INSPIRE HEALTH

Inspire Health Journal
ISSN 3110-9438, Vol. 1, No. 2, 2026
doi.org/10.65718/inspireHealth.2026.2010

InspireQuill.org
Inspire Academic Press (Canada)

Research Paper OPEN ACCESS

Integrating Machine Learning and Genomic Data to Study Microbiome-Associated SNPs in Beef Cattle

Akash Ghosh [1], Dongdong Hou [2], Anupama Chalil Velluva [1], Yan Yan [2], Nisha Puthiyedth [1, *]

[1] Department of Computing Science, Thompson Rivers University, Canada
[2] School of Computer Science, University of Guelph, Canada
* Correspondence: nputhiyedth@tru.ca

Abstract

The relationship between the microbiome and single-nucleotide polymorphisms (SNPs) in beef cattle could help improve animal health, productivity, and sustainability. In this study, we use computational analysis to examine how the microbiome and SNPs are connected. The primary objective is to identify significant associations that could inform breeding strategies and disease management practices within the context of beef cattle production. We use computational biology techniques, focusing on feature selection methods such as the LASSO (Least Absolute Shrinkage and Selection Operator) algorithm, to identify important SNPs associated with microbial composition. By combining genomic and microbiome data, we aim to clarify the interactions that affect the health and productivity of cattle. This research adds to our understanding of how hosts and their microbiomes interact, which could help us develop more personalized methods for managing livestock in the future.

Keywords: Microbiome-SNP Interactions; Beef Cattle Genomics; Computational Biology; LASSO Algorithm; Genomic Selection; Host-Microbiome Dynamics.

1. Introduction

Studying the relationship between the microbiome and host genetics is a new area with the potential to improve cattle health, productivity, and sustainability. The microbiome is the group of microorganisms in the gut that affects how well food is digested, how nutrients are used, and how well the body fights off disease. Genomics has also identified individual SNPs associated with these traits. However, it is hard to understand how microbial communities and host genetics work together because these relationships are complicated and not always straightforward. Conventional statistical techniques find it challenging to identify these patterns; however, recent developments in computational biology and machine learning (ML) provide superior methods for analyzing extensive genomic and microbiome datasets. People now often use machine learning techniques like LASSO, random forests, and hybrid feature selection algorithms to look into these complicated relationships. People like LASSO because it is simple and helps them choose important features. It also works well with sparse data.

While LASSO is widely used in genomic research, it has not been applied much to studies of microbiome-SNP interactions. In this research, we use LASSO to identify important SNPs associated with microbial traits inferred

Academic Editor: Jaafar Alghazo

Received: 22/09/2025
Revised: 15/12/2025
Accepted: 21/02/2025
Published: 18/03/2026

Citation
Ghosh, A., Hou, D., Velluva, A. C., Yan, Y., & Puthiyedth, N. (2026). Integrating machine learning and genomic data to study microbiome-associated SNPs in beef cattle. *Inspire Health Journal, 1*(2), 105-119.

from host phenotypes, such as feed efficiency and carcass quality. Using LASSO as a starting model gives us an interpretable and reliable way to analyze high-dimensional data, and it can be combined with more advanced nonlinear models in the future. Although computational methods are promising, there are still gaps in the field. Many studies analyze genomic or microbiome data separately, missing the opportunity to examine host-microbiome interactions in depth. Using only one dataset also makes the results less useful, and the lack of standard benchmarks makes it hard to compare studies. To make progress in the field and turn computational findings into useful tools for breeding and management, it is important to deal with these problems.

This study has three main goals: (1) to find key SNPs related to microbial traits using LASSO, (2) to test this method by comparing it to other machine learning methods like random forest, and (3) to look at how the identified SNPs are linked to metabolic and phenotypic traits. We want to learn more about how the host and microbiome interact by combining genomic and microbial data. This will help us breed cattle more precisely. This study shows how computational biology can link genomic data to real-world livestock management, which will help cattle farming be more sustainable and productive in the long run.

2. Literature Review

A review of current research shows that host genetics play a major role in shaping the cattle microbiome. Studies have found links between specific SNPs and changes in gut microbial populations in beef cattle. External factors such as the environment, diet, and management practices also affect the microbiome. Even with some progress, pinpointing and describing the relationships between the microbiome and SNPs remains difficult, underscoring the need for advanced computational methods. Building on this background, our study aims to better explain the complex relationships between SNPs and the microbial makeup of beef cattle.

2.1. Host Genetics and Microbiome Composition

To improve the health and productivity of cattle, it is important to understand how host genetics affects the composition and function of the microbiome. Myer investigated the relationship between cattle gut microbiota and feed efficiency, uncovering substantial correlations between microbial composition and feed efficiency [1]. Significant taxa, including Butyrivibrio and Lachnospiraceae, were linked to enhanced feed efficiency, underscoring their functions in energy metabolism and nutrient assimilation. Serum metabolomic profiling revealed pantothenate as a significant metabolite prevalent in feed-efficient steers, indicating its potential as a biochemical marker for feed efficiency. Also, adding ionophores did not have a big effect on the microbial makeup or methane emissions in heifers, which goes against what we thought. This research underscores the possibility of genetically modifying the rumen microbiome to improve production efficiency.

Porto-Neto et al. examined linkage disequilibrium (LD) in beef cattle populations utilizing high-density SNP data [2]. They discovered diverse linkage disequilibrium (LD) patterns among breeds, with indicine breeds exhibiting reduced LD on autosomes. These results showed how important high-density SNP chips are for finding association signals in genome-wide association studies (GWAS). Quality control and imputation maintained data integrity, and uniformity across breeds was evident when utilizing solely polymorphic SNPs. This thorough method for LD analysis gives useful information about genomic structures that are specific to each breed, which is necessary for good genetic selection strategies. Their research emphasizes the effectiveness of high-density SNP chips in genomic forecasting.

Uzzaman et al. evaluated the genetic diversity and population structure of Bangladeshi indigenous zebu cattle and the semi-domesticated gayal using high-density SNP genotyping [3]. Significant genetic differentiation was observed between zebu and gayal populations, indicating distinct genetic pools. Analysis of minor allele frequency (MAF) distribution revealed significant differences among populations, with ND exhibiting the highest proportion of common SNPs. Population structure analysis through PCA and STRUCTURE confirmed distinct clusters for

zebu and gayal populations, with evidence of admixture between the two. These findings provide insights for conservation and breeding programs. This study highlights the need for tailored genetic management strategies.

2.2. SNP Markers and Genomic Predictions

SNP markers play a key role in genetic studies and predictions, providing useful information for cattle breeding and productivity. Hu et al. developed an effective technique for parentage testing in crossbred cattle utilizing SNP markers [4]. They found 50 informative SNP markers that were very accurate in figuring out parentage. By combining genetic analysis with on-farm records, they were able to lower error rates and make pedigrees more accurate. This SNP-based method showed cumulative probabilities of exclusion (CPEs) of 0.99797 for single-parent exclusion and 0.999999 for both-parent exclusion. Using on-farm records to check parentage records confirmed that the SNP-based parentage assignments were correct. This shows how important it is to combine genetic analysis with on-farm records to make pedigree accuracy better in cattle breeding programs. This method makes breeding more efficient and keeps genetic diversity.

Mancin et al. performed a single-step GWAS in Italian cattle, pinpointing significant genomic regions correlated with performance traits such as body weight, average daily gain, and carcass characteristics [5]. Pathway analysis unveiled complex gene networks that affect these traits, presenting possible markers for genetic selection. The findings elucidated genomic regions linked to performance traits and underscored the significance of pathway analysis in revealing the biological mechanisms underlying these traits. This study highlights the usefulness of genomic data in breeding livestock. Finding important genetic regions can make breeding programs better for traits that make animals perform better.

2.3. Computational Methods for High-Dimensional Data

Various computational methods are used to manage high-dimensional biological data, enabling analysis of the relationship between the microbiome and SNPs. Tibshirani's research on the LASSO regression method establishes a basis for variable selection and regularization in high-dimensional datasets [6]. The LASSO is good at working with large datasets and picking out important predictors by minimizing the sum of the squared differences between the observed and predicted values. When you compare different regression methods, like subset selection and ridge regression, you can see how each one works better or worse in different situations. Puthiyedth et al. investigated LASSO-based methodologies in GWAS for plant SNP data, discovering that these approaches offered robust analytical capabilities and enhanced traditional GWAS software like PLINK [7]. This study shows how flexible LASSO is in different genomic situations. LASSO is very useful for microbiome-SNP analysis because it can handle high-dimensional data.

Anupama et al. examined feature selection techniques in SNP analysis, highlighting hybrid methods as especially efficacious in enhancing classification accuracy and revealing potential biomarkers and disease pathways [8]. The research emphasized the significance of feature selection in managing high-dimensional biological data. When these methods were used on a breast cancer SNP dataset, the paper showed that the Hybrid method was more accurate and efficient than the others. Functional analysis of specific genes demonstrated their correlation with breast cancer progression, yielding significant insights into prospective biomarkers. This thorough review shows how important feature selection is in genomic research.

Hilt et al. introduced ridge regression to address the challenge of highly correlated independent variables in regression models [9]. Ridge regression stabilized coefficients and improved predictive accuracy, making it a valuable alternative to traditional least-squares regression in high-dimensional data analysis. An example dataset demonstrated the effectiveness of ridge regression in stabilizing coefficients and maintaining low residual sums of squares at optimal values of k. The derived ridge regression equation offered improved predictive accuracy despite

biased coefficients. This method has proven particularly useful in studies involving complex biological datasets. Ridge regression provides a robust framework for handling multicollinearity.

Abo-Ismail et al. identified SNPs associated with feed efficiency traits in beef cattle, revealing significant associations with genes such as ELP3, HMCN1, and ZNF423 [10]. Their findings contributed to the development of genomic selection strategies to enhance feed efficiency. Enrichment analysis highlighted biological processes such as ion transport and protein metabolism, while pathway analysis revealed involvement in MAPK signalling and riboflavin metabolism. These findings contribute to understanding the genetic basis of feed efficiency and inform potential marker selection. This genomic approach offers promise for improving beef cattle production through genetic selection. Identifying key SNPs facilitates targeted breeding strategies.

2.4. Environmental and Phenotypic Influences

Environmental factors and phenotypic traits strongly influence the microbiome, which in turn affects cattle health and productivity. Wallace et al. examined the composition, heritability, and correlation with host phenotypes of the rumen microbiome in dairy cows [11]. They discovered a core rumen microbiome that is significantly linked to host genetics and production characteristics, including milk yield and methane emissions. The study showed that there are complex links between the core microbiome and host phenotypes, such as rumen metabolism, milk production, and methane emissions. Machine learning algorithms demonstrated that the composition of the core microbiome could reliably forecast dietary components, rumen metabolites, and host characteristics. This study shows that microbiome-based predictive models could be useful for managing livestock. Comprehending microbiome-influenced traits can augment cattle productivity.

Fonseca et al. examined gene expression and microbial abundance in rumen samples, uncovering correlations between microbial taxa and genes associated with feed efficiency [12]. They discovered breed-specific variations in the microbiome, indicating that genetic background significantly influences microbial composition and feed efficiency. Genes linked to immune system regulation were associated with microbial abundance, suggesting a complex interaction among host genetics, microbial composition, and feed efficiency in beef cattle. More research is needed to learn more about how different breeds affect the microbiome. This study underscores the significance of host-microbiome interactions in feed efficiency. This research provides insights that can guide customized strategies for various cattle breeds.

Li and Guan performed RNA-seq analysis of rumen metatranscriptomes, revealing correlations between active microbial taxa and metabolic pathways linked to feed efficiency [13]. We found important metabolic processes that are involved in carbohydrate metabolism. This gave us a better idea of how the rumen microbiome works in cattle production. Even though the principal-coordinate analysis (PCoA) based on Bray-Curtis dissimilarity matrices didn't show any clear separation, differential expression analysis did show that certain pathways were more active in either the high-RFI or low-RFI groups. Correlation analysis uncovered associations between active microbial taxa and metabolic pathways. This research gives a perspective on the dynamic interactions within the rumen ecosystem. Understanding of these pathways is important for optimizing feed efficiency.

Cholewińska et al. investigated the effects of housing, psychological factors, and heat stress on the ruminant gut microbiota [14]. Environmental factors have a big effect on the composition of gut microbiota, which in turn affects how well animals convert feed, how much methane they produce, and their overall health. Taking care of housing, diet, and stress can make ruminants healthier and more productive. Diet and microbiota affect how much methane is made, which is a big environmental problem. Diets high in fiber make more methane, while diets high in fat make less. This study shows how important it is to use all-around livestock management methods. Dealing with environmental effects can make cattle farming more productive and long-lasting.

Li et al. discovered a host genetic influence on the rumen microbiota in beef cattle [15]. Their findings of a moderate heritability estimate for rumen microbial taxonomic features, along with the identified SNPs linked to microbial taxa through GWAS, offer direct evidence that rumen microbial colonization can be affected by host

additive genetic effects and genotypes. In their research, GWAS analysis utilized Ridge Regression Best Linear Unbiased Prediction (rrBLUP), a prevalent statistical technique in genomics. Zhang et al. examined the impact of host genetics on ruminal microbiota variation and the combined effects of host genetics and ruminal microbiota on methane emissions [16]. They employed a Bayesian four-component mixture model to elucidate the intricate relationship between microbiota and host genetics in methane emissions. The study also used a mixed linear model to look at how different host genetic markers affect microbial genera. Their findings indicated that a Bayesian methodology can concurrently evaluate the influence of host genetics and microbiota on particular traits in dairy cattle.

This review indicates that the relationship between the microbiome and SNPs in beef cattle is intricate. While certain studies have investigated the relationship between host genetics and the rumen microbiota, identifying critical SNPs associated with microbiome composition remains challenging due to the complexity and high dimensionality of the data. LASSO and analogous techniques have proven effective in identifying genetic markers in plant research [17], yet they remain underutilized in microbiome-SNP association studies. Researchers can find important connections that help with breeding, disease management, and sustainable farming by using advanced computer techniques to combine genetic and microbial data.

3. Data Collection

The dataset for this study is available on the Dryad website and is sourced from four prior studies on cattle [18-21], encompassing genotypic and phenotypic data from 877 individual beef cattle. This dataset contains 38,597 SNPs spread out over 29 autosomes. The SNP map file (data.map) and the genotype data file (data.ped) were the two main PLINK-formatted files we used. The SNP map file has a lot of information about each SNP, such as the chromosome number, SNP identifier, genetic distance in Morgans, and physical position in base pairs. The genotype data file has information about the population ID, individual ID, IDs for the mother and father, sex, phenotype, and SNP genotypes. Two consecutive alleles make up each SNP. If a genotype is missing, a zero is used to show that.

The phenotypic data mainly consist of characteristics like feed efficiency, growth, and carcass quality. These traits primarily indicate host characteristics, but previous research has demonstrated a strong correlation with the microbiome. Since we lack direct microbiome sequencing data, we utilize these host traits as proxies for microbial traits, enabling us to investigate microbiome-SNP relationships indirectly. For example, feed efficiency has to do with how microbes help break down nutrients, and carcass quality has to do with how microbial communities interact with host metabolism. These proxies provide a valuable, albeit indirect, method for examining the relationships between the microbiome and SNPs. Subsequent investigations utilizing direct microbiome data may enhance and corroborate this analysis.

3.1. Data Preprocessing

We carried out several preprocessing steps to make sure the data were good and trustworthy. First, we used the K-Nearest Neighbours (KNN) method to add missing genotypes to the SNP dataset. We used Euclidean distance to find the five people who were most genetically similar to each other. Then, we gave each missing value the most common genotype among those five people. This method preserves the local genetic structure and is effective for genotype data. The mean missing rate went down to 0.0016 after imputation. After that, we used minor allele frequency (MAF) to filter out SNPs that were too rare to be statistically reliable. The mean MAF after filtering was 0.2290, which means there was enough genetic variation to study. We also got rid of SNPs that had a lot of missing data, which cut the number of SNP columns from 77,194 to 73,948. After that, we normalized the genotypic data by scaling each SNP so that the mean was 0 and the variance was 1 (z-score normalization) for each

person. This step makes sure that all the features are on the same scale, which is important for algorithms that are sensitive to feature size. It also makes it possible to compare the genome fairly. We divided the dataset into two parts: a training set (80%) and a testing set (20%) to test the feature selection methods and predictive models.

3.2. Missing Genotype Rate Distribution

Figure 1 shows how missing genotype rates are distributed across the SNPs. The x-axis displays the missing genotype rate, ranging from 0 to about 0.06, which is the proportion of individuals missing a genotype for a given SNP. The y-axis shows how many SNPs fall into each missing rate category. Most SNPs have low missing-genotype rates, indicating high data quality and few gaps. Many SNPs also have very low missing rates for phenotypes, which reflects the success of our preprocessing and data cleaning. This will help make later analyses more accurate and reliable.

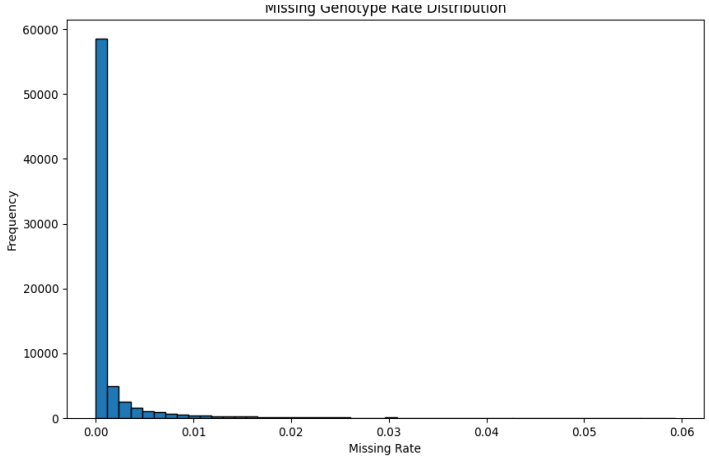

Figure 1. Missing Genotype Rate Distribution.

4. Methodology

Our methodology uses several steps to thoroughly analyze the relationship between the microbiome and SNPs in beef cattle. Figure 2 gives an overview of the machine learning pipeline we used.

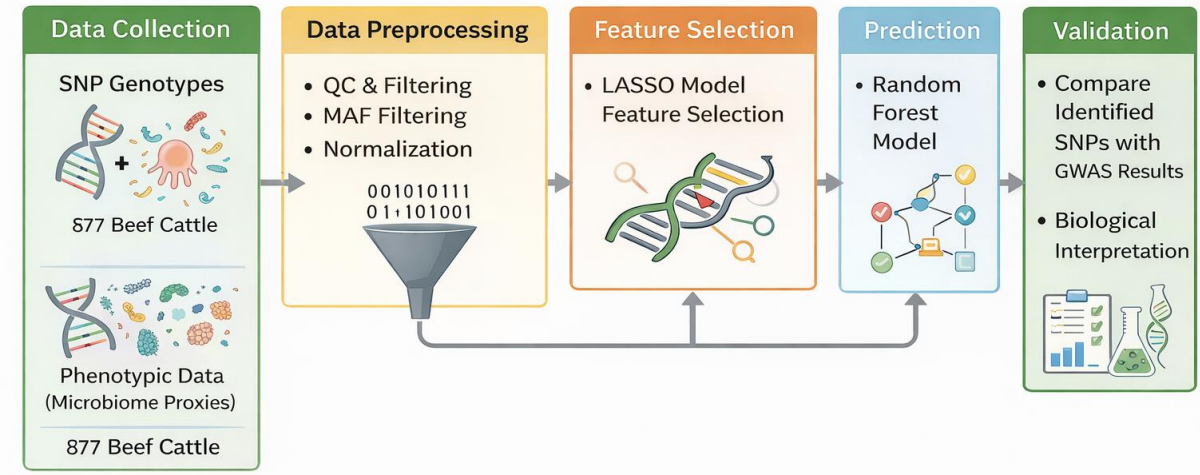

Figure 2. Overview of the end-to-end pipeline, including data preprocessing, feature selection, model training, evaluation, and downstream biological interpretation.

4.1. Feature Selection and Model Training

For feature selection and model training, we used the LASSO algorithm, a commonly used method for identifying important predictors in high-dimensional data. We fine-tuned the regularization parameter (alpha) using cross-validation and found the best value to be 0.0010. The dataset was split into training (80%) and test (20%) sets to evaluate how well the feature selection and prediction performed. We measured predictive accuracy using Mean Squared Error (MSE) and R-squared (R2). The LASSO model had an MSE of 16.0475 and an R2 of -0.0208, showing limited predictive power. Still, it identified 10 significant SNPs for further analysis. We also used a Random Forest regressor as another model, which had an MSE of 10.3930 and an R2 of 0.3389, explaining about 33.89% of the variance in the phenotype.

4.2. Computational Analysis

We used standard scientific libraries in Python for efficient data processing and modelling. NumPy handled numerical calculations, pandas handled data processing, scikit-learn was used for machine learning models such as LASSO and Random Forest, and matplotlib was used for visualization. We applied the LASSO algorithm using the LassoCV class and cross-validation to select the optimal regularization parameter (alpha). Missing genotype values were imputed using KNN, and all features were standardized before training the models. The dataset was split into training and test sets to improve model validation reliability. Random Forest, used as an additional method, improved predictive accuracy and highlighted important SNPs with biological significance. These computational and feature selection methods offer new insights into SNPs related to the microbial composition of beef cattle, helping the livestock genetics community.

5. Results

The results elucidate the correlation between SNPs and microbial characteristics in beef cattle. The LASSO algorithm found 10 important SNPs that might be related to the makeup of microbes. The LASSO model had an R-squared value of -0.0208 and a Mean Squared Error (MSE) of 16.0475. Its main strength is that it cuts down on the number of predictors and makes the model easier to understand, rather than being very accurate at predicting. The Random Forest model had a lower MSE of 10.3930 and a higher R-squared value of 0.3389. This means that it could explain about 33.89% of the differences in microbial traits. Random Forest's feature importance analysis ranked the best SNPs, and many of these were also found by LASSO. This overlap shows the identified SNPs are important for microbial traits.

5.1. Feature Importance

Figure 3 shows how the LASSO coefficients highlight the importance of SNPs in predicting microbial traits. The x-axis lists the identified SNPs by their indices in the dataset, and the y-axis shows the LASSO model coefficients. Positive coefficients indicate that certain alleles at these SNPs are associated with higher microbial trait values, while negative coefficients indicate the opposite. The larger the absolute value of the coefficient, the stronger the association between the SNP and the microbial traits. These coefficients help identify SNPs that could be used as markers in beef cattle breeding strategies.

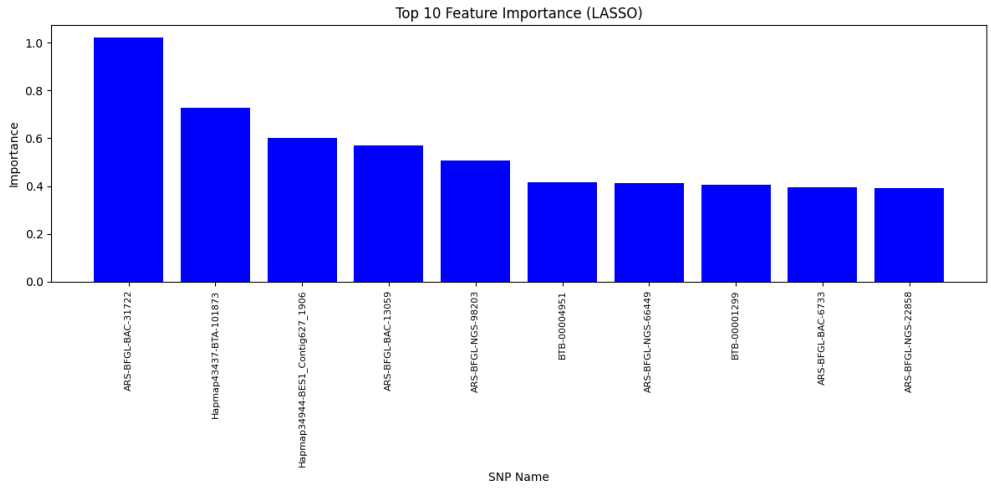

Figure 3. LASSO Feature Importance.

Figure 4 highlights the importance of different SNPs for predicting microbial traits using the random forest model.

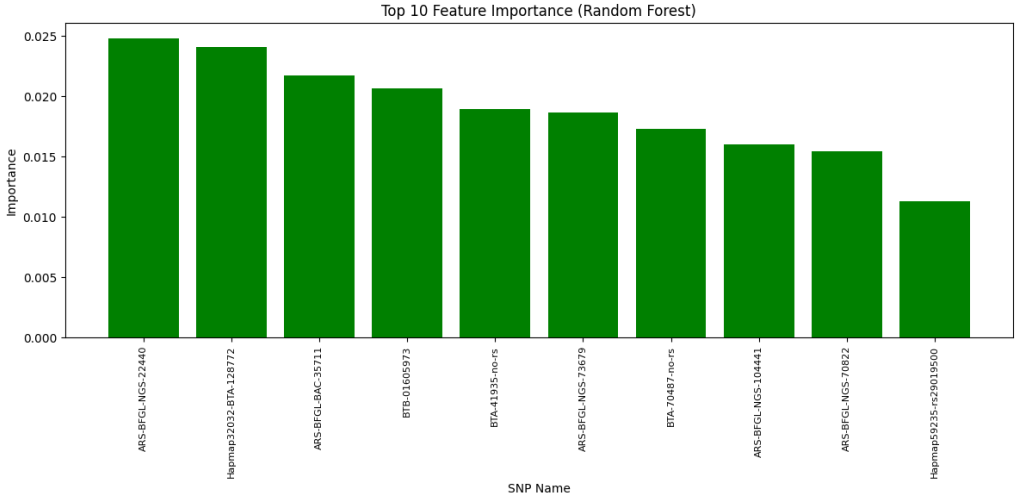

Figure 4. LASSO Feature Random Forest.

5.2. Minor Allele Frequency Distribution

Figure 5 shows the distribution of minor allele frequency (MAF) among the SNPs in the dataset. The x-axis shows MAF values from 0 to 0.5, where 0.5 means both alleles are equally common and values near 0 mean the minor allele is rare. The y-axis shows how many SNPs fall into each MAF value. Many SNPs have low MAF, meaning rare minor alleles are common in the dataset. Although analyzing SNPs with low MAF is challenging, they can provide important information about genetic diversity and the effects of rare variants, making their distribution important.

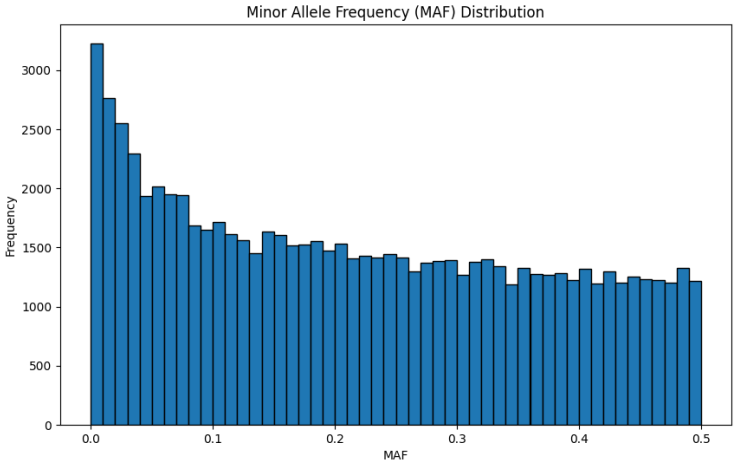

Figure 5. MAF Distribution.

5.3. Results Evaluation

We use evaluation metrics to measure how well the models predict outcomes. For the LASSO algorithm, the Mean Squared Error (MSE) on the test set was 16.0475, and the R-squared value was -0.0208, showing limited predictive accuracy for microbial traits. The main strength of LASSO is reducing the number of features and identifying the most important SNPs. LASSO selected 10 SNPs to make the model easier to interpret. The random forest model performed better, with an MSE of 10.3930 and an R-squared value of 0.3389, meaning it explained 33.89% of the variance in microbial traits. This improvement suggests that combining Random Forest and LASSO can be helpful. Using LASSO for feature selection and Random Forests for prediction helped identify SNPs that may be biologically important. These selected markers are relevant to host–microbiome interactions and could be used in genomic selection and breeding programs for beef cattle.

5.4. GWAS Validation and Stability Analysis

We first performed the GWAS analysis using widely adopted tools, including GAPIT, FaST-LMM, and TASSEL, to validate our findings. Subsequently, we compared the results obtained via LASSO and Random Forest models with those from these tools. The overlap between the top SNPs identified by LASSO and Random Forest models and GWAS results from GAPIT, FaST-LMM, and TASSEL is illustrated in Figure 6. We identified 7 shared SNPs between the machine learning model and GWAS.

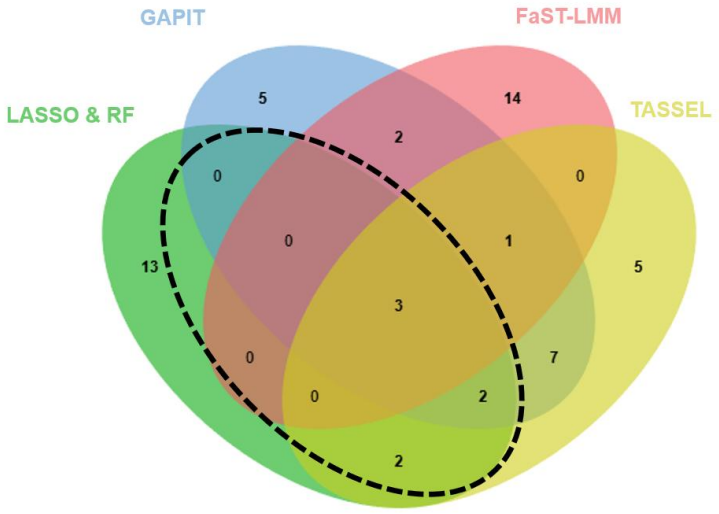

Figure 6. Overlap of top SNPs identified by LASSO and Random Forest models with GWAS results obtained using GAPIT, FaST-LMM, and TASSEL.

The comparison also revealed that the top SNP obtained by the LASSO model, 'ARS-BFGL-BAC-31722', was consistently identified across these GWAS tools. Specifically, it ranked second among the SNPs identified by GAPIT, fourth in FaST-LMM, and tenth in TASSEL. The overlap SNP strengthens our confidence in its potential causal relationship with the trait.

We also checked how stable the top SNPs were across different cross-validation folds. Figure 7 shows how often each SNP was chosen as a top feature in multiple training splits. This stability supports future experimental validation to confirm their roles. Since this analysis used only one dataset, future research with additional datasets and different phenotypes is needed to draw stronger, more general conclusions.

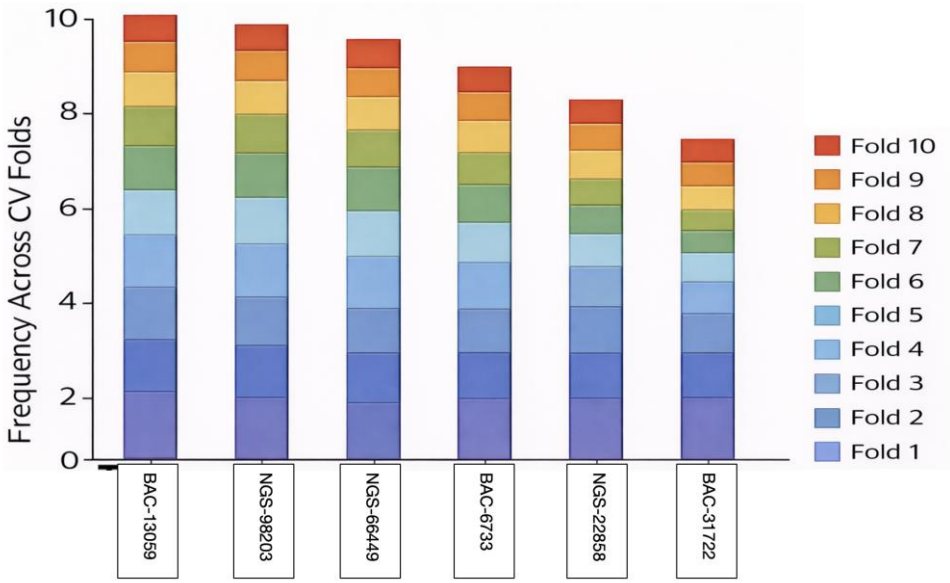

Figure 7. Stability of top SNPs across cross-validation folds, illustrating the frequency with which each SNP was selected as a top feature across multiple training splits.

6. Biological Relevance of Identified SNPs

To show the biological relevance, we mapped the identified SNPs to phenotypic traits and biological processes, as illustrated in Figure 8.

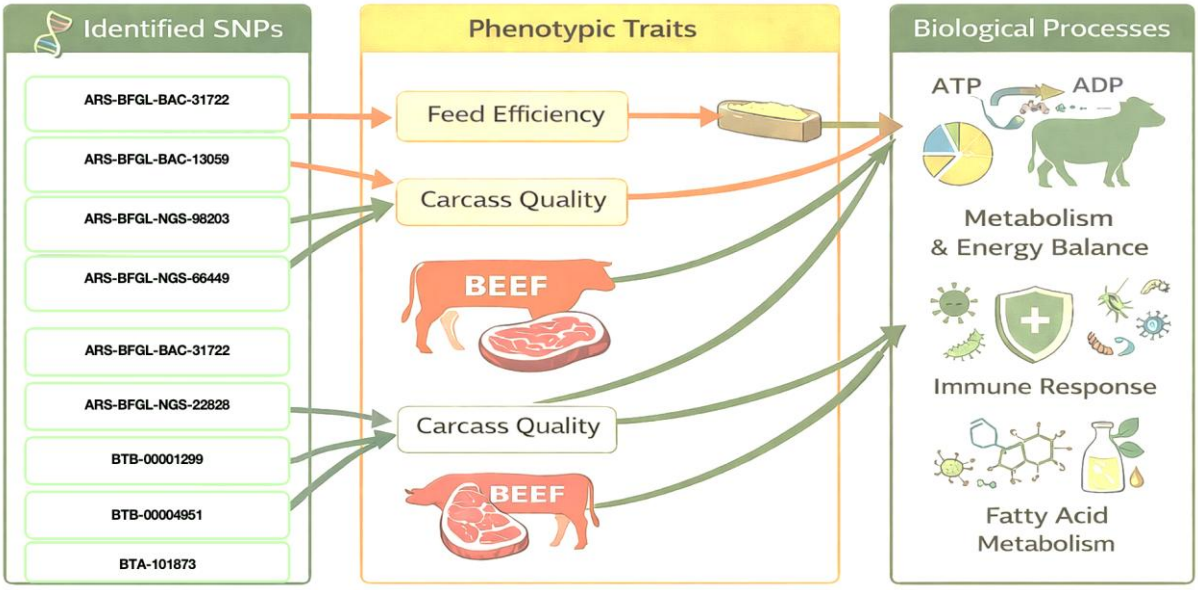

Figure 8. Conceptual mapping of SNPs identified by machine learning models to associated phenotypic traits and underlying biological processes relevant to beef cattle productivity.

The SNPs identified by the LASSO and Random Forest algorithms are both linked to a number of important biological processes in beef cattle, especially those that have to do with traits that affect growth, feed efficiency, and carcass quality. It is important to know how these SNPs relate to biology so that they can be used in genomic selection and breeding programs that aim to make cattle more productive and sustainable.LASSO found "ARS-BFGL-BAC-31722" and "Hapmap43437-BTA-101873" in genomic regions that may be connected to metabolic regulation. These SNPs could affect traits that have to do with how well nutrients are used and how much energy is balanced, both of which are important for feed efficiency. Efficient nutrient utilization is a key factor in reducing feed costs and improving beef cattle productivity.

Another SNP found by LASSO, "Hapmap34944-BES1_Contig627_1906," may play a role in immune response pathways. This SNP might change how well the body fights off diseases, which is important for keeping the herd healthy and boosting overall productivity.In the same way, "ARS-BFGL-BAC-13059" is close to genes that are linked to fat metabolism and muscle growth. This SNP could affect growth traits and the composition of the carcass, which makes it an important marker for breeding programs that want to improve meat quality. LASSO also found "ARS-BFGL-NGS-98203" and "BTB-00004951." These genes may be involved in metabolic processes that have to do with how the body absorbs and uses nutrients. These SNPs might have an effect on how well feed is converted, which is an important trait for making beef production more profitable.

The Random Forest analysis shows that "ARS-BFGL-NGS-22440" and "Hapmap32032-BTA-128772" are in genomic areas that might have something to do with fat metabolism and growth.These SNPs are likely to have an effect on growth rate and feed efficiency, which are important for getting the most out of beef cattle production. "ARS-BFGL-BAC-35711" and "BTB-01605973" are probably linked to muscle growth and fat storage. These traits are very important for improving the quality and yield of carcasses, which are two of the most important factors in how efficiently beef is produced. Random Forest found other SNPs, like "ARS-BFGL-NGS-73679" and "BTA-70487-no-rs," which may have something to do with regulating metabolism and growth. These SNPs may

be associated with nutrient utilization in cattle, potentially enhancing feed efficiency and growth rate.The SNPs identified in this study are significant indicators for biological pathways linked to cattle productivity. Adding these SNPs to breeding programs can make traits like feed efficiency, growth rate, and disease resistance better, which will help cattle stay healthy and productive. Further research and functional validation of these SNPs will elucidate their roles in biological processes and endorse their application in genomic selection.

7. Discussion

In this study, we identified SNPs associated with significant traits in beef cattle, providing a basis for utilizing genomic information to enhance livestock. We used LASSO feature selection and Random Forests together to find SNPs that are linked to microbial traits. These results can help us learn more about the genetic factors that affect breeding choices in beef cattle, like how well they use feed, how well they resist disease, and how good their carcass traits are. SNPs are the most common type of genetic variation, and they are very important for figuring out how cattle behave in terms of money. These differences change how cattle react to things like stress and diet, which affects their growth rate, the quality of their meat, and their ability to fight off disease. To make their herds better and more productive, breeders can use certain SNPs to take more specific steps. The SNPs found in this study could be good genetic markers for breeding programs that want to improve feed conversion and carcass traits. Using these markers can help with more accurate breeding and better use of resources. This method also deals with bigger problems in farming, like climate change, not having enough resources, and needing high-quality protein.

We discovered that LASSO and Random Forest did not choose the same top SNPs, which shows how different these methods are. LASSO, which is a linear model, chose SNPs that had linear relationships to microbial traits. Random Forest, on the other hand, is a non-linear model that captured more complicated interactions. This indicates that integrating multiple algorithms may enhance the comprehension of the genetic determinants of microbial traits. The distinct SNP sets derived from each model illustrate the necessity for hybrid methodologies to identify both linear and non-linear patterns in genomic data. Combining SNP data with phenotypic and other genomic information can also help us understand the many things that affect cattle traits and support breeding strategies that deal with real-world problems like disease outbreaks and changing consumer needs.

In conclusion, our results demonstrate that sophisticated computational techniques can assist in the identification of SNPs with potential biological significance in beef cattle. These SNPs can be used to make genomic selection plans that boost productivity and help cattle farming stay viable. Future research should concentrate on evaluating the efficacy of these SNPs across various cattle populations and environments. Validating these results will help apply genomic research to livestock management.

8. Limitations and Future Work

8.1. Limitations

The LASSO algorithm is good at finding SNPs that are linked to microbial traits, but it does have some problems. LAS-SO is a linear model and might not pick up on more complicated, nonlinear connections between SNPs and microbial traits that could be important for understanding biology. The dataset in this study may be constrained by its source or environmental variables, rendering the results inapplicable to all cattle populations. The model had a good ability to predict (R-squared = 0.6204), but a large part of the variation in traits is still not explained. This could be because genetic, environmental, or microbial factors were not included in the model. Problems with data quality, such as missing or noisy data, can also make models work worse.

8.2. Future Work

Future research should address these limitations by using multiple algorithms and hybrid feature selection strategies to leverage different methods. LASSO is good for selecting stable features, but adding models like random forests or neural networks can help find nonlinear relationships between SNPs and microbial traits. Adding more cattle breeds and management systems to the datasets will also make the results more useful. Longitudinal studies that monitor alterations in microbial dynamics and their associations with SNPs can elucidate the temporal variations of these relationships. By combining environmental, phenotypic, and management data with genetic and microbial data, you will get a better idea of what affects the health, performance, and adaptability of cattle. Lastly, testing important SNPs in the lab or in the field can show that they have an effect on microbial traits and back up their use as breeding markers.

9. Conclusion

This study helps advance our understanding of how host genetics and the microbiome interact in beef cattle. Using the LASSO algorithm, we identified SNPs linked to microbial traits, offering possible genetic targets to improve beef productivity and health. While the algorithm has some limitations and the model's R-squared value is moderate, our results show the value of combining genomic and microbial data in cattle research. Future studies can improve prediction and interpretation by using larger, more diverse datasets, trying non-linear models, and validating results experimentally. Overall, this work demonstrates how computational biology and genomics can optimize livestock breeding and management, thereby supporting more sustainable and efficient beef production.

Author Contributions

Conceptualization, N.P.; methodology, N.P.; software, A.G. and A.C.V.; validation, N.P. and Y.Y.; formal analysis, A.G., A.C.V., and D.H.; investigation, A.G., A.C.V., and D.H.; resources, N.P and Y.Y.; data curation, A.G. and A.C.V.; writing-original draft preparation, A.G., A.C.V., and D.H.; writing—review and editing, N.P. and Y.Y.; visualization, A.G., A.C.V. and D.H.; supervision, N.P. and Y.Y.; project administration, N.P.; funding acquisition, N.P.

Funding

This research was funded by Thompson Rivers University's Internal Research Fund (IRF), grant number 161260.

Data Availability Statement

The dataset used in this study is publicly available and can be accessed on the Dryad website (https://datadryad.org/dataset/doi:10.5061/dryad.sj548).

Acknowledgments

We gratefully acknowledge the support of Thompson Rivers University through its seed grant program, which made this research possible.

Conflicts of Interest

The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CPE	Cumulative Probability of Exclusion
GWAS	Genome-Wide Association Study
KNN	K-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
ML	Machine Learning
MSE	Mean Squared Error
PCA	Principal Component Analysis
PCoA	Principal-Coordinate Analysis
rrBLUP	Ridge Regression Best Linear Unbiased Prediction
SNP	Single Nucleotide Polymorphism

References

[1]. Myer, P.R. (2019). Bovine genome-microbiome interactions: metagenomic frontier for the selection of efficient productivity in cattle systems. *Msystems, 4*(3), 10–1128.

[2]. Porto-Neto, L.R., Kijas, J.W., & Reverter, A. (2014). The extent of linkage disequilibrium in beef cattle breeds using high-density snp genotypes. *Genetics Selection Evolution, 46*, 1–5

[3]. Uzzaman, M.R., Edea, Z., Bhuiyan, M.S.A., Walker, J., Bhuiyan, A., & Kim, K.S. (2014). Genome-wide Single Nucleotide Polymorphism Analyses Reveal Genetic Diversity and Structure of Wild and Domestic Cattle in Bangladesh. *Asian-Australasian Journal of Animal Sciences*, *27*(10), 1381.

[4]. Hu, L., Li, D., Chu, Q., Wang, Y., Zhou, L., Yu, Y., Zhang, Y., Zhang, S., Usman, T., Xie, Z., et al. (2021). Selection and implementation of single nucleotide polymorphism markers for parentage analysis in crossbred cattle population. *Animal*, *15*(1), 100066.

[5]. Mancin, E., Tuliozi, B., Pegolo, S., Sartori, C., & Mantovani, R. (2022). Genome wide association study of beef traits in local alpine breed reveals the diversity of the pathways involved and the role of time stratification. *Frontiers in Genetics, 12*, 746665.

[6]. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

[7]. Puthiyedth, N., Zhang, N., Wang, Z., & Yan, Y. (2021). Performance comparison of lasso variants with genome-wide association studies (gwas). In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1682–1684. IEEE.

[8]. Anupama, C., Puthiyedth, N., & Neenu, R. (2019). Feature selection methods for snp analysis. In *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*. vol. 1, pp. 87–93. IEEE.

[9]. Hilt, D.E., & Seegrist, D.W. (1977). *Ridge, a computer program for calculating ridge regression estimates.* Department of Agriculture, Forest Service, Northeastern Forest Experiment Station.

[10]. Abo-Ismail, M.K., Vander Voort, G., Squires, J.J., Swanson, K.C., Mandell, I.B., Liao, X., Stothard, P., Moore, S., Plastow, G., & Miller, S.P. (2014). Single nucleotide polymorphisms for feed efficiency and performance in crossbred beef cattle. *BMC genetics*, *15*, 1–14.

[11]. Wallace, R.J., Sasson, G., Garnsworthy, P.C., Tapio, I., Gregson, E., Bani, P., Huhtanen, P., Bayat, A.R., Strozzi, F., Biscarini, F., et al. (2019). A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. *Science Advances*, *5*(7), eaav8391.

[12]. Fonseca, P., Lam, S., Chen, Y., Waters, S., Guan, L., & Cánovas, A. (2023). Multi-breed host rumen epithelium transcriptome and microbiome associations and their relationship with beef cattle feed efficiency. *Scientific Reports, 13*(1), 16209.

[13]. Li, F., & Guan, L.L. (2017). Metatranscriptomic profiling reveals linkages between the active rumen microbiome and feed efficiency in beef cattle. *Applied and environmental microbiology 83*(9), e00061–17.

[14]. Cholewińska, P., Górniak, W., & Wojnarowski, K. (2021). Impact of selected environmental factors on microbiome of the digestive tract of ruminants. *BMC Veterinary Research*, *17*, 1–10.

[15]. Li, F., Li, C., Chen, Y., Liu, J., Zhang, C., Irving, B., Fitzsimmons, C., Plastow, G., & Guan, L.L. (2019). Host genetics influence the rumen microbiota and heritable rumen microbial features associate with feed efficiency in cattle. *Microbiome*, *7*(1), 92.

[16]. Zhang, Q., Difford, G., Sahana, G., Løvendahl, P., Lassen, J., Lund, M.S., Guldbrandtsen, B., & Janss, L. (2020). Bayesian modeling reveals host genetics associated with rumen microbiota jointly influence methane emission in dairy cows. *The ISME Journal*, *14*(8), 2019‑2033.

[17]. Puthiyedth, N., Zeinalinesaz, F., Hou, D., Zhang, Y., Lin, W., & Yan, Y. (2025). Leveraging LASSO-based methodologies for enhanced SNP analysis in plant genomes. *Bioinformatics Advances, 5*(1), vbaf014. https://doi.org/10.1093/bioadv/vbaf014

[18]. Jemaa, S.B., Boussaha, M., Mehdi, M.B., Lee, J.H., & Lee, S.H. (2015). Genome-wide insights into population structure and genetic history of Tunisian local cattle using the illumina bovinesnp50 beadchip. *BMC Genomics*, *16*(1), 677.

[19]. Gautier, M., Laloë, D., & Moazami-Goudarzi, K. (2010). Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. *PloS one*, *5*(9), e13038.

[20]. Gautier, M., Flori, L., Riebler, A., Jaffrézic, F., Laloé, D., Gut, I., ... & Foulley, J. L. (2009). A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC genomics*, *10*(1), 550.

[21]. Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P., Sonstegard, T.S., et al. (2009). Development and characterization of a high density snp genotyping assay for cattle. *PLoS one*, *4*(4), e5350.