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

Studying the genetic loci related to the bovine leukemia virus using random forest method and genomic data

Sara Nezhadi[1], Mohammad Moradi Shahrbabak[1], Hossein Moradi-Shahrbabak[1], Hossein Bani-Saadat[2]*

[1]*Department of Animal Science, 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:
moradim@ut.ac.ir

Received: 28 May 2025,
Received in revised form: 01 Jul 2025,
Accepted: 12 Jul 2025,
Published online: 13 Jul 2025,

ORCID
Sara Nezhadi
0009-0006-5988-1617
Mohammad Moradi Shahrbabak
0000-0002-5255-609X
Hossein Moradi Shahrbabak
0000-0002-6680-7662
Hossein Bani-Saadat
0000-0001-9034-0372

Abstract Bovine leukemia virus (BLV) is a causative agent of bovine leukosis, which, due to its long incubation period, can spread widely within a herd before clinical symptoms appear, causing significant economic losses. This study used a supervised machine learning method called random forest to identify genomic regions associated with BLV. The non-parametric nature of this method allows for the creation of predictive models without the need for initial statistical assumptions; whereas the standard Genome-wide association studies (GWAS) methods are usually based on single-variable hypothesis tests and cannot account for correlations resulting from connectivity imbalance or the combination of multiple markers. In this study, the genotyping data of 145 Holstein cows (77 BLV-positive, 68 healthy) after quality control by using the PLINK (v 1.02), which resulted in 23,910 Single nucleotide polymorphisms (SNPs) were analyzed. Random forest analyses on the mentioned data included three hyperparameters: mtry (0.5(p/3), (p/3), 2(p/3)), ntree (2000, 3000, 4000), and nodesize (5, 10, 15), where p is equal to the total number of SNPs (23,910). To find the best SNPs, the Mean Decrease Accuracy (MDA) index (> 1.89) was used which resulted in the selection of 50 SNPs. Genomic enrichment analyses showed that genes associated with the top 50 SNPs are predominantly involved in Positive Regulation, Intracellular Signaling, Apoptosis and Cell Death, Signal Transduction, Metabolic Processes, and Cell Differentiation and Development. In total, 82 genes were identified, including hub genes such as *MYC*, *RABIF*, *IRS1*, *TRAPPC9*, *MAPK8*, *HTT*, *SNX9*, *BCLAF1*, *XRN1*, and *LSM6*.

Keywords: bovine leukemia virus (BLV), dairy cattle, genomic prediction, random forest, susceptibility loci

Introduction

The bovine leukemia virus is a retrovirus that is known as the causative agent of enzootic bovine leucosis, which is a common neoplastic disease in dairy cattle. Bovine leucosis is usually disregarded in breeding processes because of its lack of specific clinical symptoms; and due to the long latency period, which may lead to widespread in the herd (Bongers, 2023). In the United States, the prevalence rate of bovine leucosis is more than forty percent, and the annual economic impact of reduced milk production is 500 million dollars. Therefore, it is important to accurately diagnose the disease in its early stages to control disease outbreaks and economic losses. (Lv et al., 2024).

Genome-wide association studies (GWAS) analyze hundreds of thousands of genetic variants across the genome (Uffelmann et al., 2021) with the main objective to identify the variants that cause a specific trait or disease, either individually or in combination (Botta et al., 2014). The most common variants studied in GWAS (Uffelmann et al., 2021) are single-nucleotide polymorphisms (SNP). The SNPs are variations in the base pairs of the DNA, and it has

been proven that SNP profiles identify various types of diseases in GWAS. The identification and selection of SNPs related to economic traits are the most important tasks in the analysis of GWAS data. However, the large dimensions of the data and the lack of association of a significant portion of SNPs with the disease make this task challenging (Nguyen et al., 2015). Since the data dimensions are large, stringent thresholds must be adopted for calculating the error rates, which leads to insufficient detection and increases the likelihood that SNPs with small effects associated with the trait will not be identified (Silva et al., 2022). Standard GWAS methods are usually based on univariate hypothesis tests and therefore cannot account for correlations resulting from linkage disequilibrium or the combination of multiple markers (Botta et al., 2014).

Several methods have been used to identify SNPs with the most distinction among thousands of markers in commercially available SNP chip tools (Schiavo et al., 2020). Random forest is one of the machine learning algorithms that has been proposed for this purpose (Enoma et al., 2022), and it can be used to estimate the importance of each SNP in classification and rank all SNPs in order of importance (Schiavo et al., 2024). It is composed of a collection of decision trees, each of which is developed with a bootstrapped subsample of the training dataset. Therefore, it is anticipated that a group of algorithms will fit well for modelling the non-linear biological functions evident in genetic data such as SNP GWAS (Enoma et al., 2022).

In simulations conducted by Meng et al. (2009), it was shown that as the genetic effect becomes stronger; the impact of LD on the performance of random forests also becomes stronger. In most genetic models that were simulated, the revised IM (importance measure) demonstrated better performance compared to the original IM when used with either the revised random forest method or the original random forest method. Additionally, SNPs in LD with noise SNPs had little effect on performance, suggesting the advantages of including all SNPs in analyses (Meng et al., 2009). Therefore, this study aimed to identify the markers and genes associated with bovine leukosis through genomic screening using the random forest algorithm.

Material and methods

Phenotypic and genotypic data

In the present research, the data (SNP chip) from the genotyping project of superior Holstein cows of a large dairy farm in Isfahan were used. The groups of sick and healthy animals were distinguished based on phenotypic data which were obtained from the ELISA blood sample test on Holstein female cows aged 3 to 7 years. In selecting animals for blood samples, non-related animals, animals that represented the diversity of the breed as much as possible (diversity in milk production and exposure to disease and stress factors), and animals that indicated different stages of leukosis were chosen. Blood samples were taken (5-7 cc) from the jugular vein and mixed with 0.5 mL of EDTA in vacuum tubes. The samples were immediately transferred to the laboratory under temperature-controlled conditions and stored at 4°C until plasma and DNA extraction. The genotyping of 145 Holstein cows (77 diseased and 68 healthy) was performed using Illumina GeneSeek arrays (GGP Bovine LD v4 30k).

Quality control

Quality control of the data was performed using the PLINK v1.02 software for 29,776 SNPs. No animals were removed with MIND > 0.1. A total of 334 SNPs with GENO > 0.05, 9 SNPs with H-W < 1e - 6, and 5,523 SNPs with MAF < 0.01 was removed. Finally, 23,910 SNPs and 145 cows were used in the random forest analyses.

Random forest in genomic screening

The random forest (RF) algorithm (Breiman, 2001) was employed to identify the SNPs associated with BLV. (1) For each tree in the forest, a training subset was drawn randomly with replacement from the original dataset, comprising approximately two-thirds of the total samples (in-bag data). The remaining one-third (out-of-bag, OOB) samples were retained for internal validation. (2) At each node of a decision tree, a random subset of SNPs (mtry) was evaluated as potential splitting variables. (3) A binary decision tree was grown by recursively partitioning the in-bag samples. At each node, the SNP that maximally reduced the MDA between the observed and predicted phenotypes (e.g., BLV infection status) was selected to split the node. This process continued until terminal nodes were reached, where no further reduction in MDA was achievable. (4) The prediction accuracy for each tree was estimated using its OOB samples, yielding an OOB error rate. The SNP variable importance (VIM) was quantified by permuting each SNP's value within the OOB samples and calculating the resultant increase in MDA. Higher MDA values indicated stronger associations between SNPs and BLV-related outcomes. (5) Steps 1–4 were repeated to generate a forest of N trees. The final importance score for each SNP was computed as the average MDA across all trees in which the SNP was included (Breiman, 2001).

To implement the random forest (RF) algorithm, we need to tune several key parameters. The first and most influential parameter is mtry - the number of candidate variables considered at each node split, which critically determines the complexity of the final model. Larger mtry values result in fewer variables being incorporated into the tree, producing sparser solutions while simultaneously reducing the variance-reducing effect of randomization. Conversely, smaller mtry values decrease the correlation between the trees and enhance the potential for variance reduction through bagging (Goldstein et al., 2011).

The second crucial parameter is ntree (number of trees in the forest). Unlike mtry, there is no definitive "optimal" value for ntree - generally, higher values yield better results. However, increasing the number of trees requires greater computational resources and demonstrates diminishing returns with very large values (Goldstein et al., 2011).

The third parameter employed in this analysis is node size. The splitting process is halted when a node's observation count falls below predefined threshold values (Goldstein et al., 2011).

The random forest (Breiman., 2013) analyses on the data were performed using the random forest package in R 4.3.0 software. This analysis includes three hyperparameters: mtry (0.5(p/3), (p/3), 2(p/3)), ntree (2000, 3000, 4000), and nodesize (5, 10, 15), where p is the total number of SNPs (23,910). Accuracy was used to select the optimal model using the largest value (0.544828). The final values used for the model were mtry = 3985 (0.5(p/3)), ntree = 2000 and nodesize = 5.

Furthermore, the Random Forest (RF) algorithm, as a machine learning method, is particularly well-suited for studies with smaller sample sizes or datasets containing outliers due to its non-parametric nature and robustness. Unlike traditional GWAS methods that rely on stringent statistical thresholds and may miss SNPs with small effects or complex interactions (Botta et al., 2014). The RF leverages Out-of-Bag (OOB) error estimates to provide robust internal validation and reduce overfitting (Breiman, 2001). Additionally, RF effectively handles high-dimensional data and outliers, which are common in genomic datasets, by averaging predictions across multiple decision trees (Enoma et al., 2022). This makes RF an ideal choice for our dataset, where the number of SNPs (23,910) far exceeded the number of samples, and potential outliers from phenotypic or genotypic data could influence the outcome.

Gene ontology

Using the MDA index (value > 1.89), we selected the top 50 SNPs. This approach is consistent with previous studies, such as Bani Saadat et al. (2024; top 10 markers), Li et al. (2018; positive-effect markers), and Schiavo et al. (2020; top 1000 MDA-ranked SNPs). To validate this selection, we identified the genes associated with these 50 SNPs (e.g., *GRK4*, *KDM1B*) using the ENSEMBL database. Subsequently, utilizing the ENSEMBL database and the VEP tool, we identified 82 genes located within 500 kilobase pairs (kbp) of the target SNPs. Finally, for functional annotation and pathway enrichment analysis of these genes, Gene Ontology (GO) analysis was performed using DAVID (https://david.ncifcrf.gov/).

Gene network

The protein-protein interaction network was constructed using the STRING database (https://string-db.org/) by uploading the list of identified genes. The resulting network was downloaded in TSV format and imported into Cytoscape (v3.10.3) for further analysis. To identify the hub genes with high confidence, we employed Maximal Clique Centrality (MCC) analysis, which is particularly effective for detecting the essential nodes based on their clustering patterns in protein networks. The MCC score for a node v was calculated as:

$$MCC(v) = \sum_{C \in S(v)} (|C| - 1)!$$

where S(v) represents the collection of maximal cliques containing v, and (|C|-1)! denotes the product of all positive integers less than |C|. When no edges exist between a node's neighbors, its MCC value equals its degree. Using this approach, we identified high-connectivity hub genes and subsequently refined and visualized the network using Cytoscape's advanced graphical tools (Chin et al., 2014).

Results and discussion

Genome-wide identification of important SNPs

Figure 1 presents the ranking of SNPs associated with bovine leukosis based on the Mean Decrease Accuracy (MDA) index, ordered from most to least significant. In the random forest analysis, the SNP importance was determined using the MDA index, which exhibited negative, zero, or positive values.Positive values indicate that random permutation of the SNP increased the prediction error (MSE) compared to its original state (suggesting SNP importance), while negative values indicated that permutation decreased the MSE, implying potential analytical complications when including such SNPs (Li et al., 2018). The analysis revealed that among all SNPs examined, 4,160 (17.39%) exerted positive effects, 14,943 (62.49%) showed neutral effects, and 4,807 (20.10%) demonstrated negative effects. The highest MDA value was observed for the SNP ARS-BFGL-NGS-117080 located at physical position 115,749,351 bp on chromosome 6, while the lowest value was associated with the SNP BovineHD2600000989 at physical position 4,872,205 bp on chromosome 26.

The Manhattan plot in Figure 2 displays significant leukosis-associated SNPs distributed across the bovine genome. Most top-ranked SNPs were located on chromosomes 1, 9, 14, 16, and 17. We identified two SNPs with MDA >4, four SNPs with MDA between 3-4, and 25 SNPs with MDA values of 2-3. The six top-ranking SNPs were located on chromosomes 6 (MDA=4.9), 17 (MDA=4.8), 9 (MDA=3.8), 20 (MDA=3.6), 17 (MDA=3.2), and 24 (MDA=3.1), respectively.

The top 50 SNPs with positive effects are listed in Table 1 in order of importance. The most important SNP here is ARS.BFGL.NGS.117080 with an MDA of 4.9, which is located on chromosome 6 and within the *GRK4* gene. In this table, the SNPs BTB.00888968, BovineHD2900004537, and ARS.BFGL.NGS.108504, ranked 6, 18, and 30 respectively, are not located on any

gene and are not near any gene. However, the SNP BovineHD2300011387 on chromosome 23 is located on both the *DEK* and *KDM1B* genes simultaneously. Additionally, the SNP BovineHD2300011392 (located on chromosome 23) is located on the *KDM1B* gene. In a study conducted on two Colombian cattle breeds, the *DEK* and *KMD1B* genes were found to be associated with cellular stress response and response to important environmental changes such as high temperature and oxidative stress, respectively (De León et al., 2021).

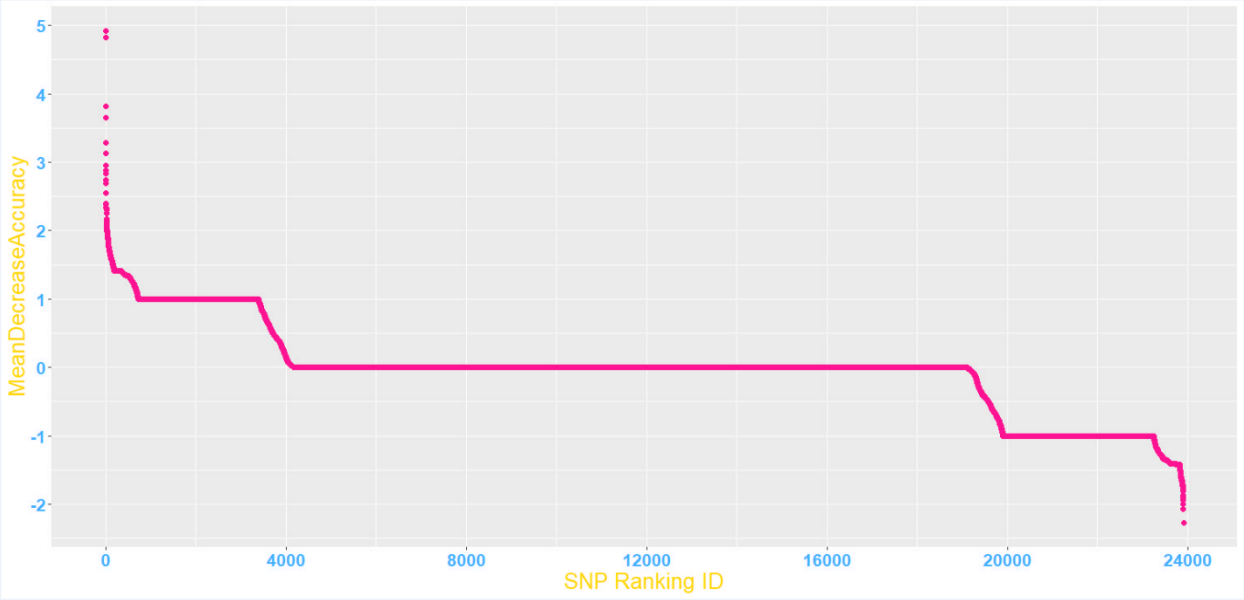

Figure 1. The distribution profiles of ranked SNP variable importance values from RF (MDA) for bovine leukosis

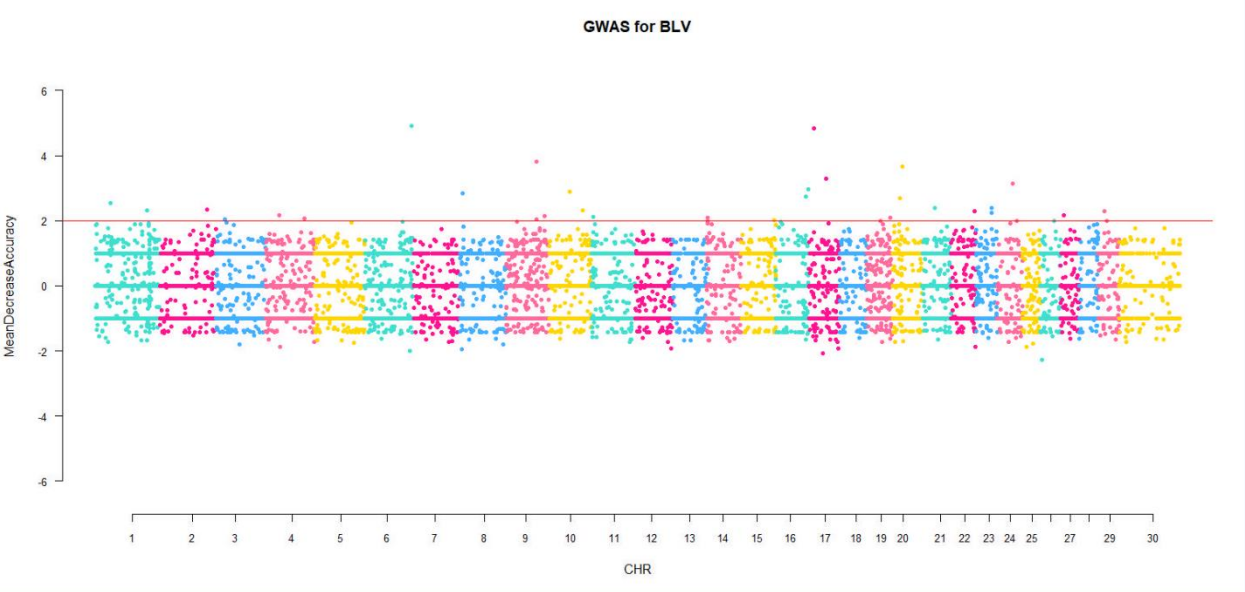

Figure 2. Manhattan plot showing the genome-wide profile of SNP variable importance values for RF (Mean Decrease Accuracy)

Other significant genes including *TDH*, *TRDN*, and *FBLN2* have been associated with various animal cancers (Mentis et al., 2010), variations in bovine intramuscular fat deposition (Sasaki et al., 2006), and carcass traits including backfat thickness (Hong et al., 2019) and milk production (Deng et al., 2024), respectively.

Matenchi et al. (2024) reported that the *CTIF* gene in cattle is associated with production traits and longissimus muscle area, a reliable indicator of growth and productivity. As a component of the CBP80/20-dependent translation initiation complex, *CTIF* binds cotranscriptionally to the cap structure of nascent mRNAs and recognizes premature termination codons (PTCs) during translation, thereby reducing expression of truncated proteins. Mutations in *CTIF* may impair this quality control mechanism, potentially increasing expression of dysfunctional or cytotoxic truncated proteins that could enhance growth of *Streptococcus*

uberis, a major bovine mastitis pathogen (Siebert, 2017). The *CTIF* gene has demonstrated antiviral functions, including suppression of viral replication (Chang et al., 2021), inhibition of HIV-1 Gag production (García-de-Gracia et al., 2021), and modulation of host responses to viral transcription (Salvucci et al., 2022). Furthermore, the *DEK* gene regulates both transcriptional responses to viral infection and maintenance of viral genetic material during infections by human immunodeficiency virus (HIV) and Kaposi's sarcoma-associated herpesvirus (KSHV) (Pease et al., 2015).

Table 1. The list of top 50 ranking SNPs related to BLV from Random Forests (RF)

Rank	Chr[1]	Marker name	Position (bp)	MDA[2]	Downstream	Gene	Upstream
1	6	ARS.BFGL.NGS.117080	115749351	4.91697	*HTT*	*GRK4*	*NOP14*
2	17	BovineHD1700003491	12114382	4.82383		*REELD1*	*LSM6*
3	9	BovineHD0900020518	73829226	3.81943			*PDE7B*
4	20	Hapmap31141.BTA.150972	24031442	3.6516	*GZMA*	*gzmA*	*GZMK*
5	17	BTB.01680332	42584965	3.28263	*PDGFC*	-	-
6	24	BTB.00888968	38518899	3.13495	-	-	-
7	16	BovineHD1600023317	79935995	2.95506	*MGAT4F*		*RABIF*
8	10	BovineHD1000030652	49507121	2.87859			*RORA*
9	8	BTA.07949.rs29027610	7954469	2.83072	*FAM167A*	*TDH*	*MTMR9*
10	16	Hapmap44461.BTA.40052	73108337	2.73872	*SERTAD4*		*SYT14*
11	20	BovineHD2000005646	18839805	2.69576			*PDE4D*
12	1	BovineHD0100010673	37269585	2.54877		*EPHA3*	
13	21	ARS.BFGL.NGS.119025	31960636	2.39442	*ISL2*	*SCAPER*	*RCN2*
14	23	BovineHD2300011387	39393720	2.37971	*RNF144B*	*DEK & KDM1B*	*TPMT*
15	2	BTB.01343453	115092998	2.33358	*IRS1*	*RHBDD1*	*COL4A4*
16	10	BovineHD1000023182	81311788	2.32759	*PLEKHD1*		*CCDC177*
17	1	ARS.BFGL.NGS.113021	126606801	2.30818		*XRN1*	
18	29	BovineHD2900004537	15377343	2.30043			
19	22	BovineHD2200016905	58371789	2.29878		*FBLN2*	*HDAC11*
20	23	BovineHD2300011392	39402448	2.25011	*DEK*	*KDM1B*	*TPMT*
21	4	BovineHD0400009689	34096022	2.17136			
22	27	ARS.BFGL.NGS.117324	7790117	2.16808	*SPCS3*		*VEGFC*
23	9	BovineHD0900026766	94393283	2.1488	*ZDHHC14*		*SNX9*
24	11	ARS.BFGL.NGS.31804	3044180	2.12191			*ZAP70*
25	14	BovineHD1400000684	3507954	2.09081		*TRAPPC9*	*KCNK9*
26	19	ARS.BFGL.NGS.39252	57422012	2.08934	*RPL38*		*SDK2*
27	4	BTB.01700995	94843636	2.07161			*MKLN1*
28	3	BovineHD0300006808	21720162	2.04681	*PDKZ1*	*GPR89A*	*GJA8*
29	9	ARS.BFGL.NGS.25071	74293236	2.03666	*BCLAF1*		*MAP7*
30	15	ARS.BFGL.NGS.108504	81255556	2.02279			
31	26	ARS.BFGL.NGS.1097	8129819	2.00063			*A1CF*
32	26	ARS.BFGL.NGS.83234	35187167	1.9972	*AFAP1L2*	*ABLIM1*	
33	29	BTA.65013.no.rs	20539330	1.9968		*LUZP20*	
34	14	BovineHD1400000779	3831667	1.99056	*KCNK9*		
35	24	BovineHD2400013489	48345676	1.98624	*ZBTB7C*	*CTIF*	*SMAD7*
36	9	Hapmap41922.BTA.63988	27684090	1.97684	*NKAIN2*	*TRDN*	
37	6	BovineHD0600026174	94436688	1.97225			*ANTXR2*
38	16	BovineHD1600003306	12378523	1.96795	*UCHL5*		
39	5	BovineHD0500025349	89333462	1.94386	*SLCO1C1*	*PDE3A*	
40	3	BovineHD0300008013	25425609	1.93565	*GDAP2*		*TENT5C*
41	24	Hapmap41219.BTA.29565	32677985	1.91473	*TTC39C*		*LAMA3*
42	14	BovineHD1400000814	3971521	1.91417	*KCNK9*		
43	17	ARS.BFGL.NGS.16934	47147387	1.91173	*FZD10*		*TMEM132D*
44	1	BTB.01744858	129942985	1.90709	*MRPS22*		
45	1	ARS.BFGL.NGS.115228	109706655	1.902	*SHOX2*		*VEPH1*
46	1	Hapmap58062.rs29012621	96059058	1.90071	*PLD1*		*TNIK*
47	14	Hapmap38378.BTA.114219	12603761	1.89927			*MYC*
48	28	BovineHD2800012053	42652683	1.89732	*PTPN20*	*FRMPD2*	*MAPK8*
49	1	Hapmap42913.BTA.33619	4052161	1.89457			*TIAM1*
50	20	BovineHD2000005045	16626569	1.89089			*IPO11*

[1]Chromosome
[2]Mean Decrease Accuracy

Gene ontology (GO) enrichment analysis

Table 2 presents the GO enrichment analysis results of the top 50 SNPs from the ENSEMBL website. Examination of the biological functions of genes associated with the top 50 SNPs (most frequently including *KDM1B*, *MYC*, *MAPK8*, *PDE3A*, *GZMA*, *EPHA3*, *BCLAF1*, and *GRK4*) revealed that these genes are primarily involved in: positive regulation of cellular process, regulation of cellular metabolic process, positive regulation of cellular metabolic process, positive regulation of macromolecule metabolic process, intracellular signal transduction, and regulation of developmental process.

Granzymes play a crucial role in inflammatory responses by mediating perforin-dependent death of virus-infected or cancerous cells targeted by cytotoxic T lymphocytes. *GZMA* is a serine protease produced by cytotoxic T and natural killer (NK) cells that induce apoptosis in target cells (Rosse et al., 2017). *GZMA* is the most highly expressed gene in endometrial CD14+ cells and functions in cellular apoptosis (Oliveira et al., 2011). This gene also plays roles in bovine tuberculosis (Bhat et al., 2023) and buffalo milk somatic cells (Ahlawat et al., 2021).

Table 2. Gene enrichment analysis for top 50 SNPs with positive variable importance values from RF

Term	Count	P-Value	Genes
GO:0048522~positive regulation of cellular process	26	2.17E-04	*KDM1B, SHOX2, HTT, RORA, FZD10, PLD1, FBLN2, BCLAF1, MAPK8, MYC, PDGFC, SNX9, ZBTB7C, AFAP1L2, GZMA, VEGFC, TENT5C, A1CF, SMAD7, SLCO1C1, GZMK, TIAM1, ISL2, PDE3A, TNIK, EPHA3*
GO:0031325~positive regulation of cellular metabolic process	17	3.29E-04	*AFAP1L2, KDM1B, VEGFC, HTT, RORA, PLD1, TENT5C, A1CF, SLCO1C1, BCLAF1, MAPK8, ISL2, MYC, PDGFC, SNX9, TNIK, ZBTB7C*
GO:0010604~positive regulation of macromolecule metabolic process	16	0.001826	*AFAP1L2, KDM1B, VEGFC, RORA, PLD1, TENT5C, A1CF, RNF144B, BCLAF1, MAPK8, ISL2, MYC, PDGFC, SNX9, TNIK, ZBTB7C*
GO:0031323~regulation of cellular metabolic process	22	0.056954	*AFAP1L2, KDM1B, SHOX2, VEGFC, MTMR9, HTT, RORA, PLD1, TENT5C, A1CF, SMAD7, SLCO1C1, CTIF, BCLAF1, MAPK8, ISL2, XRN1, MYC, PDGFC, SNX9, TNIK, ZBTB7C*
GO:0006468~protein phosphorylation	5	0.081771	*ZAP70, MAPK8, GRK4, TNIK, EPHA3*
GO:0006796~phosphate-containing compound metabolic process	10	0.044492	*ZAP70, MAPK8, GRK4, PDE4D, PTPN20, MTMR9, RORA, TNIK, PLD1, EPHA3*
GO:0006793~phosphorus metabolic process	10	0.048545	*ZAP70, MAPK8, GRK4, PDE4D, PTPN20, MTMR9, RORA, TNIK, PLD1, EPHA3*
GO:0043067~regulation of programmed cell death	8	0.030163	*GZMK, RNF144B, BCLAF1, MAPK8, MYC, RHBDD1, GZMA, HTT*
GO:0007169~cell surface receptor protein tyrosine kinase signaling pathway	6	0.003213	*ZAP70, IRS1, COL4A4, PDGFC, VEGFC, EPHA3*
GO:0043068~positive regulation of programmed cell death	6	0.003859	*GZMK, BCLAF1, MAPK8, MYC, GZMA, HTT*
GO:0019933~cAMP-mediated signaling	3	0.006016	*PDE4D, PDE3A, PDE7B*
GO:0042981~regulation of apoptotic process	8	0.026088	*GZMK, RNF144B, BCLAF1, MAPK8, MYC, RHBDD1, GZMA, HTT*
GO:1902531~regulation of intracellular signal transduction	11	0.005762	*BCLAF1, MAPK8, MYC, PDE4D, PDGFC, PDE3A, HTT, RORA, TNIK, FZD10, SMAD7*
GO:0035556~intracellular signal transduction	11	0.005858	*ZAP70, TIAM1, RABIF, MAPK8, MYC, PDE4D, PDE3A, TNIK, PDE7B, PLD1, SMAD7*
GO:0050793~regulation of developmental process	10	0.054087	*TIAM1, MYC, LAMA3, SHOX2, PDE3A, VEGFC, RORA, TNIK, SMAD7, ZBTB7C*

The ephrin-Eph gene family has well-established physiological functions in regulating mammalian reproductive performance, particularly in bovine ovarian granulosa cells (Yousuf et al., 2023). In cattle, *EPHA3* is associated with fertility, angularity (Ooi et al., 2024), nervous system function, platelet reactivity, parasite resistance and histoblood group antigens (Yang et al., 2017), while in sheep it relates to fertility traits (Yousuf et al., 2023).

The PDE type 3 plays an essential role in the meiotic resumption of bovine oocytes (Mayes and Sirard, 2002; Schwarz et al., 2014). The *PDE3A* gene is associated with semen production traits (Liu et al., 2017).

The *BCLAF1*-associated transcription factor was initially identified as a regulator of apoptosis and transcription, and has since been implicated in a wide array of biological processes including T-cell activation, lung development, muscle cell proliferation and differentiation, autophagy, and viral infections. Notably, *BCLAF1* can function as either an oncogene or tumor suppressor in carcinogenesis, depending on the cancer type (Yu et al., 2022).

Gene networks and hub genes

The gene network of identified genes was constructed using Cytoscape software version 3.10.3 (Figure 3), followed by identification of hub genes including *MYC*, *RABIF*, *IRS1*, *TRAPPC9*, *MAPK8*, *HTT*, *SNX9*, *BCLAF 1*, *XRN1*, and *LSM6* (Figure 4).

The *MYC* gene performs several functions in bovine mammary gland (Bionaz and Loor, 2007) and is associated with muscle growth and differentiation in cattle (Sheet et al., 2024). As a core component of FSH signaling, *MYC* integrates FSH signaling networks and may help investigate genome-wide transcriptional changes associated with oocyte competence acquisition (Cantanhêde et al., 2022). Elevated c-myc transcripts in BLV-induced tumors result from a series of changes induced by BLV infection that persist throughout the neoplastic process (Neoplastic diseases are conditions that cause tumor growth), even after viral expression ceases (Gupta et al., 1986).

The SNPs in *TRAPPC9* may serve as useful genetic markers for selection toward mastitis resistance (Wang et al., 2024) and improvement in milk protein (Khan et al., 2022) and fat content (Freitas et al., 2020) in dairy cattle.

MicroRNAs (miRNAs) are short non-coding RNAs that can regulate target gene expression at the post-transcriptional level. The bta-miR-145 has been reported to show differential expression across different lactation stages in bovine mammary glands. Insulin receptor substrate 1 (*IRS1*) was predicted as a potential target of miR-145. The bta-miR-145 is involved in the proliferation of bovine mammary epithelial cells by targeting *IRS1*, which is associated with the *MAPK* signaling pathway (Solodneva et al., 2022). The *IRS1* gene also plays a role in regulating milk fat metabolism in dairy cows (Jiao et al., 2020).

Downregulation of lncRNA IALNCR in host cells during Bovine Viral Diarrhea Virus (BVDV) infection suppresses *MAPK8*/*JNK1* expression at both mRNA and protein levels. This suppression indirectly activates caspase-3, triggering autonomous cell apoptosis to inhibit BVDV replication (Gao et al., 2022). The *XRN1* may serve as a critical mechanism through which various Flaviviridae family viruses (including BVDV) exert pathogenicity by disrupting cellular gene expression regulation (Moon, 2014). Following infection by Flaviviridae viruses such as Zika virus and West Nile virus, eukaryotic hosts employ the evolutionarily conserved endoribonuclease *Xrn1* to degrade viral genomic RNA (Dilweg et al., 2021).

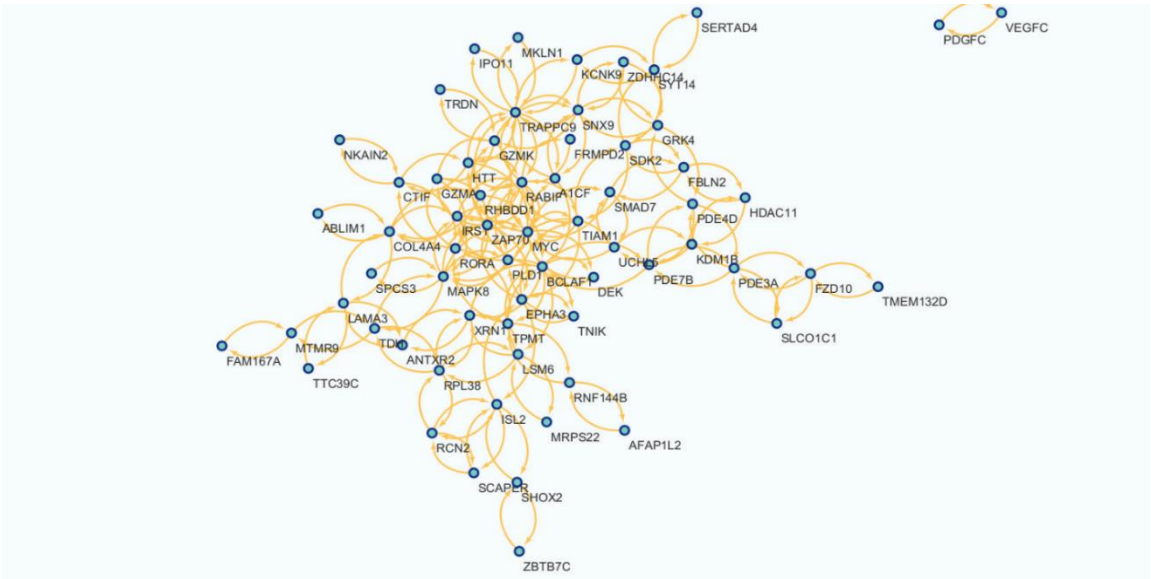

Figure 3. Gene network displaying the connections between 82 genes

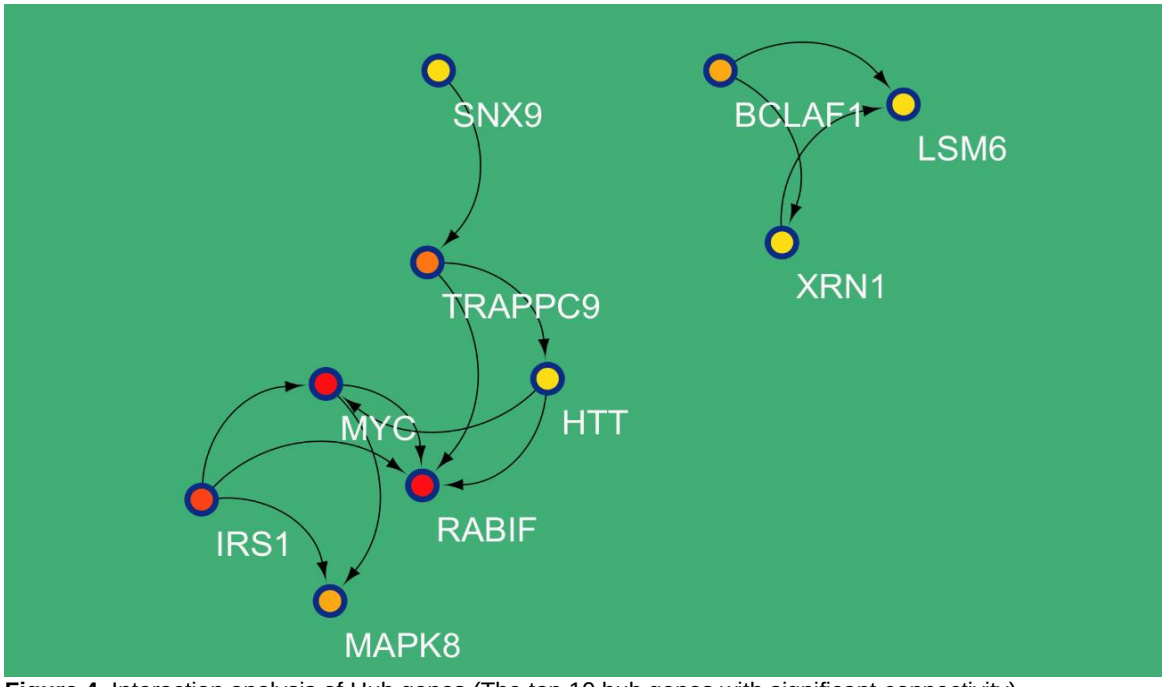

Figure 4. Interaction analysis of Hub genes (The top 10 hub genes with significant connectivity)

Identifying the susceptibility loci related to disease is one of the fundamental and important challenges in modeling complex diseases (Silva et al., 2022). Genome-wide association studies (GWAS) involve scanning genomes from many individuals to find genetic markers that correlate with observable traits or diseases. This process starts with association tests that evaluate one single-nucleotide polymorphism (SNP) at a time across the genome to identify variants with statistically significant associations with the target phenotype (Alireza et al., 2024). While a single SNP may not directly cause a particular disease, specific sequences or combinations of SNPs have predictive power for detecting the disease of interest (Wakayu et al., 2021). Machine learning (ML) methods are increasingly important in genome-wide association studies for identifying key genetic variants or SNPs that statistical methods might overlook (Alireza et al., 2024). Among these, the most successful method is random forest (Nguyen et al., 2015).

In our study, 23,910 SNPs were analyzed using the random forest. After identifying significant SNPs, we mapped their associated genes and found two genes, *GRK4* and *SCAPER*, which were consistent with a prior GWAS analysis conducted on the same dataset (Arjmand Kermani et al., 2024). In cattle, *GRK4* has been linked to the regulation of two body weight-related hormones (parathyroid hormone and adrenomedullin) (Jiang et al., 2022), while *SCAPER* is associated with spermatogenesis and fertility (Ghoreishifar et al., 2023). The discrepancy in gene identification between our study and the previous GWAS likely stems from differences in analytical models and assumptions: GWAS relies on statistical thresholds, whereas random forest is non-parametric and does not require predefined statistical assumptions. Based on these findings, we recommend integrating classical methods (e.g., GWAS) with modern machine learning techniques to optimize marker discovery in genomic analyses.

Conclusions

This study successfully identified key SNPs associated with bovine leukosis using the random forest method, with some SNPs (e.g., *GRK4* and *SCAPER*) overlapping with findings from previous GWAS study (Arjmand Kermani et al., 2024). Functional analyses revealed that critical genes such as *MYC*, *MAPK8*, and *GZMA* are involved in pathways related to apoptosis, viral response, and cellular metabolic regulation, potentially elucidating mechanisms underlying leukosis pathogenesis. The constructed gene networks further highlighted hub genes (e.g., *MYC* and *BCLAF1*) as key players in molecular interactions linked to the disease. While machine learning methods like random forest proved efficient for SNP screening, their integration with classical statistical approaches (GWAS) could enhance the precision of genetic marker identification. These findings not only provide novel insights into the genetic factors influencing bovine leukosis but also offer practical strategies for livestock breeding programs aimed at improving disease resistance. Future studies should focus on the functional validation of these SNPs and the investigation of epistatic interactions to advance our understanding of the disease's pathogenesis.

Acknowledgements

The authors gratefully acknowledge the FKA Company (Isfahan) for supplying the healthy and diseased cattle samples used in this study.

References

Ahlawat, S., Arora, R., Sharma, U., Sharma, A., Girdhar, Y., Sharma, R., Kumar, A., Vijh, R.K., 2021. Comparative gene expression profiling of milk somatic cells of Sahiwal cattle and Murrah buffaloes. *Gene* 764, 145101.

Alireza, Z., Maleeha, M., Kaikkonen, M., Fortino, V., 2024. Enhancing prediction accuracy of coronary artery disease through machine learning-driven genomic variant selection. *Journal of Translational Medicine* 22, 356.

Arjmand Kermani, F., Moradi Shahr Babak, H., Moradi Shahr Babak, M., Mohammadi, H., Jvan Nikkhah, M., Doosti, Y., 2024. Genome-wide association study to identify the loci related to resistance in Leukosis disease in Iranian Holstein cattle. *Journal of Animal Production* 26, 219-232. (In Persian)

Bhat, S.A., Elnaggar, M., Hall, T.J., McHugo, G.P., Reid, C., MacHugh, D.E., Meade, K.G., 2023. Preferential differential gene expression within the WC1.1+ γδ T cell compartment in cattle naturally infected with Mycobacterium bovis. *Frontiers in Immunology* 14, 1265038.

Bionaz, M., Loor, J.J., 2007. Identification of reference genes for quantitative real-time PCR in the bovine mammary gland during the lactation cycle. *Physiological Genomics* 29, 312-319.

Bongers, R., 2023. A genetic perspective on enzootic bovine leukosis resistance in Canadian Holstein cattle. Master of Science Thesis, University of Guelph, Guelph, Ontario, Canada.

Botta, V., Louppe, G., Geurts, P., Wehenkel, L., 2014. Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS One* 9, e93379.

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

Breiman, L., 2013. Breiman and Cutler's Random Forests for Classification and Regression. Package 'RandomForest'. Institute for Statistics and Mathematics, University of Economics and Business, Vienna.

Cantanhêde, L.F., Moura, M.T., Oliveira-Silva, R.L., Nascimento, P.S., Ferreira-Silva, J.C., Benko-Iseppon

A.M., Oliveira, M.A.L., 2022. MYC integrates FSH signalling networks in cumulus cells during bovine oocyte maturation. *Acta Veterinaria Hungarica* 70, 1-9.

Chang, J., Hwang, H.J., Kim, B., Choi, Y.G., Park, J., Park, Y., Lee, B.S., Park, H., Yoon, M.J., Woo, J.S., Kim, C., Park, M.S., Lee, J.B., Kim, Y.K., 2021. TRIM28 functions as a negative regulator of aggresome formation. *Autophagy* 17, 4231-4248.

Chin, C.H., Chen, S.H., Wu, H.H., Ho, C.W., Ko, M.T., Lin, C.Y., 2014. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology 8* (Suppl 4) S11.

De León, C., Martínez, R., Rocha, J.F., Darghan, A.E., 2021. Selection of genomic regions and genes associated with adaptation and fertility traits in two Colombian creole cattle breeds. *Genetics and Molecular Research* 20, gmr18882.

Deng, T.X., Ma, X.Y., Duan, A., Lu, X.R., Abdel-Shafy, H., 2024. Genome-wide copy number variant analysis reveals candidate genes associated with milk production traits in water buffalo (Bubalus bubalis). *Journal of Dairy Science* 107, 7022-7037.

Dilweg, I.W., Savina, A., Köthe, S., Gultyaev, A.P., Bredenbeek, P.J., Olsthoorn, R.C.L., 2021. All genera of Flaviviridae host a conserved Xrn1-resistant RNA motif. *RNA Biology* 18, 2321-2329.

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.

Freitas, P.H.F., Oliveira, H.R., Silva, F.F., Fleming, A., Schenkel, F.S., Miglior, F., Brito, L.F., 2020. Short communication: Time-dependent genetic parameters and single-step genome-wide association analyses for predicted milk fatty acid composition in Ayrshire and Jersey dairy cattle. *Journal of Dairy Science* 103, 5263-5269.

Gao, X., Sun, X., Yao, X., Wang, Y., Li, Y., Jiang, X., Han, Y., Zhong, L., Wang, L., Song, H., Xu, Y., 2022. Downregulation of the long noncoding RNA IALNCR targeting MAPK8/JNK1 promotes apoptosis and antagonizes bovine viral diarrhea virus replication in host cells. *Journal of Virology* 96, e0111322.

García-de-Gracia, F., Gaete-Argel, A., Riquelme-Barrios, S., Pereira-Montecinos, C., Rojas-Araya, B., Aguilera, P., Oyarzún-Arrau, A., Rojas-Fuentes, C., Acevedo, M.L., Chnaiderman, J., Valiente-Echeverría, F., Toro-Ascuy, D., Soto-Rifo, R., 2021. CBP80/20-dependent translation initiation factor (CTIF) inhibits HIV-1 Gag synthesis by targeting the function of the viral protein Rev. *RNA Biology* 18, 745-758.

Ghoreishifar, M., Vahedi, S.M., Salek Ardestani, S., Khansefid, M., Pryce, J.E., 2023. Genome-wide assessment and mapping of inbreeding depression identifies candidate genes associated with semen traits in Holstein bulls. *BMC Genomics* 24, 230.

Goldstein, B.A., Polley, E.C., Briggs, F.B., 2011. Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology* 10, 32.

Gupta, P., Kashmiri, S.V., Erisman, M.D., Rothberg, P.G., Astrin, S.M., Ferrer, J.F., 1986. Enhanced expression of the c-myc gene in bovine leukemia virus-induced bovine tumors. *Cancer Research* 46, 6295-6298.

Hong, M., Choi, S., Singh, N.K., Kim, H., Yang, S., Kwak, K., Kim, J., Lee, S., 2019. Genome-wide association analysis to identify QTL for carcass traits in Hanwoo (Korean native cattle). *Indian Journal of Animal Sciences* 89, 57-62.

Jiang, H., Chai, Z.X., Cao, H.W., Zhang, C.F., Zhu, Y., Zhang, Q., Xin, J.W., 2022. Genome-wide identification of SNPs associated with body weight in yak. *BMC Genomics* 23, 833.

Jiao, P., Yuan, Y., Zhang, M., Sun, Y., Wei, C., Xie, X., Zhang, Y., Wang, S., Chen, Z., Wang, X., 2020. PRL/microRNA-183/IRS1 pathway regulates milk fat metabolism in cow mammary epithelial cells. *Genes* 11, 196.

Khan, M.Z., Dari, G., Khan, A., Yu, Y., 2022. Genetic polymorphisms of TRAPPC9 and CD4 genes and their association with milk production and mastitis resistance phenotypic traits in Chinese Holstein. *Frontiers in Veterinary Science* 9, 1008497.

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.

Liu, S., Yin, H., Li, C., Qin, C., Cai, W., Cao, M., Zhang, S., 2017. Genetic effects of PDGFRB and MARCH1 identified in GWAS revealing strong associations with semen production traits in Chinese Holstein bulls. *BMC Genetics* 18, 63.

Lv, G., Wang, J., Lian, S., Wang, H., Wu, R., 2024. The global epidemiology of bovine leukemia virus: Current trends and future implications. *Animals* 14, 297.

Matenchi, Y.P., Hegarty, M., Baştanlar, E.K., 2024. Genome wide association analysis revealed novel candidate genes for body measurement traits in indigenous Gudali and crossbred Simgud in Cameroon. *Research Square* PREPRINT (Version 1).

Mayes, M.A., Sirard, M.A., 2002. Effect of type 3 and type 4 phosphodiesterase inhibitors on the maintenance of bovine oocytes in meiotic arrest. *Biology of Reproduction* 66, 180-184.

Meng, Y.A., Yu, Y., Cupples, L.A., Farrer, L.A., Lunetta, K.L., 2009. Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics* 10, 78.

Mentis, A.F., Kararizou, E., 2010. Metabolism and cancer: an up-to-date review of a mutual connection.

Asian Pacific Journal of Cancer Prevention 11, 1437-1444.

Moon, S.L., 2014. Inhibition of the host 5'-3' RNA decay pathway is a novel mechanism by which flaviviruses influence cellular gene expression. Ph.D. Thesis, Colorado State University, Fort Collins, Colorado, USA.

Nguyen, T.T., Huang, J., Wu, Q., Nguyen, T., Li, M., 2015. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics* 16 Suppl 2, S5.

Oliveira, L.J., McClellan, S., Hansen, P.J., 2010. Differentiation of the endometrial macrophage during pregnancy in the cow. *PLoS One* 5, e13213.

Ooi, E., Xiang, R., Chamberlain, A.J., Goddard, M.E., 2024. Archetypal clustering reveals physiological mechanisms linking milk yield and fertility in dairy cattle. *Journal of Dairy Science* 107, 4726-4742.

Pease, N.A., Wise-Draper, T., Privette Vinnedge, L., 2015. Dissecting the Potential Interplay of DEK Functions in Inflammation and Cancer. *Journal of Oncology* 2015, 106517.

Rosse, I.C., Assis, J.G., Oliveira, F.S., Leite, L.R., Araujo, F., Zerlotini, A., Volpini, A., Dominitini, A.J., Lopes, B.C., Arbex, W.A., Machado, M.A., Peixoto, M.G., Verneque, R.S., Martins, M.F., Coimbra, R.S., Silva, M.V., Oliveira, G., Carvalho, M.R., 2017. Whole genome sequencing of Guzerá cattle reveals genetic variants in candidate genes for production, disease resistance, and heat tolerance. *Mammalian Genome* 28, 66-80.

Saadat, H.B., Torshizi, R.V., Manafiazar, G., Masoudi, A.A., Ehsani, A. and Shahinfar, S., 2024. Comparing machine learning algorithms and linear model for detecting significant SNPs for genomic evaluation of growth traits in F 2 chickens. *Journal of Agricultural Science & Technology* 26(6), 1261-1274.

Salvucci, M., Crawford, N., Stott, K., Bullman, S., Longley, D.B., Prehn, J.H.M., 2022. Patients with mesenchymal tumours and high Fusobacteriales prevalence have worse prognosis in colorectal cancer (CRC). *Gut* 71, 1600-1612.

Sasaki, Y., Nagai, K., Nagata, Y., Doronbekov, K., Nishimura, S., Yoshioka, S., Fujita, T., Shiga, K., Miyake, T., Taniguchi, Y., Yamada, T., 2006. Exploration of genes showing intramuscular fat deposition-associated expression changes in musculus longissimus muscle. *Animal Genetics* 37, 40-46.

Schiavo, G., Bertolini, F., Bovo, S., Galimberti, G., Muñoz, M., Bozzi, R., Čandek-Potokar, M., Óvilo, C., Fontanesi, L., 2024. Identification of population-informative markers from high-density genotyping data through combined feature selection and machine learning algorithms: Application to European autochthonous and cosmopolitan pig breeds. *Animal Genetics* 55, 193-205.

Schiavo, G., Bertolini, F., Galimberti, G., Bovo, S., Dall'Olio, S., Nanni Costa, L., Gallo, M., Fontanesi, L., 2020. A machine learning approach for the identification of population-informative markers from high-throughput genotyping data: application to several pig breeds. *Animal* 14, 223-232.

Schwarz, K.R., Pires, P.R., Mesquita, L.G., Chiaratti, M.R., Leal, C.L., 2014. Effect of nitric oxide on the cyclic guanosine monophosphate (cGMP) pathway during meiosis resumption in bovine oocytes. *Theriogenology* 81, 556-564.

Sheet, S., Jang, S.S., Kim, J.H., Park, W., Kim, D., 2024. A transcriptomic analysis of skeletal muscle tissues reveals promising candidate genes and pathways accountable for different daily weight gain in Hanwoo cattle. *Scientific Reports* 14, 315.

Siebert, L.J., 2017. Identifying genome associations with unique mastitis phenotypes in response to intramammary Streptococcus uberis challenge. Ph.D. Thesis, University of Tennessee, Knoxville, Tennessee, USA.

Silva, P.P., Gaudillo, J.D., Vilela, J.A., Roxas-Villanueva, R.M.L., Tiangco, B.J., Domingo, M.R., Albia, J.R., 2022. A machine learning-based SNP-set analysis approach for identifying disease-associated susceptibility loci. *Scientific Reports* 12, 15817.

Solodneva, E.V., Kuznetsov, S.B., Velieva, A.E., Stolpovsky, Yu.A., 2022. Molecular-genetic bases of mammary gland development using the example of cattle and other animal species: I. Embryonic and pubertal developmental stage. *Russian Journal of Genetics* 58, 899-914.

Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., Posthuma, D., 2021. Genome-wide association studies. *Nature Reviews Methods Primers* 1, 1-21.

Wakayu, E.G., 2021. Machine learning analysis of single nucleotide polymorphism (SNP) data to predict bone mineral density in African American women. Master's Thesis, University of Nevada, Las Vegas, Nevada, USA.

Wang, D., Ma, S., Yan, M., Dong, M., Zhang, M., Zhang, T., Zhang, T., Zhang, X., Xu, L., Huang, X., 2024. DNA methylation patterns in the peripheral blood of Xinjiang brown cattle with variable somatic cell counts. *Frontiers in Genetics* 15, 1405478.

Yang, L., Xu, L., Zhu, B., Niu, H., Zhang, W., Miao, J., Shi, X., Zhang, M., Chen, Y., Zhang, L., Gao, X., Gao, H., Li, L., Liu, G.E., Li, J., 2017. Genome-wide analysis reveals differential selection involved with copy number variation in diverse Chinese Cattle. *Scientific Reports* 7, 14299.

Yousuf, S., Malik, W.A., Feng, H., Liu, T., Xie, L., Miao, X., 2023. Genome wide identification and characterization of fertility associated novel CircRNAs as ceRNA reveal their regulatory roles in sheep fecundity. *Journal of Ovarian Research* 16, 115.

Yu, Z., Zhu, J., Wang, H., Li, H., Jin, X., 2022. Function of BCLAF1 in human disease. *Oncology Letters* 23, 58.