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

Genome-wide identification of candidate regions associated with growth traits in Lori-Bakhtiari sheep using the Random Forest algorithm

Maryam Ghasemi[1], Mohammad Moradi Shahrebabak[1], Hossein Moradi Shahrebabak[1], Pouya Zamani[2], Mohammad Hossein Moradi[1], Hossein Banisaadat[3]*

[1]*Department of Animal Science, University College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran*
[2]*Department of Animal Science, Faculty of Agriculture, Bu-Ali Sina University, Hamedan, Iran*
[3]*Department of Animal Science, Faculty of Agriculture, Tarbiat Modares University, Iran*

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

Received: 11 May 2026,
Received in revised form: 20 Jun 2026,
Accepted: 16 Jul 2026,
Published online: 17 Jul 2026,

ORCID
Maryam Ghasemi
0009-0007-7413-0034
Mohammad Moradi Shahrebabak
0000-0002-5255-609X
Hossein Moradi Shahrebabak
0000-0002-6680-7662
Pouya Zamani
0000-0002-1051-2259
Mohammad Hossein Moradi
0000-0001-5877-0866
Hossein Banisaadat
0000-0001-9034-0372

Abstract Nowadays, with the development of arrays containing single-nucleotide polymorphism (SNPs) information for most domestic animal species, genome-wide association studies (GWAS) are some important methods for identifying candidate genomic regions associated with production traits. Therefore, the use of modern approaches, including machine learning methods, especially nonlinear algorithms such as Random Forest (RF), are essential, because these methods are able to reveal complex patterns in genomic data that are often hidden from the view of traditional statistical models. Therefore, this study aimed to identify genes and SNPs affecting body weight traits in Lori-Bakhtiari sheep using RF algorithms. A total of 132 samples of Lori-Bakhtiari sheep were genotyped using Illumina Ovine SNP 50 Bead Chip, containing 51,135 SNP markers. Plink (v 9.1) was used for quality control, where a total of 44,796 SNPs and 122 sheep were retained for downstream analyses. The RF algorithm was used to find significant SNPs associated with weaning weight and 180 days of age' weight, using the RF package in R (v 4.4.0). The top 50-ranked SNPs were detected based on the RF method. The results indicated 59 candidate genes for weaning weight and 58 candidate genes for weight at 180 days of age. Among the identified candidate genes for weaning weight, the genes including *ABL1*, *INPPL1,* and *EXOC7*, and also for the genes revealed for 180 days of age, *ROS1*, *HCK,* and *ZFAND4* genes have been identified to be hub genes associated with body weight traits in the current study. Network analysis revealed functional connections among these genes, reinforcing their candidacy for body weight regulation. Notably, the most genes and networks identified in the current study are novel and have not been previously identified by GWAS, underscoring the RF algorithm's effectiveness in discovering novel genetic markers. In conclusion, these findings offer new perspectives on the genetic basis of body weight in sheep and propose targets for enhancing genetic selection in meat-producing breeds.

Keywords: body weight, single nucleotide polymorphism (SNP), candidate gene, Lori-Bakhtiari sheep, genome-wide association study (GWAS), radom forest (RF)

Introduction

Over the past two decades in genomic technologies and the development of ovine SNP chips have provided opportunities to identify genomic regions associated with economic traits and to enable genomic selection in sheep (McRae et al., 2014). Consequently, the use of trait-associated genetic markers through marker-assisted selection is a strategic tool to boost farm animal

profitability. Although the inheritance of quantitative traits primarily follows a polygenic model with additive gene action, exceptions exist where specific genes of major effect play a critical role in determining the trait's heritable variation (Almasi et al., 2020). Genetic markers additionally facilitate the selection of traits that are otherwise nearly impossible to assess. In many practical farming scenarios, where the direct assessment of feed intake is not viable, selection can instead target economic profitability due to its high genetic and phenotypic associations with measures of feed efficiency (Zamani, 2017). Incorporating feed efficiency into breeding goals can lead to increased genetic gain and greater profitability of farm animals. As a result, improving profitability through either marker-assisted selection or indirect selection on production traits (e.g., body weight in meat animals) is a common preference among breeders (Almasi et al., 2020). A primary function of GWAS is to discover essential genes that can facilitate marker-assisted selection (Almasi et al., 2020). Therefore, in recent years, several studies have been conducted on native Iranian sheep using genomic data, for example, studying the genomic association of weaning traits in Lori-Bakhtiari sheep (Almasi et al., 2020), a genome-wide association study of birth weight trait in Lori-Bakhtiari sheep (Ghasemi et al., 2019), another study of GWAS for growth and litter size traits (Gholizadeh et al., 2015), and mapping genomic regions under selection for fat storage characteristics (Moradi et al., 2012). While linear regression serves as a standard tool for analyzing marker effects in genome scans, it suffers from critical flaws. It frequently overestimates effects, treats markers as independent and normally distributed entities, and ignores the complex interplay that may exist between them (Sun et al., 2021, Bani Saadat et al., 2024). Thus, replacing classical methodologies with artificial intelligence approaches, such as machine learning algorithms, in GWAS can help overcome existing analytical errors. This shift allows for the precise identification of pivotal genetic markers and genes that control commercially critical traits. Machine learning (ML) methods are data-driven approaches that learn patterns from existing data and apply the discovered models to predict future scenarios (Shahinfar and Kahn, 2018). The introduction of machine learning methods such as random forest and gradient boosting into GWAS discussions and the high potential and efficiency of these methods, especially for large volumes of data, has many advantages, including: it has allowed the estimation of non-additive effects such as dominance and epistasis, as well as the examination of complex relationships between variables (such as interactions between markers) (Ghafouri et al., 2019). In the context of GWAS, machine learning frameworks like Random Forest are suitable for the dual purpose of selecting relevant markers and identifying genes with major effects on economic traits. Lamb meat serves as a vital animal protein source in numerous countries. Weaning traits in sheep, such as weaning weight (90 days), constitute economically significant characteristics that influence subsequent body weight gain at later ages. Furthermore, heavier lambs at weaning demonstrate a greater survival rate in subsequent life stages (Hatcher et al., 2010). Therefore, this study aimed to utilize modern technologies such as genome-wide association studies (GWAS) coupled with Random Forest to identify genomic regions and candidate genes associated with growth traits in Lori-Bakhtiari sheep.

Material and methods

Study population and phenotypic data

The study population comprised an experimental flock of Lori-Bakhtiari sheep from Sholi Breeding Station in Shahrekord, Iran. The Lori-Bakhtiari is one of the heaviest meat sheep breeds in the Middle East, primarily raised in western and southwestern regions of Iran (Maradi et al., 2012). Comprehensive information on the study population has been published previously (Almasi et al., 2020). The phenotypic data collected in the current study included adjusted weights at 90 and 180 days of age. Given the traits under investigation (weights at 90 and 180 days), the total number of records was equal to the number of animals in the pedigree (8,634). Data were collected over a period spanning from 1989 to 2017. A more comprehensive account of the phenotypic information has been reported by Almasi et al. (2020).

Sampling, DNA extraction, and quality control

For the present study, 132 Lori-Bakhtiari sheep were used. Among the sampled animals, 66 female sheep were selected based on EBV (33 with high breeding value and 33 with low breeding value) using a case-control method, and the remaining 66 sheep (8 male sheep and 58 female sheep) were randomly selected. The body weight EBVs were estimated by the methods as described by (Almasi et al., 2020). The sampled animals were the offspring of 68 sires and 120 dams. Blood samples (10 ml per sheep) were collected from the jugular vein. Genomic DNA was extracted from whole blood using the Cinagene kit (Cinagene, Iran). The samples were genotyped using the Illumina 50K Ovine SNPChip, which contained 51,135 SNP markers.

For quality control, the animals with call rates <99% were removed from further analysis. The SNPs with call rate <0.05, minor allele frequencies (MAF) <0.01, and Hardy-Weinberg disequilibrium at P $<10^{-6}$ or whose genetic location was unknown (Moradi et al., 2012) were removed from the analysis. Plink 1.9 beta and R (v 4.4.0) software were used for various stages of the quality control (Chang et al., 2015).

Random Forest Algorithm and Bioinformatics Analysis to Identify Genes and Biological Pathways

Significant SNPs for body weight in sheep were identified using the Random Forest (RF) method. Random Forest is an ensemble algorithm that combines

multiple decision trees to improve predictive performance (Breiman,2001). Essentially, this method creates a "forest" of multiple decision trees that collectively produce more reliable predictions than a single tree. To build each tree, a bootstrap sample (drawn with replacement) is taken from the training dataset, which includes both genotypic and phenotypic records. The tree is grown using this sample. At each node, a randomly selected subset of SNPs is considered. The SNP that best splits the data based on phenotype is chosen, and the node is branched according to its genotypes, thereby separating the phenotypes into subsequent nodes or leaves (Breiman,2001). Bootstrap resampling inherently creates samples where some original observations are left out (out-of-bag), and others are replicated.

To identify SNPs significantly associated with weaning weight (90 days) and weight at 180 days, we employed the random Forest package in R (Liaw and Wiener, 2015). The relative importance of each SNP was assessed based on the percent increase in mean squared error (%IncMSE). This statistical metric was calculated by randomly permuting the genotypes of each SNP in the out-of-bag (OOB) data and measuring the resulting increase in the model's prediction error. Permutation importance (Mean Increase in Prediction Error) was used to evaluate the contribution of each SNP. For every individual SNP, its genotypes were randomly shuffled in the out-of-bag (OOB) samples — the portion of the data not included in the bootstrap sample for each tree. After permutation, the modified OOB data were passed through the trained Random Forest model, and the resulting increase in the model's prediction error was measured. SNPs whose permutation caused a larger rise in prediction error were considered more important, as this indicates a stronger predictive relationship with the trait. This process provides a robust, model-based measure of variable importance that accounts for interactions and non-linear effects. SNPs exhibiting elevated %IncMSE scores were deemed more biologically pertinent to the prediction of both weaning and 180-day weights (Li et al., 2018; Nicodemus et al., 2010). We optimized three critical RF hyperparameters: the number of trees (ntree) from 2000 to 4000 (in steps of 1000), the number of features per split (mtry) as p/3, 0.5p/3, and 2p/3 (with p being the marker count), and the minimum terminal node size (nodesize) at 5, 10, and 15. Model tuning was guided by the root mean squared error (RMSE) metric, aiming to identify the hyperparameter combination that minimized it.

The performance of the Random Forest model and hyperparameter tuning were evaluated using 3-fold cross-validation combined with grid search. The dataset was randomly partitioned into three approximately equal subsets. In each iteration, two subsets (~66% of the data) were used for model training and the remaining subset (~33% of the data) for validation. This process was repeated three times so that each subset served as the validation set exactly once. Finally, the average RMSE across the three iterations was used as the performance metric to select the best hyperparameter combination. The optimal hyperparameter combination selected was ntree = 3000, mtry = p/3, and nodesize = 10.

All markers were ranked based on their importance scores derived from the Random Forest model trained with the optimal hyperparameter set. The optimal parameter for ranking SNPs in the Random Forest method was the percentage increase in mean squared error (%IncMSE), and in this study, this index was used to identify the top 50 markers influencing body weight. Following the identification of the top markers, the genes located within ±500 kb flanking regions for each marker were annotated. This was performed by querying the Oar_v4.0 sheep reference genome assembly using the NCBI (https://www.ncbi.nlm.nih.gov) and Ensembl (https://www.ensembl.org) databases (Yates et al., 2016), consistent with previous studies in sheep under similar genotyping density (Moradi et al., 2012). To explore the biological processes and molecular functions of the identified genes, Gene Ontology (GO) enrichment analysis was performed using the online DAVID bioinformatics resource and the NCBI database. Subsequently, gene network analysis was conducted using Cytoscape software (V 3.7.2).

Results

Phenotypic data and Quality control

The adjusted phenotype values and their descriptive statistics are presented in Table 1.

Table 1. Descriptive statistics of growth traits in Lori-Bakhtiari sheep

Phenotype	N	Mean (kg)	SD (kg)	CV%	Min	Max
W90	132	31.32	4.33	13.82	21.15	44.20
W180	132	39.32	5.35	13.60	27.99	68.82

Quality control filters were applied to the genomic data, and two samples with a genotype call rate < 0.99 were removed, along with eight more during file preparation. At the SNP level, 1,434 markers failed the missingness threshold (call rate < 0.95), 1,823 had a low MAF (< 0.01), and one violated HWE ($P < 1\times10^{-6}$). The final analytic dataset consisted of 44,796 SNPs genotyped across 122 individuals.

SNP identification and ranking

Figure 1 displays the Manhattan plots for weaning weight (90-day weight) and 180-day weight, alongside the genome-wide distribution of SNP importance values

based on the percentage increase in mean squared error (%IncMSE). Figure 1 shows that 50 SNPs exceeded the predefined importance cutoff (% IncMSE > 2) and are distributed across various chromosomes. The analysis identified two SNPs with the greatest influence on weaning weight: OAR8_93131673.1 on chromosome 8 (%IncMSE = 6.80) and OAR4_14230002.1 on chromosome 4 (%IncMSE = 5.45). For 180-day weight, the top-ranking SNPs were OAR12_59600031.1 on chromosome 12 and s56411.1 on chromosome 11, both with a %IncMSE value of 6.26. The detailed information for the top 10 SNPs associated with each weight trait is provided in Table 2. The RF-based ranking of SNP effects is visualized in Figure 2. The analysis classified SNP effects as positive, neutral, or negative, revealing distinct distributions for each trait: weaning weight (26.91% positive, 45.27% neutral, 27.80% negative) and 180-day weight (40.33% positive, 19.43% neutral, 40.23% negative).

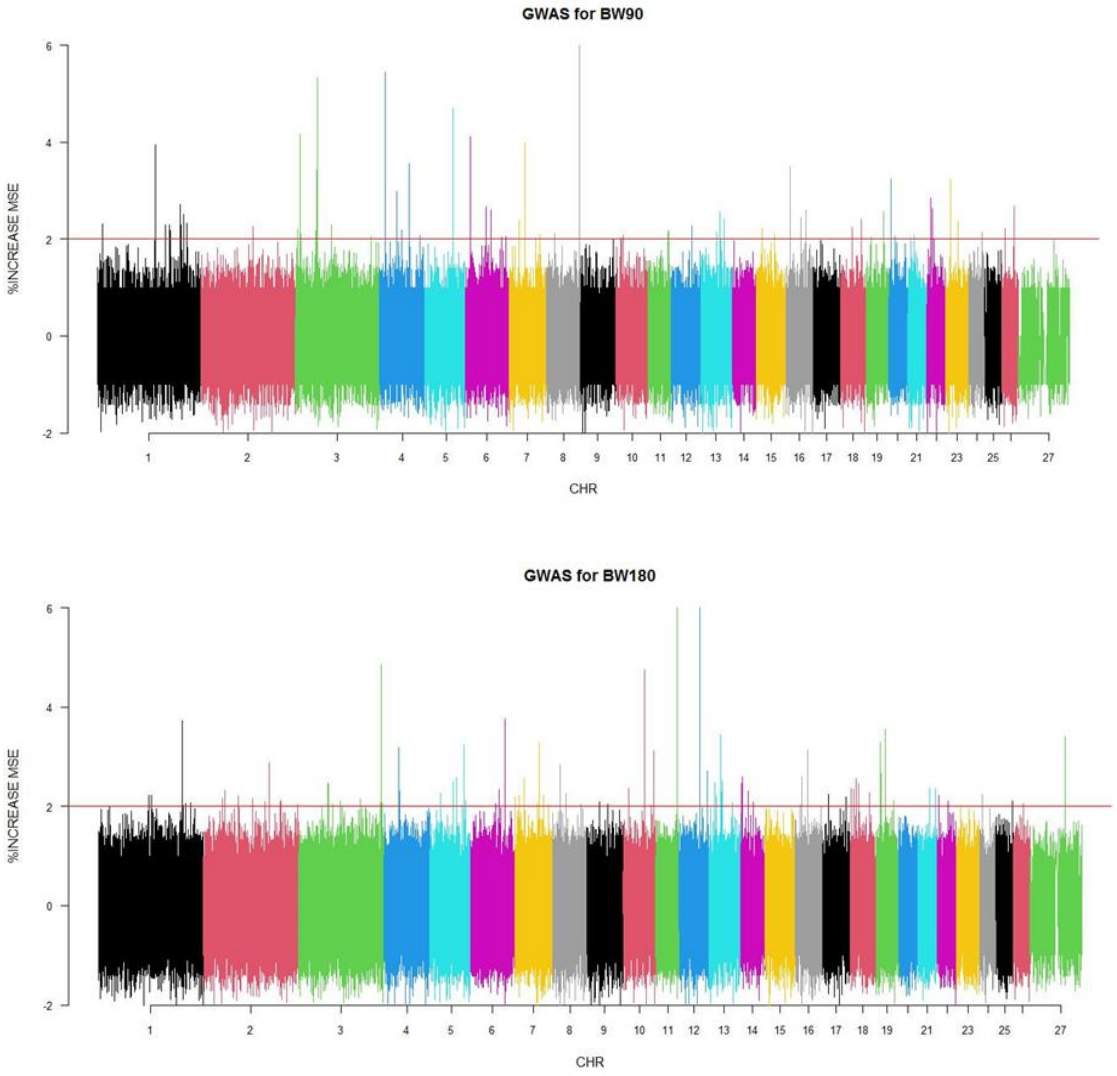

Figure 1. Manhattan plots showing the genome-wide association results for weaning weight and 180 days of age based on Random Forest analysis: SNP position in the genome is shown on the X-axis, and % Increase in MSE values is plotted on the Y-axis. The red line indicates the significance threshold (%IncMSE > 2) and was later examined for further analysis.

Detected genes and functional analysis

We identified 59 and 58 candidate genes for weaning weight and 180-day weight, respectively, by investigating genes located within or near the 50 significant SNPs. Among the 50 significant SNPs identified for each weight trait, Table 2 provides detailed information on the top 10 markers, including their chromosomal position, importance values (%IncMSE), and associated upstream/downstream genes. The positional mapping revealed high-potential candidate genes harboring significant SNPs: *DYNC1I1*, *PAX8*, and *KGD4* for weaning weight, and *CNTN6*, *MYRIP*, and *CACNA1I* for 180-day weight. These genes are

proposed as strong candidates influencing body weight in sheep.

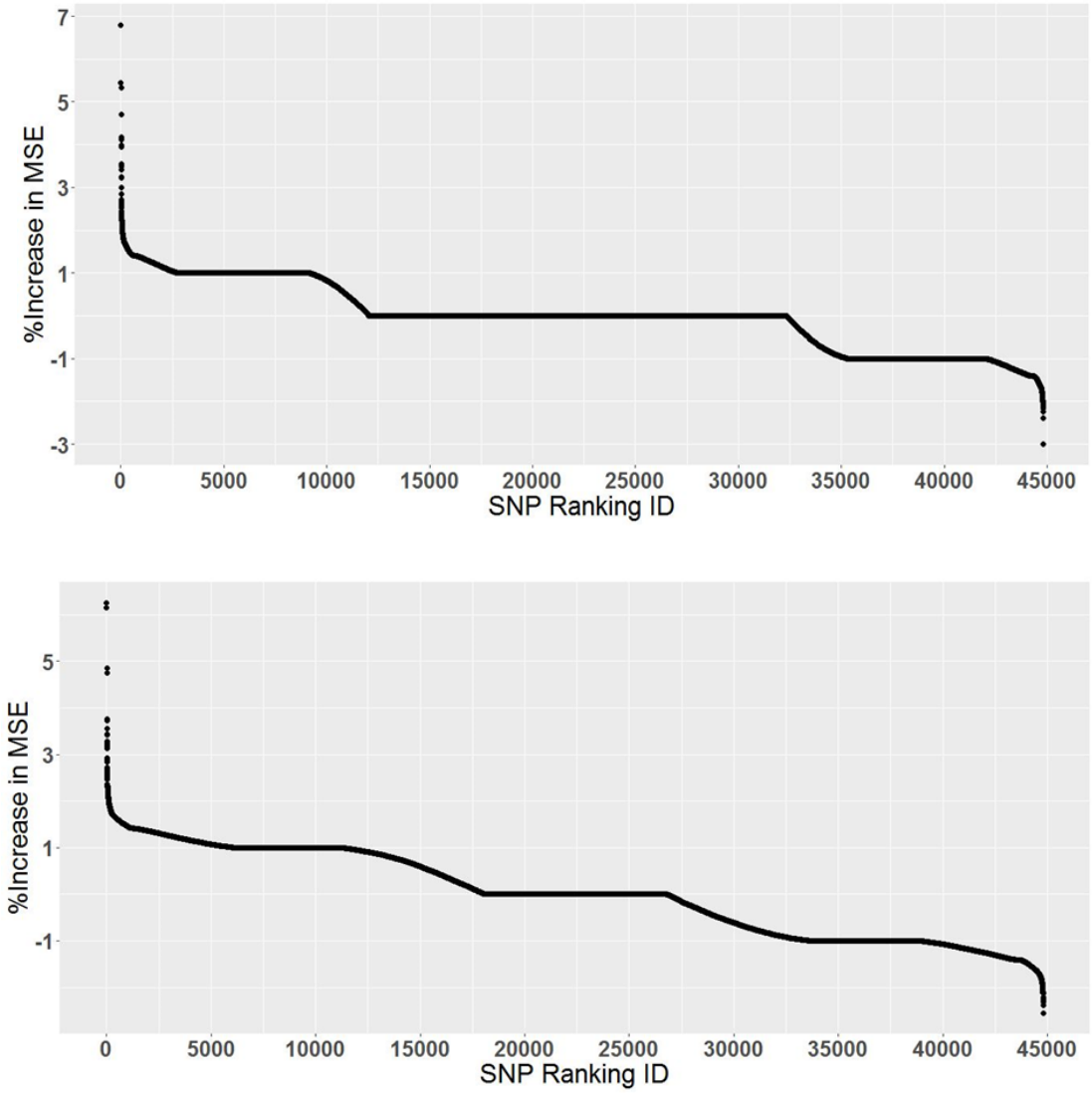

Figure 2. Distribution of importance values (%IncMSE) of SNPs using random forest analysis based on ranking from highest %IncMSE to lowest: The vertical axis represents the value (%IncMSE) in the range (-3 to 7), indicating the relative importance of each SNP for predicting weaning weight and 180-day weight, respectively. The horizontal axis represents the ranking of all 44,796 SNP markers from highest to lowest importance.

We performed gene ontology (GO) enrichment analysis on the candidate gene sets (59 genes for weaning weight and 58 for 180-day weight). This analysis identified significant associations with biological processes involved in weight regulation, as summarized in Table 3. Gene ontology analysis revealed several highly significant biological processes, such as post-embryonic development (P=.0019) and protein tyrosine kinase signaling at the cell surface (P=0.0026), alongside other relevant pathways, including cell-cell junction organization (P=0.0172).

Gene network analysis

The gene network, derived from our analysis and visualized with Cytoscape software, is depicted in Figure 3. A functional interaction network highlighting the 10 central genes with maximal connectivity was generated from the combined candidate genes for both weight traits (Figure 3). These central genes act as critical hubs within the network, typically playing pivotal roles in biological pathways and regulating gene expression networks. In the gene network analysis, the core genes identified for weaning weight were *ABL1* (ABL proto-oncogene 1, non-receptor tyrosine kinase) and *INPPL1* (inositol polyphosphate phosphatase like 1). For 180-day weight, the central hubs were ROS1 (ROS proto-oncogene 1, receptor tyrosine kinase), *HCK* (HCK proto-oncogene, Src family tyrosine kinase), and *ZFAND4* (zinc finger AN1-type containing 4). These genes formed a tightly interconnected cluster (Figure 3). These findings indicate that the identified hub genes likely act in concert

within common regulatory pathways affecting weaning and 180-day weights.

Table 2. Top 10 SNPs associated with weaning weight and 180-day weight identified by random forest analysis, chromosomal location, significance values (%IncMse), and candidate genes reported in these genomic regions

Rank	Chr	SNP	Position (bp)	%IncMse	Downstream gene	Gene	Upstream gene
BW90							
1	8	OAR8_93131673.1	86431305	6.798691	-	-	-
2	4	OAR4_14230002.1	14016885	5.450899	-	*DYNC1I1*	*SLC25A13*
3	3	OAR3_63334035.1	59728134	5.328323	-	*PAX8*	*PSD4*
4	5	OAR5_82093890.1	74551788	4.704493	-	-	-
5	3	s08158.1	13765098	4.168318	*NDFA8*	-	*TTLL11*
6	6	OAR6_17489056.1	14706831	4.115636	-	-	-
7	7	OAR7_47161420.1	42551974	3.993598	-	-	-
8	1	OAR1_165108584_X.1	153375890	3.950621	-	-	-
9	4	OAR4_83461583.1	78666709	3.552691	*NUDCD3*	-	*CAMK2B*
10	16	OAR16_11445590.1	10583923	3.488471	*CDK7*	*KGD4*	*CENPH*
BW180							
1	12	OAR12_59600031.1	53614033	6.257759	-	-	-
2	11	s56411.1	56536218	6.154743	-	-	-
3	3	s50765.1	216832845	4.850936	-	*CACNA1I*	-
4	10	OAR10_58051560.1	56906392	4.751834	-	-	-
5	6	s69065.1	91631661	3.75515	*SDAD1*	-	*RAT3*
6	1	OAR1_238200651.1	220937683	3.722695	-	-	-
7	19	OAR19_26747265.1	25306966	3.558213	-	*CNTN6*	-
8	13	OAR13_34123985.1	30972822	3.439149	*CUNB*	-	*VIM*
9	27	OARX_109922971.1	90782534	3.412637	-	-	-
10	19	s24573.1	13068309	3.288347	-	*MYRIP*	*EIF1B*

Table 3. Ontology analysis of reported genes within or adjacent to 50 significant SNP markers identified using the random forest method for the traits weaning weight and 180-day weight

Category	Term	Count	%	P-Value	Genes
BW90					
GOTERM_BP_ALL	GO:0009791~post-embryonic development	3	5.555556	0.001942	*ABL1*, *INPPL1*, *PLAGL2*
GOTERM_BP_ALL	GO:0051641~cellular localization	11	20.37037	0.003784	*NUDCD3, IMMP2L, RIMS3, EXOC7, AGK, ABL1, SYNDIG1, TM9SF4, DYNC1I1, SEC31A, LRRC1*
GOTERM_BP_ALL	GO:0051649~establishment of localization in cell	8	14.81481	0.010409	*RIMS3, AGK, ABL1, SYNDIG1, TM9SF4, DYNC1I1, SEC31A, LRRC1*
GOTERM_BP_ALL	GO:0008104~protein localization	8	14.81481	0.011968	*NUDCD3, IMMP2L, EXOC7, AGK, ABL1, TM9SF4, SEC31A, LRRC1*
GOTERM_BP_ALL	GO:0006629~lipid metabolic process	7	12.96296	0.017116	*BDH1, AGK, SCD5, INPPL1, PLCE1, PLAGL2, PLBD1*
GOTERM_BP_ALL	GO:0051179~localization	14	25.92593	0.018429	*HIP1, EXOC7, AGK, INPPL1, TM9SF4, LRRC1, NUDCD3, IMMP2L, RIMS3, SYNDIG1, ABL1, PLCE1, DYNC1I1, SEC31A*
GOTERM_BP_ALL	GO:0070371~ERK1 and ERK2 cascade	2	3.703704	0.031186	*ABL1, INPPL1*
GOTERM_BP_ALL	GO:0072698~protein localization to microtubule cytoskeleton	2	3.703704	0.057444	*NUDCD3, ABL1*
BW180					
GOTERM_BP_ALL	GO:0007169~cell surface receptor protein tyrosine kinase signaling pathway	5	9.259259	0.002628	*SHC4, EFNB2, HCK, HGF, ROS1*
GOTERM_BP_ALL	GO:0007167~enzyme-linked receptor protein signaling pathway	5	9.259259	0.009469	*SHC4, EFNB2, HCK, HGF, ROS1*
GOTERM_BP_ALL	GO:0034330~cell junction organization	1	7.407407	0.017242	*GABRB2, PCLO, CDH11, PCDH15*
GOTERM_BP_ALL	GO:0000902~cell morphogenesis		7.407407	0.036796	*EFNB2, HGF, CDH11, PCDH15*
GOTERM_BP_ALL	GO:0042325~regulation of phosphorylation	4	7.407407	0.066934	*IL7, HGF, ROS1, SOCS4*
GOTERM_BP_ALL	GO:0032502~developmental proces	11	20.37037	0.070681	*SHC4, EFNB2, GABRB2, HCK, RBFOX1, PCLO, IL7, HGF, CDH11, PCDH15, ROS1*
GOTERM_BP_ALL	GO:0019220~regulation of phosphate metabolic process	4	7.407407	0.092528	*IL7, HGF, ROS1, SOCS4*

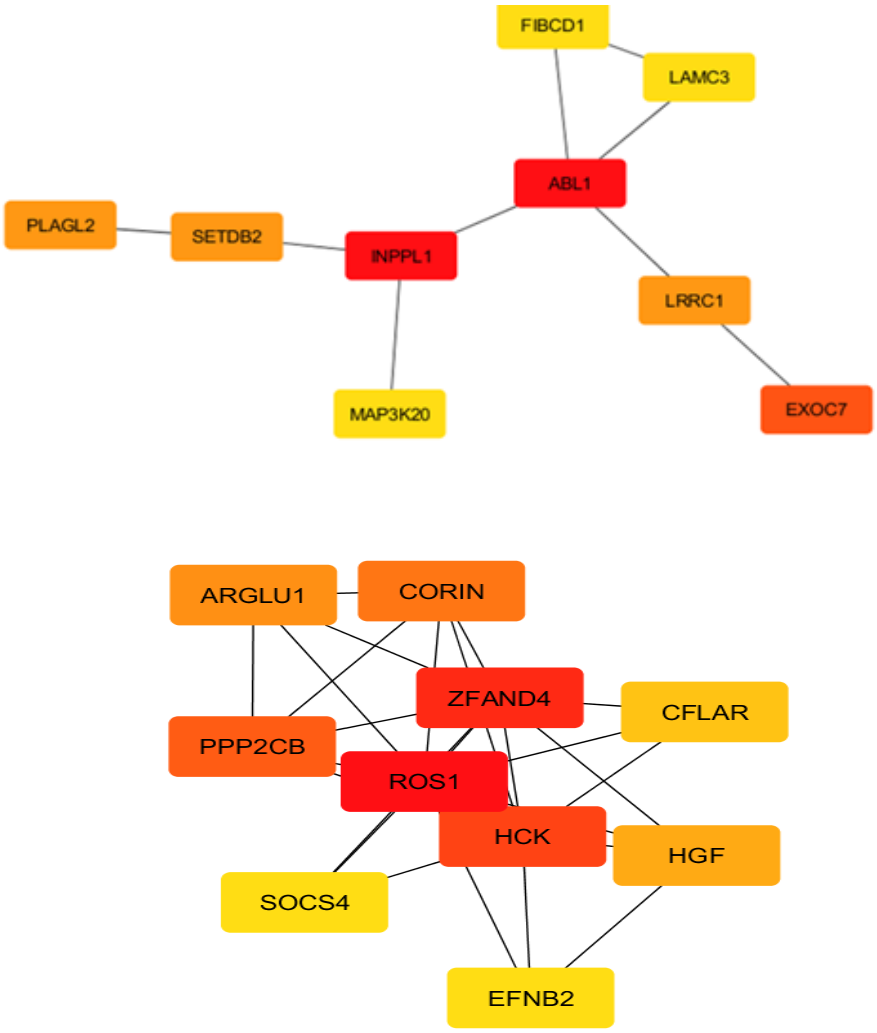

Figure 3. Interaction network of candidate genes associated with weaning weight and 180 days of age in Lori-Bakhtiari sheep. The network was constructed using 10 central genes identified through random forest analysis. Nodes represent genes (colored by functional clusters).

Discussion

Body weight is a pivotal composite trait in sheep, fundamentally influencing profitability in meat production systems. It serves as a key determinant of growth rate, feed efficiency, and carcass yield, directly impacting farm income (Kenyon et al., 2014). However, its genetic architecture is complex, being controlled by numerous quantitative trait loci (QTLs) and genes with small additive effects, alongside significant environmental influences (Montaldo et al., 2011). Unraveling this complexity is crucial for effective genomic selection. Recent advances in high-throughput genotyping and machine learning, such as Random Forest analysis, have proven powerful in identifying candidate genes and SNPs associated with polygenic traits like weight, offering a refined approach beyond traditional GWAS (Gonzalez-Recio et al., 2014). Identifying key genetic variants, as performed in this study, provides valuable markers for selection programs aimed at enhancing growth traits while maintaining genetic diversity.

Among the top 10 SNPs and the genes identified around them for weaning weight, three important genes, including *DYNC1I1*, *PAX8*, and *KGD4*, *CNTN6*, and *MYRIP* for 180-day weight associated with the top-most significant SNPs in the RF method, are involved in biological processes of the body weight traits. The *DYNC1I1* gene, dynein cytoplasmic 1 intermediate chain 1, is located on chromosome 4 of sheep and is involved in gene ontology for cellular localization and establishment of localization in the cell. Although the *DYNC1I1* gene has not been explicitly studied in the context of farm animal growth, its inclusion as a candidate is supported by multiple lines of indirect evidence. First, its product is a core component of the cytoplasmic dynein complex, a master regulator of

intracellular transport and cell division processes fundamental to tissue growth and development in all mammals (Roberts et al., 2013). Second, systems biology approaches in livestock have identified broader cellular pathways, such as vesicle transport and intracellular protein trafficking, as being significantly enriched for genes affecting body composition traits (Zhang et al., 2025), placing *DYNC1I1* within a critical functional network. The *PAX8* gene, paired box 8, is located on chromosome 3 of sheep and is involved in a master transcriptional regulator essential for the development and function of the thyroid gland. In ruminants, the thyroid axis is a well-documented key regulator of postnatal growth and development (López-Corrales et al., 1999). The *KGD4* gene (alpha-ketoglutarate dehydrogenase subunit 4) in gene ontology was involved in cell proliferation and tissue growth. While this specific gene has not been directly studied in farm animal genomics. The identification of *CNTN6* as a candidate gene for 180-day weight suggests a potential neuroregulatory mechanism influencing growth. *CNTN6* (contactin-6) encodes a neuronal adhesion molecule critical for axon guidance and synapse formation in the developing brain (Zuko et al., 2013). Notably, the hypothalamus, a key center for regulating appetite, energy balance, and the growth hormone axis, expresses such molecules. Genetic variation in *CNTN6* could therefore affect the development or function of hypothalamic circuits that govern systemic growth. This hypothesis is supported by genomic studies in humans revealing pleiotropy between loci involved in neural development and body mass regulation (Watanabe et al., 2019). Another candidate gene, *MYRIP*, was identified for 180-day weight. *MYRIP* (myosin VIIA and Rab interacting protein) encodes a protein that links the actin cytoskeleton to the plasma membrane in skeletal muscle cells, playing a crucial role in structural stability and force transmission during contraction (El-Amraoui et al., 2002). The *ABL1* gene also emerged as a key candidate. Beyond its established role in cellular growth signaling, this gene has been implicated in adipogenesis and fat metabolism in Yorkshire pigs (Li et al., 2021) and is associated with the meat-to-fat ratio in swine (Falker-Gieske et al., 2019), suggesting a pleiotropic influence on body composition. *INPPL1* acts as a key modulator in the insulin signaling pathway process, and its dysregulation is associated with metabolic disorders such as insulin resistance in model organisms (Lehtonen, 2020). The receptor tyrosine kinase gene *ROS1* was identified as a top candidate for 180-day weight. Although not previously reported in livestock growth studies, *ROS1* is a well-characterized oncogene that activates major mitogenic and anti-apoptotic pathways such as MAPK/ERK and PI3K/AKT upon ligand binding (Davies & Doebele, 2012). These pathways are fundamental regulators of cell proliferation and hypertrophy in skeletal muscle. Furthermore, systems genetics approaches in cattle have highlighted the importance of tyrosine kinase signaling networks in growth and carcass traits (Okada et al., 2018). The *HCK* gene represents not merely a fundamental immunological component, but also a promising candidate marker for genomic selection. In livestock, naturally occurring SNPs within the HCK locus have been shown to modulate the efficiency of this innate immune pathway. This functional impact is corroborated by genetic association studies that link specific *HCK* polymorphisms to enhanced resistance against major diseases, such as bovine tuberculosis (Finlay et al., 2012). Emerging evidence from livestock genomics implicates *ZFAND4* as a novel candidate gene influencing reproductive efficiency. In cattle, transcriptomic analysis has revealed that *ZFAND4* is differentially expressed in the endometrium during the critical period of maternal recognition of pregnancy, suggesting a potential role in uterine receptivity and early embryonic development (Mohan et al., 2020). Furthermore, integrative systems genetics analysis in sheep has positioned *ZFAND4* within a key co-expression network module strongly associated with prolificacy, indicating its involvement in pathways governing ovulation rate or embryonic survival (Abdoli et al., 2021). While direct functional studies in farm animals are still lacking, the conserved biological role of *ZFAND4* as an AN1-type zinc finger protein involved in cellular stress response and protein homeostasis (Kushwaha et al., 2021) provides a plausible mechanistic link.

Compared to the study by Almasi et al., 2020, which used the classic GWAS method with Plink on the same data and five SNPs on chromosomes 2, 3, 4, 12, and 22 were identified with the *ANGPTL7***,** *mTOR* and *WDR11* genes, all these markers exhibited a positive increase in mean square error (%IncMSE) using the present random forest method; however, their rankings differed. This discrepancy stems from the random forest method's focus on nonlinear relationships and interaction effects, potentially highlighting novel genomic regions, whereas linear methods primarily emphasize independent effects (Sun et al., 2021). Together, these findings suggest that integrating multiple methods can yield a more comprehensive understanding of the genetic architecture underlying growth traits.

This investigation demonstrates that machine learning algorithms, including Random Forest, serve as effective tools in genomic screening for pinpointing significant genetic markers and discerning genes that influence economically valuable traits. In contrast to conventional statistical approaches, this technique operates without prerequisites such as defining the underlying population distribution, calculating estimators (e.g., confidence intervals or p-values), or testing null hypotheses, thereby constituting a distribution-free and assumption-lean methodology. The single-nucleotide polymorphisms (SNPs) and candidate genes uncovered through this process provide a critical understanding of the genetic foundations underlying the trait. Consequently, these results pave the way for the implementation of more accurate and targeted genetic

enhancement strategies within sheep breeding programs (Bani Saadat et al., 2024).

While the findings of this study provide a valuable genomic framework, several important limitations must be acknowledged to guide future research. The scope of the investigation was constrained by a limited sample size and a lack of racial diversity within the cohort, suggesting that validation in larger, independent, and more genetically diverse populations is essential to conclusively confirm the identified candidate genes. Beyond validation, a critical and necessary next step is to move from genetic association to functional characterization. Dedicated experimental studies are urgently required to elucidate the precise biological roles of the highlighted genes and to determine the mechanistic impact of their associated single-nucleotide polymorphisms (SNPs) on the trait in question.

Conclusions

This research, leveraging body weight data and a genome-wide panel of 51,135 single-nucleotide polymorphisms (SNPs) in Lori-Bakhtiari sheep, establishes the efficacy of machine learning algorithms, notably RF, in pinpointing key genetic determinants of growth. The analysis yielded the ten most influential SNPs, alongside 59 and 58 candidate genes linked to weaning weight and 180-day weight, respectively, with multiple genes representing novel discoveries. In conclusion, these findings provide a critical genetic framework for future investigations, facilitating the refinement of genomic regions associated with vital production traits and advancing the comprehension of their complex biological underpinnings. The study underscores the significant value of integrating advanced analytical techniques, such as RF, with traditional GWASs to elucidate intricate, non-linear genetic architectures in indigenous sheep populations.

Conflict of interest

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

Acknowledgments

The authors express their sincere gratitude to the faculty and staff of the University of Tehran's College of Agricultural Sciences and Natural Resources for their valuable assistance and support throughout this study. This research was made possible through the financial support of Bu-Ali Sina University, Iran, which is gratefully acknowledged.

References

Abdoli, R., Mirhoseini, S.Z., Ghavi Hossein-Zadeh, N., Zamani, P., 2021. Genome-wide association study and gene co-expression network reveal molecular pathways and candidate genes associated with litter size in sheep. *Journal of Animal Breeding and Genetics* 138, 442-455.

Almasi, M., Zamani, P., Mirhoseini, S.Z., Moradi, M.H., 2020. Genome-wide association study of weaning traits in Lori-Bakhtiari sheep. *Annals of Animal Science* 20, 811-824.

Bani Saadat, H., Vaez Torshizi, R., Manafiazar, G., Masoudi, A.A., Ehsani, A., Shahinfar, S., 2024. Comparing machine learning algorithms and linear model for detecting significant SNPs for genomic evaluation of growth traits in F2 chickens. *Journal of Agricultural Science and Technology* 26(6), 1261-1274.

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

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4, 1-16.

Davies, K.D., Doebele, R.C., 2012. Molecular pathways: ROS1 fusion proteins in cancer. *Clinical Cancer Research* 18, 4040-4045.

El-Amraoui, A., Schonn, J.S., Küssel-Andermann, P., Blanchard, S., Desnos, C., Henry, J.P., Wolfrum, U., Darchen, F., Petit, C., 2002. MYRIP, a novel Rab effector, enables myosin VIIa recruitment to retinal melanosomes. *EMBO Reports* 3, 463-470.

Falker-Gieske, C., Blaj, I., Preuß, S., Bennewitz, J., Thaller, G., Tetens, J., 2019. GWAS for meat and carcass traits using imputed sequence level genotypes in pooled F2-designs in pigs. *G3: Genes, Genomes, Genetics* 9, 2823-2834.

Finlay, E.K., Berry, D.P., Wickham, B., Gormley, E.P., Bradley, D.G., 2012. A genome-wide association study of bovine tuberculosis resistance in dairy cattle. *BMC Genomics* 13, 631.

Ghafouri, F., Alipour, S., Mohammadian Jashouqani, P., 2019. Application of machine learning approach and its subset algorithms in estimating genomic breeding values. *Scientific-Extensional (Professional) Domestic* 20, 19-29. [In Persian].

Ghasemi, M., Zamani, P., Vatankhah, M., Abdoli, R., 2019. Genome-wide association study of birth weight in sheep. *Animal*13, 1797-1803.

Gholizadeh, M., Rahimi-Mianji, G., Nejati-Javaremi, A., 2015. Genome wide association study of body weight traits in Baluchi sheep. *Journal of Genetics* 94, 143-146.

Gonzalez-Recio, O., Forni, S., Gianola, D., Weigel, K.A., 2014. Genome-wide prediction of discrete traits with machine learning methods. *Journal of Animal Science* 92, 52-52.

Hatcher, S., Eppleston, J., Thornberry, K.J., Watt, B., 2010. High Merino weaner survival rates are a function

of weaning weight and positive post-weaning growth rates. *Animal Production Science* 50, 465-472.

Kenyon, P.R., Blair, H.T., 2014. Foetal programming in sheep – Effects on production. *Small Ruminant Research* 118, 16-30.

Lehtonen, S. 2020. SHIPping out diabetes-Metformin, an old friend among new SHIP2 inhibitors. *Acta Physiologica* 228, e13349.

Liaw, A., Wiener, M., 2015. randomForest: Breiman and Cutler's random forests for classification and regression. *R package version* 4, 1-14.

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.

Li, D., Huang, M., Zhuang, Z., Chen, Z., Wang, S., Zhang, H., Huang, L., Ding, X., 2021. Genomic analyses revealed the genetic difference and potential selection genes of growth traits in two duroc lines. *Frontiers in Veterinary Science* 8, 725367.

López-Corrales, N.L., Sonstegard, T.S., Smith, T.P.L. 1999. Physical mapping of the bovine, caprine and ovine homologues of the paired box gene PAX8. *Cytogenetics and Cell Genetics* 84, 179-181.

McRae, K.M., McEwan, J.C., Dodds, K.G., Gemmell, N.J., 2014. Signatures of selection in sheep bred for resistance or susceptibility to gastrointestinal nematodes. *BMC Genomics* 15, 637.

Mohan, H., Tomar, A.S., Kumar, S., Rautela, R., Singh, S.P., Bharti, V.K., Maurya, V.P., 2020. Transcriptomic profiling of endometrial tissues at the time of maternal recognition of pregnancy in buffalo (*Bubalus bubalis*). *Animal Biotechnology* 33, 1265-1276.

Montaldo, H.H., Meza-Herrera, C.A., 2011. Current challenges and future opportunities for genetic improvement of small ruminants in developing countries. *Livestock Science* 136, 1-12.

Moradi, M.H., Nejati-Javaremi, A., Moradi-Shahrbabak, M., Dodds, K.G., 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, 1-15.

Nicodemus, K.K., Malley, J.D., Strobl, C., Ziegler, A., 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics* 11, 1-16.

Okada, D., Endo, S., Matsuda, H., Ogawa, S., Taniguchi, Y., Katsuta, T., Watanabe, T., Iwaisaki, H., 2018. An intersection network based on combining SNP coassociation and RNA coexpression networks for feed utilization traits in Japanese Black cattle. *Journal of Animal Science* 96, 2553-2566.

Roberts, A.J., Kon, T., Knight, P.J., Sutoh, K., Burgess, S.A., 2013. Functions and mechanics of dynein motor proteins. *Nature Reviews Molecular Cell Biology* 14, 713-726.

Shahinfar, S., Kahn, L., 2018. Machine learning approaches for early prediction of adult wool growth and quality in Australian Merino sheep. *Computers and Electronics in Agriculture* 148, 72-81.

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

Watanabe, K., Stringer, S., Frei, O., Mirkov, M.U., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., Posthuma, D., 2019. A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics* 51, 1339-1348.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., 2016. Ensembl 2016. *Nucleic Acids Research* 44, 710-716.

Zamani, P., 2017. Statistical properties of proportional residual energy intake as a new measure of energetic efficiency. *Journal of Dairy Research* 84, 248-253.

Zhang, Y., Cai, W., Zhang, Q., Li, Q., Wang, Y., Peng, R., Yin, H., Hu, X., Wang, Z., Zhu, B., Gao, X., Chen, Y., Gao, H., Xu, L., Li, J., Zhang, L., 2025. Integrated analyses of genomic and transcriptomic data reveal candidate variants associated with carcass traits in Huaxi cattle. *Journal of Integrative Agriculture* 24, 3169-3184.

Zuko, A., Kleijer, K.T., Oguro-Ando, A., Kas, M.J., van Daalen, E., van der Zwaag, B., Burbach, J.P., 2013. Contactins in the neurobiology of autism. *European Journal of Pharmacology* 719, 63-74.