Elucidating the multiple genetic lineages using SNP data in livestock

Marcos Vinicius Gualberto Barbosa da Silva, Saulo Moraes Villela, Luiz Afonso Glatzl Junior

UNIVERSIDADE FEDERAL DE JUIZ DE FORA

Abstract

Differences observed among populations of the same species regarding a trait can be attributed to differences in the environments in which they are found or to genetic differences among them. When seeking greater efficiency in animal production, there are two paths. The first one is to make improvements in general management, and the second one, slower and with a permanent and cumulative character, is a genetic improvement, carried out through the selection of desirable phenotypes and mating. In traditional animal breeding, the additive genetic value derived from the phenotype is estimated. With the advancement of Biotechnology and Genomics, the use of molecular markers such as SNP (Single Nucleotide Polymorphism) has resulted in more accurate selection methodologies. Relationship presupposes similarity of genotypes and its measurement is fundamental in the correct use of the selection and mating tools. The main purpose of models based on machine learning is to obtain generic conclusions through a particular dataset. With supervised learning, classifiers can be generated using data previously labeled for adjustment. As an alternative to traditional animal relationship identification techniques, this work aims to develop a lineage classifier based only on the animal's molecular markers. The dataset used in this study is composed of the genotype of 14, 242 Gir Leiteiro animals, allocated to the GGP Indicus SNP Chip, considering only markers from autosomal chromosomes, resulting in 33, 336 SNPs. Initially, quality control was performed using filters for minor allele frequency and call rate, resulting in 1, 380 markers. The five lineages with the largest number of individuals genotyped were identified, resulting in 1, 061 animals. In model development, each SNP was coded using one-hot-encoding and the dataset was split into training and test sets in the proportion of 70% and 30%. The training set was balanced through oversampling, being subdivided into 5 parts to perform a cross-validation. Four different machine learning models were adjusted, k-Nearest Neighbors (k-NN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Random Forest (RF). The metrics used in the model selection were the accuracy and standard deviation of the test sets, and the area under the ROC curve. The best performance was obtained through SVM. From the developed work, it was possible to verify the importance of evaluating new models in animal breeding problems, which helps the development of methodologies with less computational effort and greater accuracy results.

Funding:

Link to Video: