Letters in Animal Biology

Journal homepage: www.liabjournal.com

Integrated transcriptomic and machine learning analysis of the candidate biomarker potential of ACSL6 in bovine mastitis

Lutfi Bayyurt

Department of Animal Science, Faculty of Agriculture, Tokat Gaziosmanpasa University, Tokat 60240, Turkey

Article info

Received: 22 April 2026
Received in revised form: 22 May 2026
Accepted: 23 May 2026
Published online: 01 June 2026

Keywords

Bovine mastitis
ACSL6
Differential gene expression
Machine learning
Random Forest
XGBoost

* *Corresponding author:*
Lutfi Bayyurt
Email: lutfi.bayyurt@gop.edu.tr

Reviewed by:

Consent to publish the name of the reviewers could not be obtained

Abstract

Bovine mastitis is a complex inflammatory disease that has a substantial negative influence on milk productivity and costs the dairy industry greatly. Identifying strong molecular biomarkers for early diagnosis is still a key area of research. In this study, integrated analyses of transcriptomic data and machine-learning approaches were used to identify candidate bovine mastitis-associated genes. The GSE15020 and GSE24217 data sets were used as discovery cohorts, while the third dataset GSE50685 served as an independent validation set. Differential gene expression analysis following quality control, normalization and batch effect adjustment using the ComBat method identified 1143 significant genes in the discovery cohort. Candidate biomarkers were subsequently ranked and selected using Random Forest and XGBoost algorithms. All candidate genes were evaluated in an independent dataset, where ACSL6 showed the highest discriminative power ($p < 0.001$) relative to all other candidates and PTX3 also showed potential as a good classifier. The robustness of the model was evaluated using ROC curve, confidence intervals, and permutation tests. Functional enrichment analysis showed that the DEGs were mainly involved in immune response, inflammatory pathways and defense response. Despite the limited sample size of the validation cohort, findings of present study support ACSL6 as a potential biomarker for bovine mastitis. However, larger independent cohorts and experimental validation are needed to support its diagnostic relevance and biological significance. This analysis collectively demonstrates that the combination of transcriptomic profiling and machine learning is a valuable means to identify candidate biomarkers for bovine mastitis.

1. Introduction

Bovine mastitis is one of the most common and economically important diseases in dairy cattle. It causes significant economic damage through decreased milk production, decreased milk quality, increased treatment costs and early removal from the herd (Seegers et al. 2003; Halasa et al. 2007). Mastitis is a multifactorial disease as a result of a complex interaction between different management, environmental, and susceptible host-related factors which predispose cattle to infection (Bradley 2002; Rainard and Riollet 2006). This complexity makes precise diagnosis and effective control challenging, emphasizing the need for sensitive, molecular-based methods of diagnostics to facilitate early detection and ultimately better disease management. Mastitis occurs in clinical and subclinical forms, and subclinical mastitis constitutes a major challenge for early diagnosis due to the absence of obvious clinical signs. Parameters such as somatic cell count are widely used indicators of inflammation but do not fully reflect the molecular level dynamics of the disease progression (Halasa et al. 2007; Schukken et al. 2003). Therefore, in recent years, studies aimed at identifying molecular biomarkers for early diagnosis of mastitis and better understanding of its pathogenesis have increased considerably (Lutzow et al. 2008; Hasin et al. 2017; Lim et al. 2025).

With the development of high-throughput omics technologies, gene expression profiling has become an important tool in the investigation of disease mechanisms. In particular, microarray and RNA-seq based studies have shown that genes involved in immune response, inflammation, cellular stress, and metabolic processes are significantly altered during mastitis (Lutzow et al. 2008; Moyes et al. 2009; Kosciuczuk et al. 2017; Loor et al. 2011). For example, cytokines, chemokines and acute phase proteins related to the innate immune response have been reported to be significantly regulated during mastitis (Naserkheil et al. 2022). However, a significant portion of these studies has been conducted using single datasets with limited sample sizes, and the generalizability of the results remains limited. In recent years, approaches based on the integration of multiple datasets have enabled the generation of more robust and reliable results in biomarker discovery. Meta-analysis and data integration methods allow the combined evaluation of data obtained from different experimental conditions and increase the statistical power (Ramasamy et al. 2008). However, one of the most important challenges of such approaches is batch effects arising from technical variations. Batch effects can lead to systematic differences between datasets, masking biological signals, producing false-positive results, and leading to incorrect biological interpretations (Leek et al. 2010; Lazar et al. 2013). The ComBat method, which is based on empirical Bayes approaches, is widely used for the removal of batch effects. Johnson et al. (2007) stated that this method effectively reduces technical variation in microarray data and increases compatibility between datasets. This method is critically importance, especially in the integration of data obtained from different sources and ensures the preservation of biological signals.

Although differential gene expression analysis is a fundamental approach for identifying disease-associated genes, it may not be sufficient on its own due to the high dimensional structure of the data.

These analyses generally produce a large number of candidate genes, but it is often difficult to determined which of these genes can be used as biomarkers. At this point, machine learning methods provide powerful tools for identifying important features and modeling of high-dimensional data (Guyon and Elisseeff 2003). In particular, the Random Forest algorithm is widely used in biological data to identify important features through variable importance measures (Breiman 2001). Similarly, the XGBoost algorithm produces high accuracy predictions with the gradient boosting principle and provides successful results in biomedical data analysis (Chen and Guestrin 2016). The combined use of these two methods is thought to provide important advantages in terms of both accuracy and reliability in biomarker selection. Nevertheless, one of the, if not the, most crucial steps in biomarker discovery is validation against an independent dataset. There is a tendency so far for many investigations to be limited only to the original discovery dataset, resulting in both high risk of overfitting (Ioannidis 2005; Varma and Simon 2006; Vabalas et al. 2019; Libbrecht and Noble 2015) and a lack of generalizability of the model. Because biomarkers may not truly reflect underlying biological processes, their independent validation is critical (Ioannidis 2005). Additionally, confidence interval estimation or permutation testing should be applied to properly evaluate model performance, especially in studies with small sample size.

In this study, an integrative analytical framework combining differential gene expression analysis with machine learning algorithms was applied to bovine mastitis related microarray datasets. Genes nominated by this approach were further validated using an external dataset. Receiver operating characteristic (ROC) analysis was conducted with confidence interval estimation to rigorously assess predivtive performance. Permutation testing was performed to determine the probability that the results obtained were due to random chance, thereby assessing their statistical significance. In addition, functional enrichment analyses of the identified genes were conducted to further investigate their potential biological roles in relevant cellular and inflammatory processes. The overall study design and analytical workflow are summarized in Fig. 1. The primary objective of this study was to identify biologically relevant and potentially reproducible biomarker candidates that can be used in the diagnosis of mastitis and to contribute to the existing methodological gap in this field.

2. Materials and Methods

2.1 Data acquisition and study design

The gene expression data used in this study were obtained from the Gene Expression Omnibus (GEO) database. The GSE15020 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15020) and GSE24217 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24217) datasets used in the discovery stage, and the GSE50685 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50685) dataset analyzed for independent validation were evaluated within the scope of this study. All datasets were generated on the same microarray platform (GPL2112), which contributed to reducing platform-related technical variation during data integration. The discovery dataset included a total of 31 samples (14 control, 17 mastitis), while the validation dataset includes 10 samples (5 control, 5 mastitis). The overall study design and analytical workflow are presented in Fig. 1.

2.2 Data preprocessing and normalization

All analyses were performed in the R statistical environment. GEO datasets were downloaded using the GEOquery package and converted into gene expression matrices. Since all datasets were generated using the same GPL2112 microarray platform, platform-related variability was minimized; however, dataset-specific technical variation was still considered during preprocessing. Expression values were initially inspected to determine whether log transformation was required. When necessary, log2 transformation was applied to stabilize variance and reduce the influence of highly expressed probes. To ensure comparability between samples, GEO-provided normalized expression matrices generated using standard platform preprocessing procedures were used (Bolstad et al. 2003), and additional between-sample distribution checks were performed using boxplots and density plots. Samples showing extreme distributional deviation were visually inspected before downstream analysis. Only genes common to all discovery and validation datasets were retained for the integrative analysis. When multiple probes corresponded to the same gene symbol, probe-level expression values were summarized at the gene level. Probe summarization was performed by retaining the probe with the highest median absolute deviation across samples, rather than simple averaging, because this approach preserves the most informative probe and reduces the potential dilution of gene-level signal (Irizarry et al. 2003; Hackstadt and Hess 2009). Genes with low expression variability were removed before differential expression and machine learning analyses. Specifically, genes within the lowest 20% of variance across discovery samples were excluded to reduce low-information noise and improve downstream statistical and machine learning model stability.

2.3 Evaluation and correction of batch effects

To reduce non-biological technical variation arising from the integration of independent microarray datasets, batch effect correction was performed using the ComBat function implemented in the sva package based on an empirical Bayes framework (Johnson et al. 2007). Batch effects represent systematic technical differences between datasets that may obscure underlying biological signals and influence downstream statistical analyses (Parker et al. 2014). Expression matrices from the discovery datasets were merged according to common gene symbols prior to correction. The ComBat model was adjusted with batch information corresponding to the origin of the dataset, whereas biological group information (mastitis versus control), which was directly relevant to adjustment, was retained. The goal of this approach was to reduce dataset-specific technical variation while retaining biologically relevant differences in expression related to disease status. Principal component analysis (PCA) was used to evaluate the batch effects and the role of batching correction. PCA was used to explore clustering both before and after batch correction (Ringnér 2008). The samples clustered almost exclusively by their dataset origin before the ComBat adjustment, indicative of significant technical variation. Together, these results suggest that batch correction greatly reduced technical variability and at the same time preserved biologically relevant expression patterns. Since PCA provides only a qualitative representation of the global variance landscape, the evaluation of batch correction success should be performed cautiously and treated as indicative but not definitive evidence for complete removal of batch effects. Such a viewpoint correlates well with previous recommendations for careful evaluation of batch correction methods when integrating data from different transcriptomic studies (Leek et al. 2010; Lazar et al. 2013).

2.4 Differential gene expression analysis

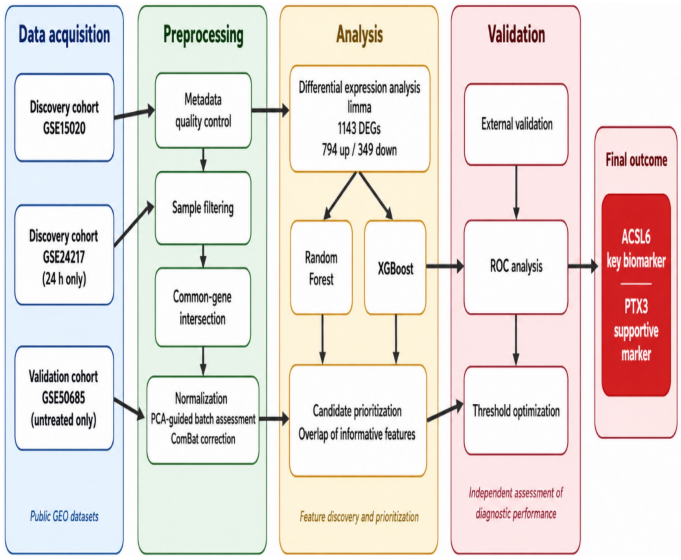

Fig. 1. *Study design and analytical workflow of the integrated transcriptomic and machine learning framework*

Batch-corrected differential gene expression analysis was performed on the discovery cohort using limma, a R package that uses linear modeling with empirical Bayes variance moderation for microarray data analysis (Smyth 2004; Ritchie et al. 2015). This improves stability of variance estimation and increases statistical power, especially in studies where the sample size is relatively small. The design matrix was based on the biological group status (mastitis versus control) and linear models for each gene were fit across all retained genes. Differential expression analysis was performed using Moderated t-statistics obtained from empirical Bayes shrinkage of the variance estimates, refining differential expression assessment. P-values were adjusted for multiple hypothesis testing using the Benjamini–Hochberg false discovery rate (FDR) correction procedure (Benjamini and Hochberg 1995). The genes with $p_{adj} < 0.05$ and $|log2FC| \geq 1$ were considered significantly differentially expressed ones. These thresholds ensured that the selected genes had an appropriate balance of statistical significance and biologically relevant changes in expression. The differential expression analysis therefore served as the first statistical filtering step of the two-step feature prioritization framework designed here. The machine learning algorithms in such a framework were not meant to supplant traditional statistical inference but rather supplement techniques that rank genes in terms of their relative contribution to class discrimination in the dataset. This approach aimed to improve biomarker candidate selection whilst sustaining the statistical requirements of transcriptomic analysis workflows.

2.5 Feature selection using machine learning

In additional analyses, differentially expressed genes identified in the discovery cohort were used as the input feature set for downstream machine learning-based prioritization. All feature selection procedures were performed in the Discovery Cohort only, to avoid information leakage and introducing optimistic bias into the analysis. Differential expression analysis, feature selection, model training, hyperparameter tuning and candidate gene prioritization (Ambroise and McLachlan 2002) using the independent validation dataset were kept completely separate. Two tree-based ML algorithms, Random Forest and XGBoost, were used to rank candidate genes based on their relative ability to discriminate classes. The Random Forests algorithm was set with 500 trees, and variable importance was measured using the mean decrease in Gini impurity (Breiman 2001). The number of variables sampled at each split was governed by the square-root rule, which is the default in all classification tasks, except when internal resampling suggested further optimization. XGBoost was specified as a gradient boosting model with binary logistic objective function (Friedman 2001). The main hyperparameters were a maximum depth of 3, learning rate of 0.05, subsample ratio of 0.8, column sample ratio and boosting rounds as 100. Repeated cross-validation (Steyerberg et al. 2001) was applied in the discovery cohort to evaluate model performance and feature selection stability as a measure of mitigation of overfitting whilst enhancing robustness in ranking features. All work was done with a fixed random seed to ensure computational reproducibility. Candidate biomarkers were consistently identified as highly ranked features by Random Forest and XGBoost. In this architecture, machine learning algorithms served as complementary prioritization tools rather than heuristic competitors and diagnostic classifiers. This approach limited the possibility of model-specific selection bias by first identifying candidate genes with both biological and statistical significance for independent evaluation.

2.6 Model construction and validation

Candidate genes prioritized by Random Forest and XGBoost analyses in the discovery cohort were subsequently evaluated in an independent validation dataset. To minimize information leakage and optimistic bias, the validation dataset was not involved in differential expression analysis, feature selection, model development, or candidate prioritization procedures (Varma and Simon 2006; Vabalas et al. 2019). A logistic regression model was used to evaluate the discriminative potential of the selected candidate genes in the independent cohort. Logistic regression was preferred because of its interpretability, statistical robustness, and suitability for binary outcome modeling in studies with relatively limited sample sizes (Hosmer et al. 2013; Menard 2001). The model estimated the probability of mastitis status based on candidate gene expression profiles. Validation was aimed not at creating a final and absolute diagnostic classifier but by as assessment of whether the candidate genes identified during discovery retained their discrimination ability when applied to an independent dataset. Due to small sample size of the validation cohort, performance metrics were interpreted cautiously as exploratory markers only and not as evidence of clinical utility. Confidence interval estimation and permutation testing were done to provide a robustness assessment of the models as well as to minimize overestimation risk (Ojala and Garriga 2010; Good 2005). This validation framework was designed to provide a statistically conservative evaluation of classification performance within relatively small independent validation cohorts.

2.7 ROC analysis and performance metrics

Receiver operating characteristic (ROC) curve analysis was employed to assess classification performance in the independent validation dataset, and area under the ROC curve (AUC) estimates were computed. The receiver operating characteristic (ROC) analysis characterized the ability of candidate transcriptomic biomarkers to discriminate between mastitis cases and control samples at different classification thresholds (Fawcett 2006; Hanley and McNeil 1982). AUCs were calculated with 95% confidence intervals reported using the non-parametric DeLong method (DeLong et al. 1988). Confidence intervals were necessary to represent the statistical uncertainty of performance estimates given the small size of the validation cohort. In

total, model evaluation included not only point estimates of AUC, sensitivity, specificity and accuracy but also the range of confidence intervals along with unavoidable dataset limitations. The Youden index was used to determine the optimal classification threshold. In addition, threshold optimization was performed to assess the stability of classification performance over a spectrum of decision boundaries beyond only assessing the default threshold. The pROC package in R (Robin et al. 2011) was used to conduct all ROC analyses and calculations. Given the small sample size in the independent validation cohort, ROC-based performance metrics were interpreted with caution and viewed as providing supportive evidence of possible discriminatory ability but not definitive proof of clinical diagnostic utility.

2.8 Permutation test

To internally further test whether the observed classification performance was likely to have occurred by chance, permutation testing was used as a robustness criterion (Ojala and Garriga 2010; Good 2005). In this approach, class labels were randomly reassigned but the original expression matrix was left intact, and model performance was recalculated with each iteration of permutations. Each time the AUC was calculated for each permutation and then compared to the AUC of the original, non-permuted dataset. This resulted in a null distribution of AUC values that allows us to compute the chance of obtaining the same or better classification accuracy by assigning labels randomly. Finally, a permutation-based p-value was calculated based on the proportion of permuted models with AUCs greater than or equal to that of the original model. The permutation testing served as an additional statistical tool to reduce the chance that performance is simply by random class separation. However, due to the relatively small sample size of this independent validation dataset, these findings should be interpreted with caution and were exploratory rather than final proof of performance at a clinical level. In addition, statistical distribution of permutation-derived AUC values was to determine if model performance remained distinguishable from baseline null distribution produced by random labelling. This analysis gave yet additional statistical support for the reliability and stability of the classification patterns detected.

2.9 Functional enrichment analysis

To gain further insights into the biological relevance of the differentially expressed genes identified using the discovery cohort, functional enrichment analyses were performed. Gene Ontology (GO)-based enrichment analyses were conducted using the clusterProfiler package in R (Yu et al. 2012) with separate analyses for biological process (BP), molecular function (MF), and cellular component (CC). In addition to GO analysis, pathway-specific functional interpretations relevant to bovine mastitis were explored in pathways regulating the immune response, inflammatory signaling, host defense mechanisms and potential metabolic disorders. Differentially expressed genes also were systematically evaluated for their participation in previously described inflammation pathways and immune-associated molecular functions relevant to mastitis transcriptomic research. The statistical significance of enrichment was assessed using hypergeometric testing, and p-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) correction method. Adjusted p-values < 0.05 were considered statistically significant. Because the present study primarily focused on candidate biomarker prioritization rather than comprehensive pathway reconstruction, enrichment analyses were interpreted as supportive biological evidence for the identified candidate genes. ACSL6 has been implicated in mastitis due to other experimental evidence available, indicating the importance of pathway-level analyses; however, further studies integrating large-scale transcriptomics with rigorous validation are needed before specific biological roles are established.

3. Results

3.1 General properties of the datasets

Gene expression data related to bovine mastitis was obtained from the Gene Expression Omnibus (GEO) database in the present study. The datasets were grouped into the discovery (GSE15020 and GSE24217) and validation (GSE50685) cohorts. All datasets were generated using a similar microarray platform, ensuring analytical consistency and comparability across datasets. Table 1 provides information on sample sizes and the division of groups for analysis. The discovery datasets had 31 samples (14 controls and 17 mastitis cases), while the validation dataset included 10 samples of the same distribution (5 controls and 5 mastitis cases). This design allowed independent validation separate from model development.

3.2 Data preprocessing and removal of batch effects

Before proceeding to data analysis, all samples were subjected to quality control, identification of a common gene set, and normalization steps. This preprocessing procedure was applied to minimize technical differences between datasets and to make biological variation more distinguishing from technical variation. In this context, the presence of technical variation arising from different datasets was evaluated using principal component analysis (PCA) and is presented in Fig. 2, which shows that, before batch correction, the samples were clustered according to the datasets. In contrast, after correction using the ComBat method, this clustering was markedly reduced and the samples were more clearly separated by biological condition (mastitis and control). These findings suggest that dataset-related technical variation was substantially reduced after ComBat correction and that the datasets became more suitable for integrated downstream analysis.

3.3 Differential gene expression analysis

As a result of the differential gene expression analysis performed on the batch-corrected discovery dataset, a total of 1143 significant

Table 1. Summary of datasets used in the study

Dataset	Platform	Total samples	Control	Mastitis	Role
GSE15020	GPL2112	16	7	9	Discovery
GSE24217	GPL2112	15	7	8	Discovery
GSE50685	GPL2112	10	5	5	Validation

Discovery datasets (GSE15020 and GSE24217) and the independent validation dataset (GSE50685), including total sample sizes and group distributions

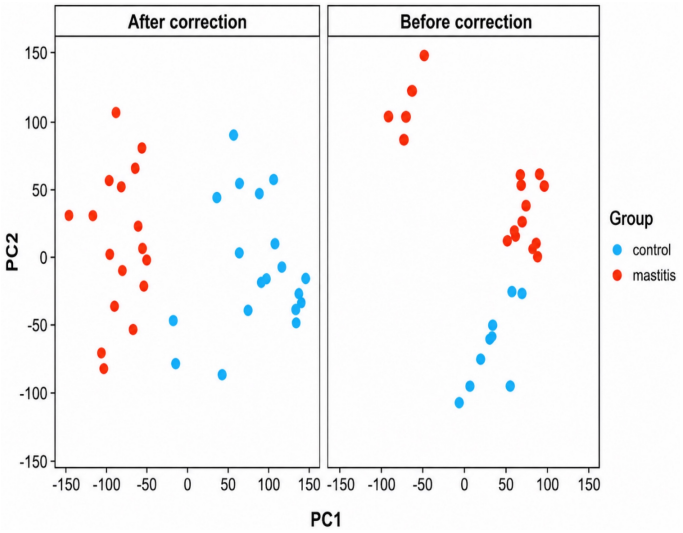

Fig. 2. Principal component analysis (PCA) before and after batch effect correction

differentially expressed genes were identified under the defined statistical thresholds (p_{adj} <0.05 and $|\log_2 FC| \geq 1$). Of these genes, 794 were upregulated and 349 were downregulated. The distribution of identified genes is shown in the volcano plot presented in Fig. 3A. In the plot, it can be observed that significant genes are distributed relative to the predefined threshold values. In addition, the heatmap constructed using genes with the highest absolute $\log_2 FC$ and significance values is shown in Fig. 4, which reveals that the samples are separated into two main groups based on gene expression profiles.

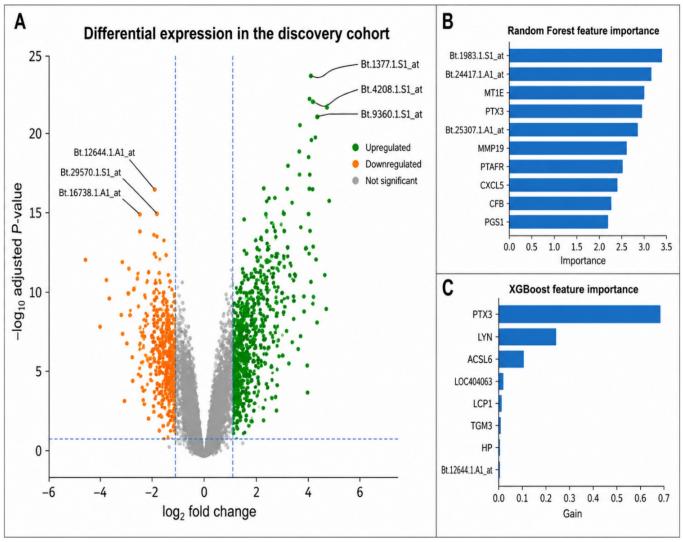

Fig. 3. Identification of candidate genes through differential expression analysis and machine learning-based feature prioritization

These findings indicate that there are distinct gene expression profiles between mastitis and control groups. The complete list of differentially expressed genes is provided in Supplementary Table S1.

3.4 Feature selection using machine learning

Differentially expressed genes were used as input variables for machine learning models to identify biomarker candidates. Random Forest and XGBoost algorithms were subsequently applied and variable importance scores were computed per model. The results from this analysis are shown in Fig. 3B and Fig. 3C, with substantial agreement on gene ranking between the two models; many genes were given high importance by both methods. Of these, two candidates ACSL6 and PTX3 were present at the top of the list in each run, suggesting that they may be biologically relevant. These findings suggest that the genes captured in the analyses may significantly contribute to class separation for these datasets, warranting further biological confirmation and experimental follow-up.

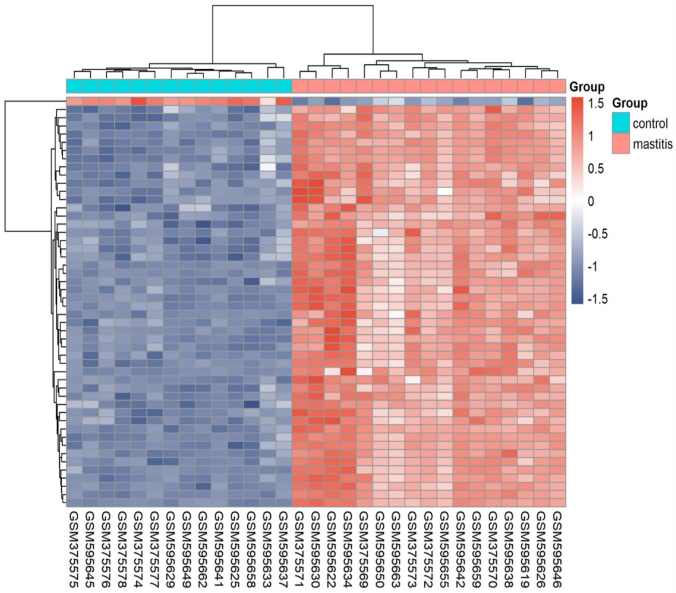

Fig. 4. Heatmap of top differentially expressed genes in the discovery cohort

3.5 Model Performance and Independent Validation

The evaluation of classification model derived from machine-learning analysis was performed using the independent validation dataset. Receiver operating characteristic (ROC) analysis was performed to evaluate model performance, with results shown in Fig. 5. Table 2 shows numerical performance metrics for this criterion. Among the candidate genes, *ACSL6* showed better bootstrapped classification performance compared with *PTX3* in the validation cohort. In addition,

Table 2. Classification performance of candidate biomarkers

Model/ Gene panel	AUC	95% CI	Sensitivity	Specificity	Accuracy
ACSL6	1.00	(0.85–1.00)	1.00	1.00	1.00
PTX3	0.80	(0.44–0.97)	0.80	0.80	0.80
Combined panel	1.00	(0.85–1.00)	1.00	1.00	1.00

Receiver operating characteristic (ROC)-based performance metrics of individual candidate genes and the combined gene panel evaluated in the independent validation cohort. AUC values are presented with 95% confidence intervals calculated using the DeLong method

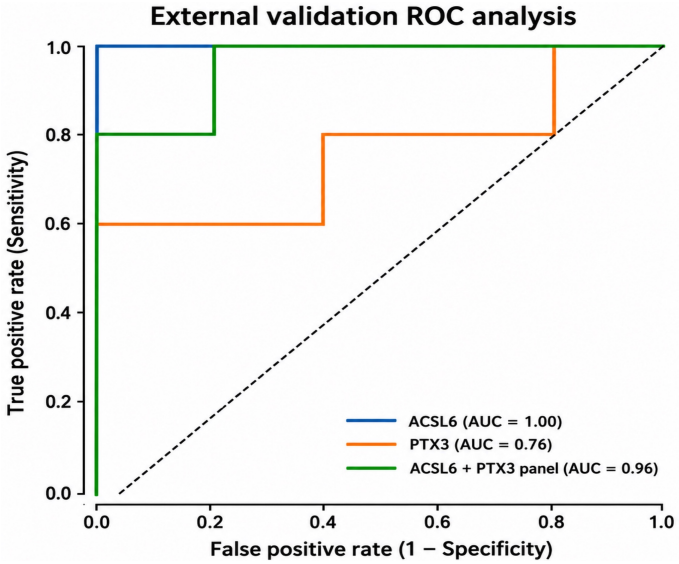

Fig. 5. ROC analysis of candidate biomarkers in the independent validation cohort

the combined gene panel model performed better than the individual gene models resulting in an area under the curve (AUC) of 1.00 (95% CI, 0.85–1.00). Although this suggests a high level of discrimination in the validation dataset, the large confidence interval and small sample size means that these performance estimates should be interpreted with caution. The classification results for the default decision threshold of 0.5 were compared with those determined from ROC-based threshold optimization. Optimization resulted in improved values across major performance metrics. Fig. 6A presents the final classification results with confusion matrix analysis, while Fig. 6B describes threshold optimization results. The overall performance indicators are summarized in Table 2.

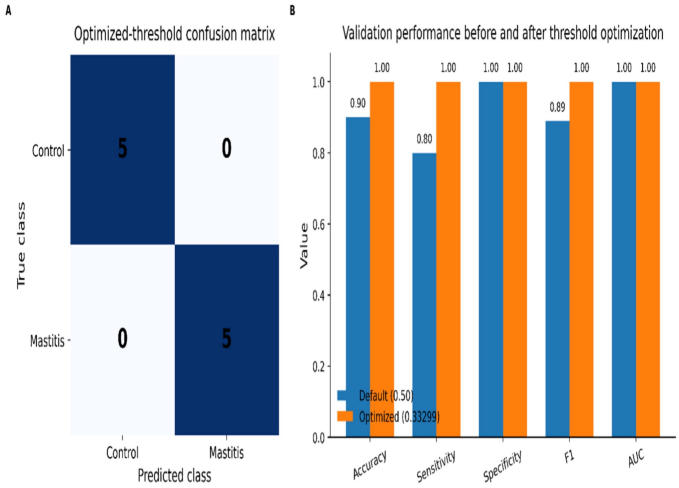

Fig. 6. Classification performance and threshold optimization in the validation cohort

3.6 Gene Level Validation and Correlation Analysis

Expression profiles of the selected candidate genes were scrutinized to validate the model outputs at a gene level. Among the analyzed genes, ACSL6 exhibited significantly higher expression in the mastitis group compared with controls ($p < 0.05$) (Fig. 7). The Spearman correlation analysis was also used to exam interrelationships among candidate genes, and a positive correlation ($r = 0.18$) between ACSL6 and PTX3 was observed (Fig. 7). This suggests a weak positive correlation of ACSL6 and PTX3 expression levels for the samples; however, the biological significance of this association is unknown.

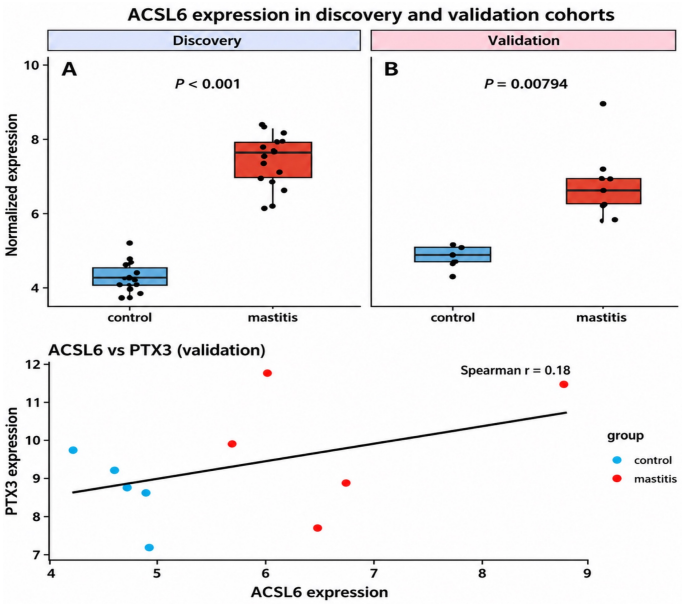

Fig. 7. Gene-level validation and correlation analysis of ACSL6 and PTX3

3.7 Permutation test results

A permutation test was applied to evaluate whether the observed model performance occured at random. The distribution of AUC values obtained from randomly labeled datasets is shown in Fig. 8. The numerical distribution of AUC values obtained from the permutation test is presented in Supplementary Table S3, and this distribution supports that the observed model performance differs from a random distribution. It was observed that the AUC value obtained from the real

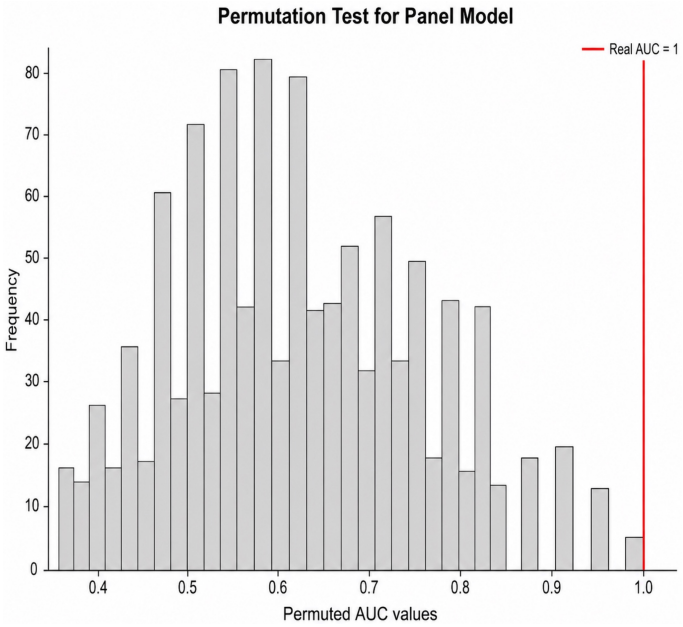

Fig. 8. Permutation test for evaluating model robustness of the candidate biomarker panel

model is located in the right tail of the AUC distribution obtained from permutations. The p-value obtained from the permutation test was found to be less than 0.01. This finding indicates that the probability of obtaining the observed model performance under random labelling is low. These results suggest that model performance reflects the underlying data structure rather than the random labelling.

3.8 Functional enrichment analysis

Gene Ontology (GO) enrichment analysis, performed on the differentially expressed genes, revealed that the significantly enriched biological processes (BP) were mainly related to immune response, defense-mechanisms, and chemotaxis (Fig. 9). The detailed list of enrichment results is presented in Supplementary Table S2. These enrichment findings support the involvement of immune- and inflammation-related biological processes in bovine mastitis and provide supportive biological context for the identified candidate genes.

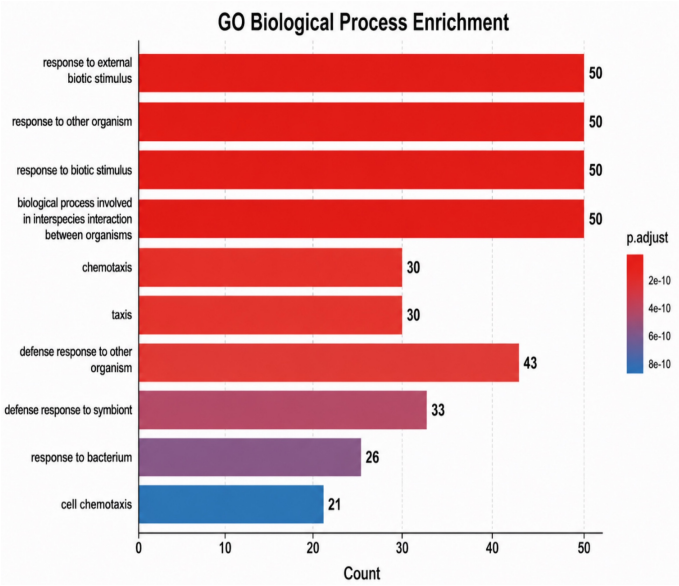

Fig. 9. *GO biological process enrichment analysis of differentially expressed genes*

4. Discussion

In this study, an integrative approach combining transcriptomic analysis with machine-learning methods was applied to identify potential biomarkers associated with bovine mastitis. The findings indicate that, following the integration of different datasets and the removal of technical variation, a consistent transcriptomic signal related to mastitis could be identified. In particular, the clearer separation of samples according to their biological conditions after batch correction using the ComBat method supports the importance of data integration as a critical step in such studies. This is consistent with previous studies emphasizing that technical variation in microarray data may mask biological signals (Leek et al. 2010; Lazar et al. 2013; Johnson et al. 2007). Many genes identified in differential gene expression analysis were known to be involved in pathways related to inflammation and immune response, which reflected the pathophysiological process of mastitis. Cytokines, chemokines and acute phase proteins are key players in innate immune response during mastitis (Rainard and Riollet 2006; Sharifi et al. 2018; Bannerman 2009). In this context, the functional enrichment analyses for immune- and defense-related biological processes support the reliability of differential gene expression results. According to the reported literature, multi-omics data integration improves signal resolution and facilitates the biological interpretation of complex traits (Ramasamy et al. 2008). In addition to differential expression analysis, machine learning methods enabled a more systematic and reproducible discovery of biomarker candidates in this study. The use of both Random Forest and XGBoost algorithms allowed for variable importance metrics to be compared, providing greater confidence in prioritizing important genes. This prioritization strategy is in agreement with previous demonstrating the success of machine-learning approaches in identifying biologically relevant aspects in high-dimensional datasets (Libbrecht and Noble 2015; Chen and Guestrin 2016; Kourou et al. 2015).

An important finding of the machine-learning analyses is the consistent identification of ACSL6 and PTX3. ACSL6 is an acyl-CoA synthetase of the long-chain family of enzymes that catalyzes the activation of long-chain fatty acids, a key step in lipid metabolism. Recent studies have implicated lipid metabolic pathways as important players in the modulation of pro- and anti-inflammatory/immune responses and potential immunometabolic functions (Calder 2015; O'Neill et al. 2016). Because of the clear interrelationship between inflammatory signalling and metabolic adaptation during mastitis pathophysiology, the consistent ranking of ACSL6 across a variety of analytical approaches provides an indirect biological plausibility that this gene may be associated with biologically relevant processes involved in mastitis. However, direct experimental evidences testing the role of ACSL6 in bovine mammary immunity or mastitis progression are lacking, and its specific biological functions remain ambiguous (Calder 2015; O'Neill et al. 2016). In a similar way, PTX3 encodes a long pentraxin protein that is an important component of the innate immune system and has well-documented functions mediating infection and inflammatory responses (Garlanda et al. 2005; Bottazzi et al. 2010). Overall, these observations are both compatible with existing biological paradigms and corroborate the validity of the analytical findings.

An important methodological aspect of this study was the use of an independent validation dataset that allows comparison of the discriminatory power of candidate genes identified within the discovery cohort with their ability to discriminate in samples from outside this cohort. Independent validation is widely advocated for in biomarkers discovery and validation research, due to both its role in reducing optimistic bias as well as improving generalizability (Ioannidis 2005). However, the small validation sample in this study should be taken into consideration when interpreting the classification performance. Small validation cohorts tend to overestimate performance and do not always reflect the full range of biological variability that would be encountered in real-world bioinformatics cohorts (Varma and Simon 2006; Vabalas et al. 2019). This finding suggests that relying only AUC is not sufficient for evaluating model performance. Confidence interval and ROC permutation tests provide a comprehensive and better evaluation of model performance. AUC confidence intervals measure uncertainty in the estimated performance, whilst permutation testing tests whether or not any observed accuracy of a model is statistically greater than what results from chance. Permutation testing can be especially useful for assessing model robustness on datasets with small sample sizes (Ojala and Garriga 2010; Good 2005). Thus, the use of these statistical measures further

strengthens the methodological rigour and robustness of the results from this study. A weak positive correlation between the expression levels of ACSL6 and PTX3 was determined in the samples analysed. Although both genes are involved in biological pathways linked to inflammation and immune responses, the observed correlation coefficient was low suggesting a cautious interpretation. The biological relevance of this association is yet unknown and will need to be validated using larger, experimentally defined cohorts.

This study has several limitations, despite the use of a well-reasoned integrative analytical framework. Above all, the comparatively small number of samples in the datasets may limit how well the predictive performance of the model generalizes. This analytics framework only utilized microarray data, and additional omics platform validation of findings such as RNA-sequencing were not included. Previous studies noted that the combination of multi-omics data points can provide a more rigorous and in-depth understanding of biology (Wang et al. 2009). Collectively, the observations of this study suggest that integrated differential gene expression and machine-learning approaches may represent a promising novel approach for biomarker discovery. Interestingly, the fact that ACSL6 emerged as a consistently prioritized candidate gene among others, suggesting that it may serve as a new biomarker associated with bovine mastitis. However, further experimental validation and evaluation in larger independent cohorts are needed prior to its diagnostic use or clinical application.

5. Conclusions

In this study, an integrated analytical framework combining differential gene expression analysis and machine learning methods to identify novel candidate genes in bovine mastitis. The study merged several transcriptomic datasets and performed stringent batch-effect correction in order to identify biologically relevant expression changes associated with mastitis. Within the independent validation dataset, ACSL6 displayed superior discriminatory power among candidate genes evaluated and PTX3 provided complementary classification information. Although the classification performance observed across the analyzed datasets was encouraging, the small sample size of validation cohort means that the results should be interpreted with caution. Therefore, the results should be considered exploratory and hypothesis-generating rather than definitive clinical diagnostic evidence. Future studies on larger independent cohorts, profiling more RNA-seq datasets, and functional assays will be critical for elucidating the biological significance of ACSL6 as well as understanding its role as a diagnostic biomarker for mastitis. Taken together, these findings suggest a promising approach for biomarker prioritization through the integration of transcriptomic analyses with machine-learning methods in bovine mastitis research.

Declarations

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Conflict of interest: The author declare that they have no conflicts of interest

Acknowledgements: None

Ethics approval: Not applicable

Supplementary data

The supplementary tables are available on journal website. Click the link to access the supplementary tables (Supplementary Tables)

References

Ambroise C, McLachlan GJ. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences of the United States of America 99: 6562–6566.https://doi.org/10.1073/pnas.102102699

Bannerman DD. (2009). Pathogen-dependent induction of cytokines and other soluble inflammatory mediators during intramammary infection of dairy cows. Journal of Animal Science 87: 10–25. https://doi.org/10.2527/jas.2008-1187

Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B: Statistical Methodology 57: 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bolstad BM, Irizarry RA, Astrand M, Speed TP. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185–193. https://doi.org/10.1093/bioinformatics/19.2.185

Bottazzi B, Doni A, Garlanda C, Mantovani A. (2010). An integrated view of humoral innate immunity: pentraxins as a paradigm. Annual Review of Immunology 28: 157–183. https://doi.org/10.1146/annurev-immunol-030409-101305

Bradley AJ. (2002). Bovine mastitis: an evolving disease. The Veterinary Journal 164: 116–128. https://doi.org/10.1053/tvjl.2002.0724

Breiman L. (2001). Random forests. Machine Learning 45: 5–32. https://doi.org/10.1023/A:1010933404324

Calder PC. (2015). Functional roles of fatty acids and their effects on human health. Journal of Parenteral and Enteral Nutrition 39: 18–32. https://doi.org/10.1177/0148607115595980

Chen T, Guestrin C. (2016). XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: Association for Computing Machinery. Pp. 785–794. https://doi.org/10.1145/2939672.2939785

DeLong ER, DeLong DM, Clarke-Pearson DL. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3): 837–845. https://doi.org/10.2307/2531595

Fawcett T. (2006). An introduction to ROC analysis. Pattern Recognition Letters 27(8): 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Friedman JH. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics 29: 1189–1232. https://doi.org/10.1214/aos/1013203451

Garlanda C, Bottazzi B, Bastone A, Mantovani A. (2005). Pentraxins at the crossroads between innate immunity, inflammation, matrix deposition, and female fertility. Annual Review of Immunology 23: 337–366. https://doi.org/10.1146/annurev.immunol.23.021704.115756

Good P. (2005). Permutation, parametric, and bootstrap tests of hypotheses. 3rd edition. Springer, New York, USA. https://doi.org/10.1007/b138696

Guyon I, Elisseeff A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research 3: 1157–1182.

Hackstadt AJ, Hess AM. (2009). Filtering for increased power for microarray data analysis. BMC Bioinformatics 10: 11. https://doi.org/10.1186/1471-2105-10-11

Halasa T, Huijps K, Østerås O, Hogeveen H. (2007). Economic effects of bovine mastitis and mastitis management: a review. Veterinary Quarterly 29(1): 18–31. https://doi.org/10.1080/01652176.2007.9695224

Hanley JA, McNeil BJ. (1982). The meaning and use of the area under a receiver operating characteristic curve. Radiology 143: 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hasin Y, Seldin M, Lusis A. (2017). Multi-omics approaches to disease. Genome Biology 18: 83. https://doi.org/10.1186/s13059-017-1215-1

Hosmer DW, Lemeshow S, Sturdivant RX. (2013). Applied logistic regression. 3rd edition, John Wiley & Sons, Hoboken, USA. https://doi.org/10.1002/9781118548387

Ioannidis JPA. (2005). Why most published research findings are false. PLoS Medicine 2: e124. https://doi.org/10.1371/journal.pmed.0020124

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research 31: e15. https://doi.org/10.1093/nar/gng015

Johnson WE, Li C, Rabinovic A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kosciuczuk EM, Lisowski P, Jarczak J, Majewska A, Rzewuska M, Zwierzchowski L, Bagnicka E. (2017). Transcriptome profiling of staphylococci-infected cow mammary gland parenchyma. BMC Veterinary Research 13: 161. https://doi.org/10.1186/s12917-017-1088-2

Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal 13: 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Nowé A. (2013). Batch effect removal methods for microarray gene expression data integration: a survey. Briefings in Bioinformatics 14: 469–490. https://doi.org/10.1093/bib/bbs037

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Irizarry RA. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11: 733–739. https://doi.org/10.1038/nrg2825

Libbrecht MW, Noble WS. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics 16: 321–332. https://doi.org/10.1038/nrg3920

Lim B, Kim DY, Seo YJ, Lee JY, Kim JM. (2025). Systems biology and integration of multi-omics data. In: Kim JM, Pathak RK, editors, Bioinformatics in Veterinary Science. Springer, Singapore. Pp. 163–183. https://doi.org/10.1007/978-981-97-7395-4_8

Loor JJ, Moyes KM, Bionaz M. (2011). Functional adaptations of the transcriptome to mastitis-causing pathogens. Journal of Mammary Gland Biology and Neoplasia 16: 305–322. https://doi.org/10.1007/s10911-011-9232-2

Lutzow YCS, Donaldson L, Gray CP, Vuocolo T, Pearson RD, Reverter A, Tellam RL. (2008). Identification of immune genes and proteins involved in the response of bovine mammary tissue to *Staphylococcus aureus* infection. BMC Veterinary Research 4: 18. https://doi.org/10.1186/1746-6148-4-18

Menard S. (2001). Applied logistic regression analysis. 2nd edition. Sage Publications, Inc., Thousand Oaks, USA. https://doi.org/10.4135/9781412983433

Moyes KM, Larsen T, Friggens NC, Drackley JK, Ingvartsen KL. (2009). Identification of potential markers in blood for mastitis. Journal of Dairy Science 92: 5419–5428. https://doi.org/10.3168/jds.2009-2088

Naserkheil M, Ghafouri F, Zakizadeh S, Pirany N, Manzari Z, Ghorbani S, Min KS. (2022). Multi-omics integration and network analysis reveal potential hub genes regulating bovine mastitis. Current Issues in Molecular Biology 44(1): 309–328. https://doi.org/10.3390/cimb44010023

O'Neill LAJ, Kishton RJ, Rathmell J. (2016). A guide to immunometabolism for immunologists. Nature Reviews Immunology 16: 553–565. https://doi.org/10.1038/nri.2016.70

Ojala M, Garriga GC. (2010). Permutation tests for studying classifier performance. Journal of Machine Learning Research 11: 1833–1863.

Parker HS, Corrada Bravo H, Leek JT. (2014). Removing batch effects for prediction problems with frozen surrogate variable analysis. PeerJ 2: e561. https://doi.org/10.7717/peerj.561

Rainard P, Riollet C. (2006). Innate immunity of the bovine mammary gland. Veterinary Research 37: 369–400. https://doi.org/10.1051/vetres:2006007

Ramasamy A, Mondry A, Holmes CC, Altman DG. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Medicine 5: e184. https://doi.org/10.1371/journal.pmed.0050184

Ringnér M. (2008). What is principal component analysis? Nature Biotechnology 26: 303–304. https://doi.org/10.1038/nbt0308-303

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43: e47. https://doi.org/10.1093/nar/gkv007

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12: 77. https://doi.org/10.1186/1471-2105-12-77

Schukken YH, Wilson DJ, Welcome F, Garrison-Tikofsky L, Gonzalez RN. (2003). Monitoring udder health and milk quality using somatic cell counts. Veterinary Research 34: 579–596. https://doi.org/10.1051/vetres:2003028

Seegers H, Fourichon C, Beaudeau F. (2003). Production effects related to mastitis and mastitis economics in dairy cattle herds. Veterinary Research 34: 475–491. https://doi.org/10.1051/vetres:2003027

Sharifi S, Pakdel A, Ebrahimi M, Reecy JM, Fazeli Farsani S, Ebrahimie E. (2018). Integration of machine learning and meta-analysis identifies the transcriptomic bio-signature of mastitis disease in cattle. PLoS One 13(2): e0191227. https://doi.org/10.1371/journal. pone.0191227

Smyth GK. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3: 3. https://doi.org/10.2202/1544-6115.1027

Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJC, Vergouwe Y, Habbema JDF. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology 54(8): 774–781. https://doi.org/10.1016/S0895-4356(01)00341-9

Vabalas A, Gowen E, Poliakoff E, Casson AJ. (2019). Machine learning algorithm validation with a limited sample size. PLoS One 14(11): e0224365. https://doi.org/10.1371/journal.pone.0224365

Varma S, Simon R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7: 91. https://doi.org/10.1186/1471-2105-7-91

Wang Z, Gerstein M, Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63. https://doi.org/10.1038/nrg2484

Yu G, Wang LG, Han Y, He QY. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology 16: 284–287. https://doi.org/10.1089/omi.2011.0118

Citation

Bayyurt L. (2026). Integrated transcriptomic and machine learning analysis of the candidate biomarker potential of ACSL6 in bovine mastitis. Letters in Animal Biology 06(2): 38 – 46. *https://doi.org/10.62310/liab.v6i2.364*