Iranian Society of Animal Science

Journal of Livestock Science and Technologies

Shahid Bahonar University of Kerman

ISSN: 2322-3553 (Print) ISSN: 2322-374X (Online)

Paper type: Original Research

Identification of genes affecting litter size trait in Zandi sheep using random forest method

Roghaye Tanegonbadi[1], Hossein Moradi Shahrbabak[1], Mohammad Moradi Shahrbabak[1], Hossein Banisaadat[2]

[1]*Department of Animal Sciences, Faculty of Agriculture, College of Agriculture and Natural Resources, University of Tehran, Alborz Province, Karaj, Iran*
[2]*Department of Animal Science, Faculty of Agriculture, Tarbiat Modares University, Tehran, Iran*

[*]Corresponding author,
E-mail address:
hmoradis@ut.ac.ir

Received: 05 May 2026,
Received in revised form: 08 Jul 2026,
Accepted: 28 Jul 2026,
Published online: 30 Jul 2026,

ORCID
Roghaye Tanegonbadi
0009-0003-5502-1052
Hossein Moradi Shahrebabak
0000-0002-6680-7662
Mohammad Moradi Shahrebabak
0000-0002-5255-609X
Hossein Banisaadat
0000-0001-9034-0372

Abstract Knowledge of the association of single nucleotide polymorphisms with economically important traits is an important tool for breeding programs in the livestock industry. The aim of the present study was to identify important and influential markers of litter size trait in Zandi sheep using genomic data and random forest method. Significant markers including s59645, OAR14_49049707, s51972, OAR12_19965932, OAR14_40046813, s61144 and OAR2_53988637 were identified by the machine learning method (random forest) for the litter size. The identified markers were associated with new genes including *STEAP2*, *CCNP*, *RAB11FIP2, GPATCH2*, *URI1*, *ROPN1L* and *SKAP1*. Gene ontology study showed that genes discovered by the random forest play roles in key biological functions such as cellular metabolic processes, DNA-templated transcription, biosynthetic process, embryonic skeletal system morphogenesis, embryonic skeletal system development, regulation of gene expression, organ morphogenesis, skeletal system development, and chordate embryonic development. The identified candidate genes provided additional insights into the genetic mechanisms underlying the litter size in Zandi sheep. Furthermore, the important SNP markers identified by the random forest analysis could serve as potential candidates for future functional validation and marker-assisted or genomic selection programs.

Keywords: litter size machine learning, random forest method, single nucleotide polymorphisms, Zandi sheep

Introduction

Reproductive traits are the most important factors affecting the economic profitability of sheep production systems and are significantly influenced by the ewe ability to produce more lambs annually (Abdoli et al., 2019). Litter size is an economically important reproductive trait, and increasing the number of lambs born per ewe can improve production efficiency and contribute to the economic sustainability of sheep farming (Zhang et al., 2022). Zandi sheep is a medium-weight breed native to Iran with several desirable characteristics, including relatively low subcutaneous fat deposition and tolerance to harsh environmental conditions, such as temperature fluctuations, poor nutritional conditions, and low-quality pastures (Bohlouli et al., 2013). In recent years, molecular biology techniques have been developed at a rapid pace, leading to the completion of various whole-genome sequencing projects, including sheep, and subsequently, various panels of nano arrays with different densities have been designed, enabling whole-genome scanning studies. Genome scanning studies can help to accurately identify the genes associated with economic traits, and

their findings can be useful for marker-assisted selection (Johnston et al., 2011).

In genome scanning studies, the existence of a large number of single nucleotide polymorphisms (SNPs) at the genome level, along with phenotype are used to analyze and identify the relation between genotype and phenotype, genes and regulatory elements associated with important economic traits (Hayes, 2013). In these studies, the most commonly used method for examining the effect of markers is the linear regression method (Linear Model) and P-value. In various studies, the problems of using this method have been pointed out as increasing the type I error rate and overestimating the effect of markers, not considering linkage disequilibrium between markers, incorrectly assuming that markers are independent of each other, assuming all genomic variables are from a normal distribution, and not considering the interaction between markers. (Sun et al., 2021; Enoma et al., 2022). Consequently, a suitable alternative to solving these problems is to conduct genome scanning studies based on machine learning methods.

Machine learning methods have been widely used in genomics studies as a tool to optimize predictive power in large datasets for animal science studies (Li et al., 2018; Hu et al., 2020). These methods are data-driven learning algorithms that can be trained using existing data and use the discovered patterns to predict future scenarios. Machine learning uses a training set to obtain the desired output, depending on the type of training model. This training set includes correct inputs and outputs that allow the model to learn over time (Baştanlar and Özuysal, 2014).

Studies have shown that machine learning methods, such as the random forest in genomic scanning studies is an appropriate method for selecting important markers and identifying genes affecting economic traits (Enoma et al., 2022). Unlike other statistical models, these methods do not require determining the distribution of the studied population, evaluating estimators (p-values) or testing null hypotheses, and are a free statistical method. In addition, these methods are able to identify markers with no effect or a negative effect on the trait and consider nonlinear relationships between markers and phenotype, and can be easily used in large-scale data (Li et al., 2018).

Factors such as linkage disequilibrium, non-additive genetic effects, allele frequency distribution, complex nonlinear relationships, and interactions among genetic markers can influence the identification of loci associated with complex traits (Li et al., 2018; Enoma et al., 2022). Conventional regression-based GWAS methods are primarily designed to detect additive linear effects and may have limited ability to capture complex genetic interactions. Random Forest (RF) is a non-parametric ensemble learning algorithm capable of modeling nonlinear relationships and interactions among genetic markers without requiring assumptions about the underlying data distribution (Breiman, 2001). In sheep, RF has been successfully applied as a complementary approach to GWAS for validating associated loci and reducing false-positive associations (Kijas et al., 2013). More recently, machine learning-enhanced GWAS frameworks incorporating RF have demonstrated their usefulness for identifying biologically relevant candidate genes associated with economically important traits in sheep (Arzık et al., 2025). In addition, RF has shown excellent performance in genomic prediction studies, achieving high prediction accuracy for breeding values in sheep (Hamadani et al., 2022). However, the application of RF for identifying genomic regions associated with litter size in sheep remains limited. Therefore, the present study aimed to identify important SNP markers and candidate genes associated with litter size in Zandi sheep using the Random Forest method.

Materials and methods

Phenotypic and genotypic data

The data for this research were obtained from the National Genomic Project of Zandi sheep, supervised by the Animal Breeding and Improvement Center of Iran and the Agricultural Jihad Organization of Tehran Province, and implemented by the knowledge-based company Saina Gostar Alborz. Samples were collected from the Zandi Sheep Breeding Station in Varamin, Tehran Province. The station maintains a closed nucleus flock under a pedigree recording and genetic improvement program. The breeding season extends from 15 September to 15 November, during which mating is carried out according to the station's breeding management program. The criterion for selecting ewes was based on twin or multiple births. Prolificacy was defined as the average number of lambs born per lambing, calculated as the total number of lambs born per ewe divided by the ewe's total number of lambings. Ewes with an average litter size greater than 1.3 lambs per lambing were classified as prolific. A total of 99 blood samples were collected from the station. Genomic DNA was extracted from blood using the optimized salt method (Helms, 1990). After DNA extraction at the Animal Science Laboratory of the University of Tehran and quality assessment of the samples, they were transferred to the Genomics Laboratory of the University of Kentucky, USA, for subsequent genotyping. The samples were genotyped using the Illumina OvineSNP50 BeadChip (Kijas et al.,2009).

Quality control

To perform quality control, samples with a call rate below 99% were first excluded. In addition, SNPs with a missing genotype rate greater than 1% were removed. In the dataset used in this study, the genotyping call quality was generally high, and most samples had a call rate close to 100%. Therefore, applying the 99% threshold resulted in the removal of only two animals with no significant impact on the sample size.

Subsequently, the SNP markers with a call rate below 95% and a minor allele frequency (MAF) below 0.01 were excluded, resulting in the removal of 1,386 and 2,095 SNPs based on call rate and MAF, respectively. In addition, 120 SNPs that deviated from Hardy-Weinberg equilibrium (P-value < 10^-6) or had unknown genomic locations (Moradi et al., 2012) were excluded from the final analysis. PLINK v1.9 (Chang et al., 2015) was used to perform all quality control procedures. Finally, after quality control and the removal of sex chromosomes to avoid potential biological and technical biases, 48,812 SNPs across the 26 autosomal chromosomes and 96 sheep were retained for the final analysis.

Random forest in genomic screening

The random forest (RF) method is a hybrid algorithm of several decision trees and create a set of decision trees in a forest that can make better predictions than an individual tree: each tree is grown from a bootstrap sample (sampling with replacement) of the original training data. The training data set includes genotypic information and phenotypic values (Breiman, 2001). To form each tree based on genotypic and phenotypic information, a sample is taken with replacement from the training data and the tree is grown on the sample taken and at each node from several SNPs, which are randomly selected: an SNP is selected according to the best resolution and the node is branched based on the genotype of that SNP and the phenotypes are separated based on that SNP to the next leaf node. In the next steps and at other nodes from other SNP samples, an SNP is selected and the branching of the nodes continues until the tree is complete and all phenotypes are in the leaves of the tree. In the same way, subsequent trees are formed. The random forest algorithm predicts the result by averaging the outputs from all trees in the forest (Breiman, 2001). In each bootstrap of the data, some data are never sampled, and some are sampled several times. In other words, any input data for some trees will be out-of-bag data that does not participate in the creation of some trees. In each bootstrap sample, some observations are not selected and therefore constitute the out-of-bag (OOB) samples. These OOB samples serve as an internal validation set, allowing the prediction error (OOB error) to be estimated for each tree. To evaluate the importance of each SNP, a permutation-based approach was applied. First, the OOB prediction error was calculated using the original genotypic values. Next, the genotypic values of the target SNP were randomly permuted among the OOB samples while all other SNPs remained unchanged, and the OOB prediction error was recalculated. The increase in OOB prediction error after permutation, relative to the original OOB prediction error, was averaged across all trees in the forest and used as the mean decrease accuracy (MDA) value for that SNP (Li et al., 2018). Higher MDA values indicate that permuting an SNP has a greater negative effect on prediction accuracy, suggesting that the SNP makes a more important contribution to the predictive performance of the model.

Prediction accuracy was used as the main criterion for model evaluation. To optimize the model hyperparameters, a grid search strategy combined with 3-fold cross-validation was applied. The dataset was randomly divided into three approximately equal subsets. In each iteration, two subsets (approximately 66% of the data) were used for model training and the remaining subset (approximately 33%) was used for validation. This process was repeated three times so that each subset served once as the validation set. The average accuracy across the three validation folds was used to evaluate each hyperparameter combination. The combination of ntree = 3000, mtry = p/3, and nodesize = 5, which achieved the highest mean accuracy, was selected for the final random forest model. The RF analysis was performed in R using the caret package with the randomForest engine (Kuhn, 2008).

Gene ontology (GO) enrichment analysis

All markers were ranked based on their importance after running the random forest method based on the best combination of parameters. After ranking, the top eleven markers for the multiple fecundity trait were identified. In addition, genes located within 500 Kb of the top 10 markers, which were detected in the genomic region by each method, were identified using the NCBI and Ensemble databases of the sheep reference genome (ovis-aries) (Yates et al., 2016). To explore the biological and functional processes of the genes, gene ontology was performed using the online software DAVID (http://david.abcc.ncifcrf.gov/). Then, we drew the gene network using Cytoscape 3.10.3 software (Shannon et al., 2003).

Gene networks and hub genes

A gene interaction network was constructed for the 29 functional and positional candidate genes associated with litter size in Zandi sheep. Protein–protein interaction information was obtained from the STRING database, and the resulting network was visualized using Cytoscape. In the network, genes were represented as nodes and interactions between genes as edges. Hub genes were subsequently identified based on their network connectivity.

Results

Genome-wide identification of important SNPs

The rank of SNPs from the most to the least important for litter size is shown in Figure 1.

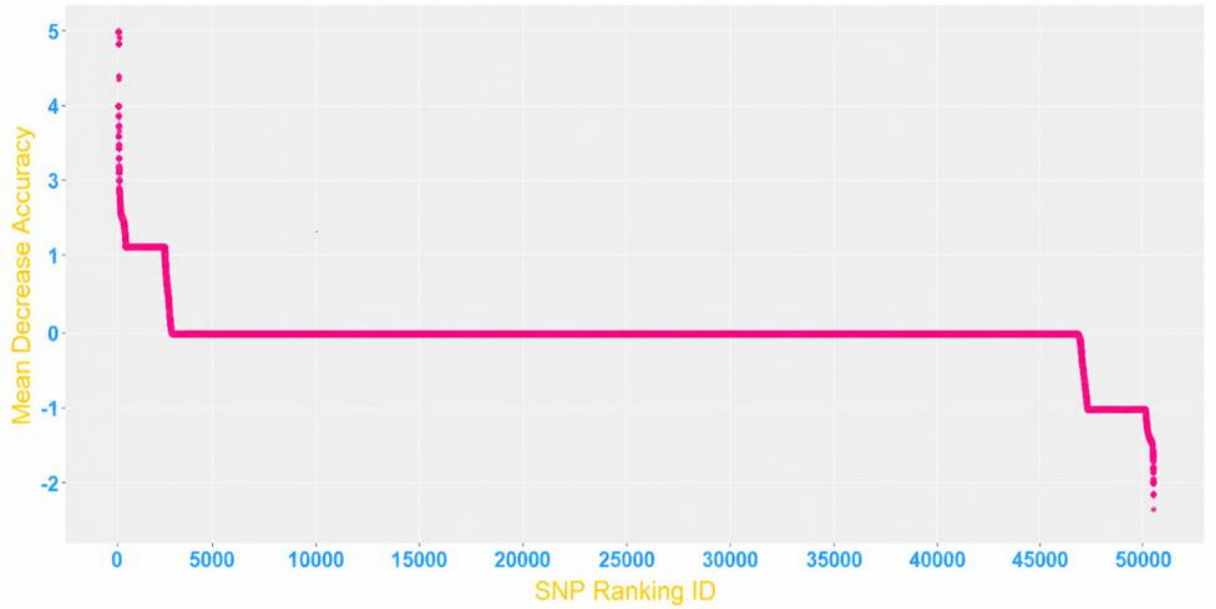

Figure 1. The distribution of ranked SNPs for the litter size trait RF method

In random forest analysis, the importance of SNPs is determined as negative, zero, and positive by the mean decrease accuracy (MDA) index. In Figure 1, the ranking of SNPs based on this index is shown from the most important to the least important, where the higher the index value indicates greater the importance of the SNP. In this analysis, 2581 (5.28%) SNPs had a positive effect, 42590 (87.25%) SNPs had no effect, and 3641 (7.45%) SNPs had a negative effect. The MDA rank range was from 3.73 to -2.37. A negative MDA value indicates that its presence in the analysis increases the error, a positive MDA value reduces the error and a zero value has a neutral effect on the analysis.

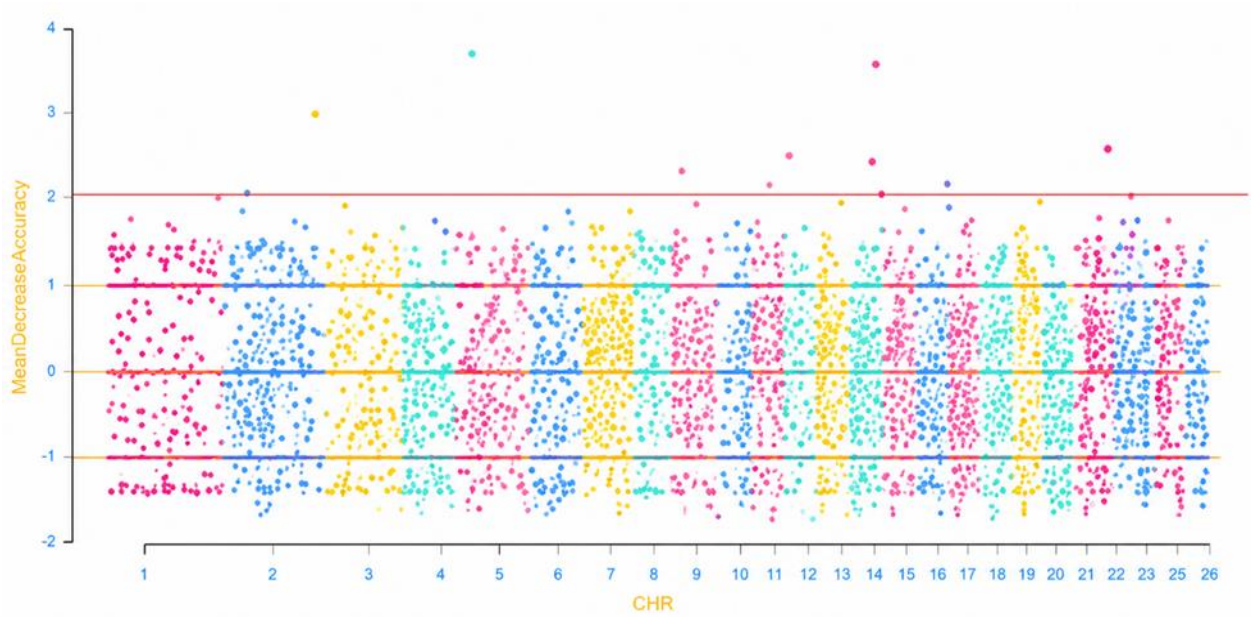

Figure 2. Manhattan plot for litter size trait of Zandi sheep using the random forest method

The Manhattan plot with random forest indicates how markers with significance and influence on a trait are distributed across chromosomes (Figure 2). In this study, the MDA threshold (≥2) was determined empirically according to the distribution of variable importance values. This cutoff retained 11 highly informative SNPs for subsequent candidate gene identification. The significant markers were located on chromosomes 4, 14, 2, 22, 12, 9, 16, and 11, respectively, with three markers on chromosome 14, two markers on chromosome 2, and one marker on the remaining chromosomes.

Table 1. The list of the top 11 ranking SNPs from random forests (RF)

Rank	Chr[1]	Marker name	Position(bp)	MDA[2]	Downstream	Gene	Upstream
1	4	s59645	75773352	3.735497	*STEAP1*	*STEAP2*	*CFAP69*
2	14	OAR14_49049707	49049707	3.604072	*PSMC4*	*CCNP*	*TTC9B*
3	2	OAR2_197496174	197496	2.90064			
4	22	s51972	37809693	2.489533	*EMX2*	*RAB11FIP2*	*FAM204A*
5	12	OAR12_19965932	19965932	2.41585		*GPATCH2*	*SPATA17*
6	14	OAR14_40046813	40046813	2.351949	*CCNE1*	*URI1*	*ZNF536*
7	9	OAR9_18170750	18170750	2.252417			
8	16	s61144	63383307	2.115934	*CCT5*	*ROPN1L*	*SEMA5A*
9	11	OAR11_38029087	38029087	2.097924			
10	2	OAR2_53988637	53988637	2.014958	*HOXB1*	*SKAP1*	*SNX11*
11	14	OAR14_61077517	61077517	2.000662			

[1] Chromosome
[2]Mean decrease accuracy

The top 11 SNPs are listed in Table 1 in order of importance. In this table, the SNPs OAR2_197496174.1, OAR9_18170750.1, OAR11_38029087.1 and OAR14_61077517.1, with ranks 3, 7, 9 and 11 respectively, were not located on any gene and were not near any gene.

Gene ontology (GO) enrichment analysis

Table 2 presents the results of gene ontology (GO) enrichment analysis of the candidate genes identified within ±500 kb upstream and downstream of the top 11 SNPs using the Ensembl database. Functional annotation showed that these candidate genes are mainly involved in the regulation of cellular metabolic processes, regulation of DNA-templated transcription, regulation of biosynthetic processes, embryonic skeletal system morphogenesis, embryonic skeletal system development, regulation of gene expression, animal organ morphogenesis, skeletal system development, chordate embryonic development, and embryonic development ending in birth.

Table 2. Gene enrichment analysis for the top 62 SNPs with positive variable importance values from RF

Term	Count	Raw P-value	Genes
GO:0031323~regulation of cellular metabolic process, regulation of DNA-templated transcription, regulation of biosynthetic process	31	2.77E-10	*URI1, PLEKHF1, HOXB13, SERTAD3, SERTAD1, TBX21, PRX, ROPN1L, CCT5, CDK5RAP3, EMX2, CBX1, DYRK1B, ATPSCKMT, HOXB9, SP2, CCNE1, SP6, HOXB4, MAP3K10, HOXB3, ZNF536, HOXB2, HOXB1, HOXB8, CCNP, HOXB7, SKAP1, HOXB6, NFE2L1, HOXB5*
GO:0010468~regulation of gene expression	24	1.26E-06	*EMX2, URI1, CBX1, DYRK1B, HOXB13, SERTAD3, HOXB9, SP2, SERTAD1, TBX21, PRX, SP6, HOXB4, MAP3K10, HOXB3, ZNF536, HOXB2, HOXB1, HOXB8, HOXB7, SKAP1, HOXB6, NFE2L1, HOXB5*
GO:0048568~embryonic organ development, skeletal system development, chordate embryonic development, embryonic morphogenesis, embryo development, embryonic organ morphogenesis, embryonic skeletal system morphogenesis, embryonic skeletal system development	6	5.34E-05	*HOXB9, HOXB4, HOXB3, HOXB1, HOXB7, HOXB5*
GO:0048731~system development, animal organ morphogenesis	11	0.004006	*SEMA5A, SPTBN4, EMX2, HOXB9, PRX, HOXB4, HOXB3, HOXB1, HOXB13, HOXB7, HOXB5*
GO:0032502~developmental process	15	0.007999	*SEMA5A, SPTBN4, EMX2, HOXB13, HOXB9, TBX21, PRX, SP6, HOXB4, HOXB3, ROPN1L, HOXB1, HOXB7, HOXB6, HOXB5*

Gene networks and hub genes

Gene network showing the connections among 29 interconnected functional and positional candidate genes associated with litter size in Zandi sheep is presented in Figure 3. In the network, genes are represented as nodes, whereas edges represent gene–gene interactions. Hub genes were identified based on their high degree of connectivity within the network, which include *PNPO, NFE2L1, CCNE1, PSMC4, HOXB6, HOXB9, HOXB8, HOXB5, HOXB4, HOXB7, HOXB2, HOXB7, HOXB13, HOXB1, SNX11, COPZ2.* The genes *HOXB4, MAP3K10, HOXB3, HOXB2, HOXB1, HOXB8, HOXB7,* and *SKAP1* are involved in the regulation of DNA transcription. *HOXB9, HOXB4, HOXB3, HOXB1, HOXB7* and *HOXB5* are involved in the development of the skeletal system.

Discussion

The most significant SNP identified in this study, s59645.1, was located on chromosome 4 and associated with the *STEAP2* gene. *STEAP2* plays a role in fat metabolism and has been reported to be

significantly upregulated in myogenic precursors differentiating into skeletal muscle fibers, suggesting a relationship between myogenesis, adipogenesis, and intramuscular fat deposition (Qiu et al., 2018).

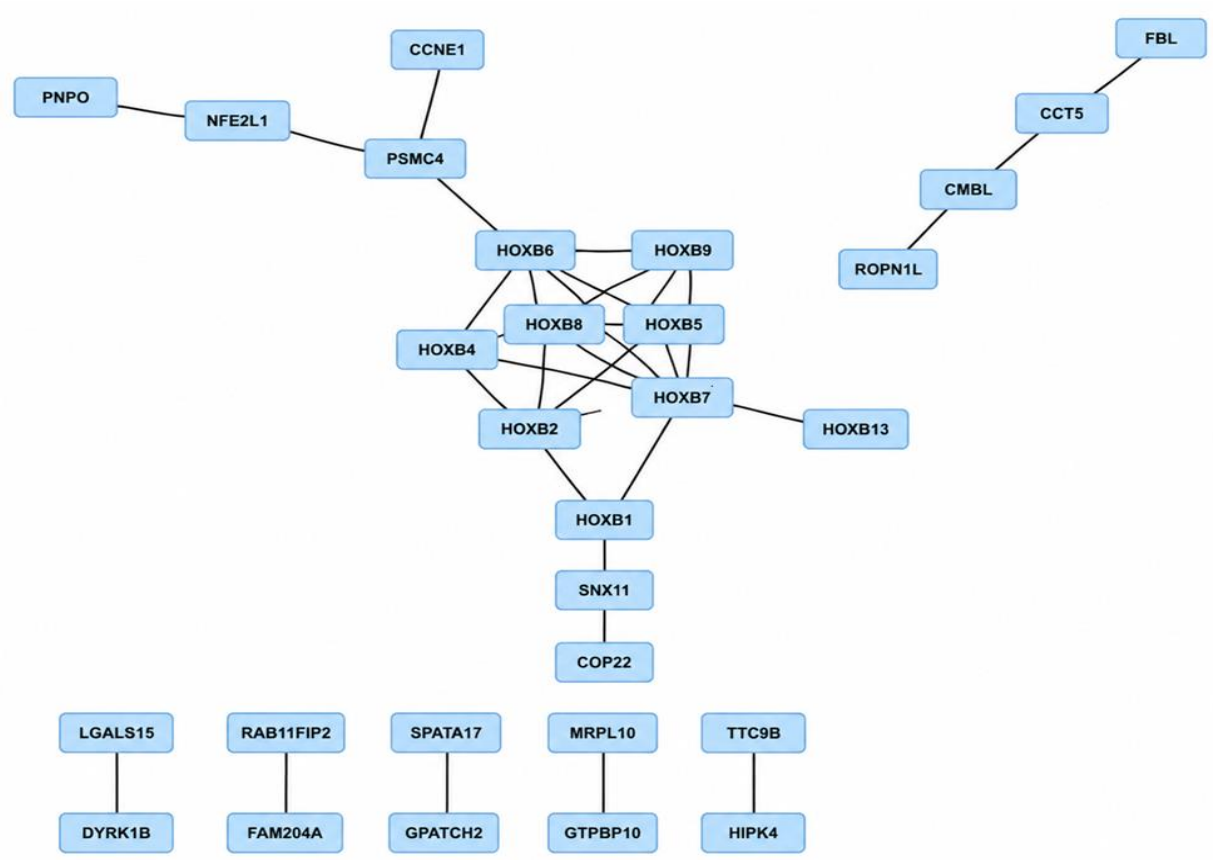

Figure 3. Gene network displaying the connections between 29 genes in Zandi sheep

Several genes identified in the significant genomic regions are associated with reproductive performance. The *SKAP1* gene functions as an immune cell adaptor involved in T-cell immune responses (Smith et al., 2016), which are critical for ovarian function (Jiao et al., 2021).

The *URI1* gene encodes a protein involved in regulating cell growth, survival, and reproduction. Although *URI1* is not directly involved in reproductive cell development, its roles in transcriptional regulation and embryogenesis suggest indirect effects on reproductive processes (Kirchner et al., 2008).

Mammals have nearly 40 members of the *HOX* gene family, which are divided into four clusters: *HOXA, HOXB, HOXC, and HOXD*. The formation of these genes is mainly due to the duplication of ancestral clusters and subsequent gene loss or duplication (Krumlauf, 2018; Duboule, 2007).

The *HOXB13* gene is significantly associated with growth traits in sheep. The HOX gene encodes transcriptional regulatory proteins that control bilateral axial patterns in all animals (Mallo et al., 2010; Garcia-Fernàndez et al., 2005). Hox genes contribute to anterior-posterior (A-P) patterning during embryonic development in vertebrates and play an important role in the axial growth pattern of organisms (Godsave et al.1994). *HOXB13* is the most distal of the 5 genes in the *HOXB* cluster, which is related to tail formation (Gofflot et al., 1997). There is a significant association between The HOXB13 gene has been associated with pelvic width in Luxi sheep (Zhu et al., 2024). Pelvic width is an important morphological trait that reflects body conformation and may be related to reproductive performance in sheep. A wider pelvic structure may provide a larger pelvic cavity, potentially facilitating parturition and contributing to reproductive performance. During development, Hoxb-13 is expressed in the tail and posterior regions of the spinal cord, gastrointestinal tract, and genitourinary tract (Economides et al., 2003). *SKAP1*, an immune cell adaptor, is involved in T-cell immune responses (Smith et al., 2016), which are critical for ovarian function (Jiao et al., 2021). The *NFE2L1* gene in sheep is involved in meat production (Trukhachev et al., 2016).

The *PSMC4* gene plays a crucial role in early embryogenesis, specifically in blastocyst development, and is part of the 26S proteasome, a complex involved

in protein degradation. Studies in mice have shown that *PSMC4* is essential for early embryogenesis, and its deficiency leads to developmental defects before implantation. Furthermore, *PSMC4*'s involvement extends to various biological processes, including DNA replication, chromatin assembly, and mitotic prophase (Zhu et al., 2025).

The *SP6* gene plays a crucial role in controlling cytotrophoblast cell fate and trophoblast stem cell maintenance during human placenta development (Xiao et al., 2002).

The *EMX2* gene plays a critical role in both the development of the reproductive tract and in regulating endometrial physiology during reproduction. It acts as a transcription factor, influencing cell proliferation and differentiation within the uterus, particularly in the endometrium, the lining of the uterus (Hugh Xiaolan., 2005).

Conclusions

In the present study, a genomic analysis of the litter size trait in Zandi sheep was performed using the random forest method to identify significant markers using a 50k cutoff of single nucleotide polymorphisms. Significant markers including s59645, OAR14_49049707, s51972, OAR12_19965932, OAR14_40046813, s61144, and OAR2_53988637 were identified by the machine learning method (random forest) for the litter size trait in Zandi sheep. The markers identified in the machine learning methods were associated with new genes including *STEAP2*, *CCNP*, *RAB11FIP2*, *GPATCH2*, *URI1*, *ROPN1L*, and *SKAP1*. Gene ontology analysis indicated that these candidate genes are involved in several biological processes potentially related to reproductive performance and litter size. These findings provide additional insights into the genetic architecture of litter size in Zandi sheep and demonstrate the usefulness of the Random Forest algorithm as a feature selection approach for identifying candidate genomic regions associated with complex traits.

Conflict of interest

The authors confirm that there are no conflicts of interest with the conduct or publication of this research.

Acknowledgements

We would like to express our gratitude to the Saina Gostar Alborz Knowledge-Based Company as the implementer of the National Genomics of Zandi Sheep Project and for financing the project, and for the great efforts they have made in sampling, DNA extraction, and sending samples abroad to determine the genotype of the livestock, as well as the National Animal Breeding Center, which supported the project.

References

Abdoli, R., Mirhoseini, S.Z., Ghavi Hossein-Zadeh, N., Zamani, P., Moradi, M.H., Ferdosi, M.H., Gondro, C., 2019. Genome-wide association study of first lambing age and lambing interval in sheep. *Small Ruminant Research* 178, 43-45.

Baştanlar, Y., Özuysal, M., 2014. Introduction to machine learning. *Methods in Molecular Biology* 1107, 105-128.

Bohlouli, M., Mohammadi, H., Alijani, S., 2013. Genetic evaluation and genetic trend of growth traits of Zandi sheep in semi-arid Iran using random regression models. *Small Ruminant Research* 114, 195-201.

Breiman, L., 2001. Random forests. *Machine Learning* 45, 5-32.

Duboule, D., 2007. The rise and fall of Hox gene clusters. *Development* 134, 2549-2560.

Economides, K.D., Zeltser, L., Capecchi, M.R., 2003. HOXB13 mutations cause overgrowth of caudal spinal cord and tail vertebrae*. Developmental Biology* 256, 317-330.

Emrani, H., Vaez Torshizi, R., Masoudi, A.A., Ehsani, A., 2017. Identification of new loci for body weight traits in F2 chicken population using genome-wide association study. *Livestock Science* 206, 125–131.

Enoma, D.O., Bishung, J., Abiodun, T., Ogunlana, O., Osamor, V.C., 2022. Machine learning approaches to genome-wide association studies. *Journal of King Saud University - Science* 34, 101847.

Garcia-Fernández, J., 2005. Hox, ParaHox, ProtoHox: Facts and guesses. *Heredity* 94, 145-152.

Godsave, S., Dekker, E.J., Holling, T., Pannese, M., Boncinelli, E., Durston, A., 1994. Expression patterns of Hoxb genes in the Xenopus embryo suggest roles in anteroposterior specification of the hindbrain and in dorsoventral patterning of the mesoderm. *Developmental Biology* 166, 465-476.

Gofflot, F., Hall, M., Morriss-Kay, G.M., 1997. Genetic patterning of the developing mouse tail at the time of posterior neuropore closure. *Developmental Dynamics* 210, 431-445.

Hamadani, A., Ganai, N.A., Mudasir, S., Shanaz, S., Alam, S. and Hussain, I., 2022. Comparison of artificial intelligence algorithms and their ranking for the prediction of genetic merit in sheep. *Scientific Reports* 12, 18726.

Hayes, B., 2013. Overview of statistical methods for genome-wide association studies (GWAS). *Methods in Molecular Biology* 1019, 149-169.

Helms, C. (1990). Salting out procedure for human DNA extraction. Available at: http://humgen.wustl.edu/hdk_lab_manual/dna/dna2.html. Accessed 20 April 2010.

Hu, T., Darabos, C., Urbanowicz, R., 2020. Machine learning in genome-wide association studies. *Frontiers in Genetics* 11, 593958.

Jiao, X., Zhang, X., Li, N., Zhang, D., Zhao, S., Dang, Y., Zanvit, P., Jin, W., Chen, Z.-J., Chen, W., Qin, Y., 2021. Treg deficiency-mediated TH1 response causes human premature ovarian insufficiency through apoptosis and steroidogenesis dysfunction of granulosa cells. *Clinical and Translational Medicine* 11, 448.

Johnston, S.E., McEwan, J.C., Pickering, N.K., Kijas, J.W., Beraldi, D., Pilkington, J.G., Pemberton, J.M., Slate, J., 2011. Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Molecular Ecology* 20, 2555-2566.

Kijas, J.W., Serrano, M., McCulloch, R., Li, Y., Salces Ortiz, J., Calvo, J.H., Pérez-Guzmán, M.D. and the International Sheep Genomics Consortium., 2013. Genomewide association for a dominant pigmentation gene in sheep. *Journal of Animal Breeding and Genetics* 130, 250-255.

Kijas, J.W., Townley, D., Dalrymple, B.P., Heaton, M.P., Maddox, J.F., McGrath, A., Wilson, P., Ingersoll, R.G., McCulloch, R., McWilliam, S., Tang, D., McEwan, J., Cockett, N., Oddy, V.H., Nicholas, F.W., Raadsma, H., 2009. A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. *PLoS ONE* 4, e4668.

Kirchner, J., Vissi, E., Gross, S., Szoor, B., Rudenko, A., Alphey, L., White-Cooper, H., 2008. Drosophila Uri, a PP1α binding protein, is essential for viability, maintenance of DNA integrity and normal transcriptional activity. *BMC Molecular Biology* 9:36.

Krumlauf, R., 2018. Hox genes, clusters and collinearity. I*nternational Journal of Developmental Biology* 62, 659-663.

Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28(5), 1-26.

Li, B., Zhang, N., Wang, Y.G., George, A.W., Reverter, A., Li, Y., 2018. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Frontiers in Genetics* 9, 237.

Mallo, M., Wellik, D.M., Deschamps, J., 2010. Hox genes and regional patterning of the vertebrate body plan. *Developmental Biology* 344, 7-15.

Moradi, M.H., Nejati-Javaremi, A., Moradi-Shahrbabak, M., Dodds, K.G. and McEwan, J.C, 2012. Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *BMC Genetics* 13, 10

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81, 559-575.

Qiu, K., Zhang, X., Wang, L., Jiao, N., Xu, D., Yin, J., 2018. Protein expression landscape defines the differentiation potential specificity of adipogenic and myogenic precursors in the skeletal muscle. *Journal of Proteome Research* 17, 3853-3865.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research* 13, 2498-2504.

Smith, X., Taylor, A., Rudd, C.E., 2016. T-cell immune adaptor SKAP1 regulates the induction of collagen-induced arthritis in mice. *Immunology Letters* 176, 122-127.

Sun, S., Dong, B., Zou, Q., 2021. Revisiting genome-wide association studies from statistical modelling to machine learning. *Briefings in Bioinformatics* 22, 263.

Taylor, H.S., Fei, X., 2005. Emx2 Regulates Mammalian Reproduction by Altering Endometrial Cell Proliferation. *Molecular Endocrinology* 19, 2839-2846.

Trukhachev, V., Skripkin, V., Kvochko, A., Kulichenko, A., Kovalev, D., Selionova, M., Aybazov, M., Yatsyk, O., Krivoruchko, A., 2016. Sequencing of the NFE2L1 gene in sheep and evaluation influence of gene polymorphisms on meat production*. Journal of Animal & Plant Sciences* 26, 1262-1267.

Xiao, J., Xu, M., Li, J., Chang Chan, H., Lin, M., Zhu, H., Zhang, W., Zhou, Z., Zhao, B., Sha, J., 2002. NYD-SP6, a novel gene potentially involved in regulating testicular development/spermatogenesis. *Biochemical and Biophysical Research Communications* 291, 101-110.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C.G., 2016. Ensembl 2016. *Nucleic Acids Research*- 44 (Database issue).

Zhang, Z., Sui, Z., Zhang, J., Li, Q., Zhang, Y., Wang, C., Li, X., Xing, F., 2022. Identification of signatures of selection for litter size and pubertal initiation in two sheep populations. *Animals* 12, 2520.

Zhu, L., Li, Y., Xu, Y., Lei, J., 2025. The role of PSMC4 in non-small cell lung cancer: implications for prognosis, diagnosis, and immune microenvironment modulation. *Frontiers in Oncology* 15, 1503466.

Zhu, Q., Liu, P., Zhang, M., Kang, Y., Lv, L., Xu, H., Zhang, Q., Li, R., Pan, C., Lan, X., 2024. The detection of a functional 168 bp deletion of the HOXB13 gene determining short tail and its association with senior growth traits in sheep breeds worldwide. *Animals* 14, 1617.