D., Arsenos, G., Triantafyllidis, A. and Banos, G. (2021). Understanding the seasonality of performance resilience to climate volatility in mediterranean dairy sheep. Sci Rep. 11: 1889. doi.org/10.1038/s41598-021-81461-8.

Tsartsianidou, V., Sanchez-Molano, E., Kapsona, V.V., Basdagianni, Z., Chatziplis, D., Arsenos, G., Triantafyllidis, A. and Banos, G. (2021). A comprehensive genome-wide scan detects genomic regions related to local adaptation and climate resilience in Mediterranean domestic sheep. Genet Sel Evol. 53: 90. doi.org/10.1186/s12711-021-00682-7.

Weihua, C., Guochao, N., Hao, L., Wang, J. (2022). Integrative transcriptomic and proteomic analysis of ovaries at different physiological periods in dolang sheep. Indian Journal of Animal Research. 56(12): 1492-1498.