

Augmenting Association Analysis in Alfalfa Biomass Yield & Forage Quality by Multivariate Model & Machine Learning

Amna Sajjad, New Mexico State University

Harpreet Kaur, University of Minnesota

Donovan Bailey, New Mexico State University

Christopher Pierce, New Mexico State University

Ian Ray, New Mexico State University

Mohammad Abdullah Bari, New Mexico State University

Alfalfa (*Medicago sativa* L.) is the most important forage legume where improving biomass yield and forage quality traits are critical for sustainable and profitable livestock production. These economically important traits are complex, often quantitatively inherited, and can be highly correlated, making their genetic dissection challenging. Genome-wide association studies (GWAS) provide a powerful tool to identify loci associated with traits of interest in unrelated populations by leveraging historical recombination and linkage disequilibrium decay, resulting in greater mapping resolution compared to traditional biparental QTL mapping. In this study, we analyzed 215 parental genotypes using 11K genotyping-by-sequencing (GBS) SNP markers from the New Mexico Genomic Study (NMGS). Their half-sib families were evaluated under field conditions for 15 forage nutritional traits under deficit irrigation management, Las Cruces, New Mexico, during 2018 and 2019. Population structure and relatedness were accounted for using principal component analysis and mixed-model approaches to control spurious associations. Single-trait GWAS implemented in GWASpoly identified 6 significant SNP markers associated with 7 forage quality and nutritional traits, including adjusted neutral detergent fiber, neutral detergent fiber, lignin, and crude protein among others.

However, conventional GWAS methods typically analyze one trait at a time, potentially limiting the detection of shared genetic architecture and pleiotropic loci. To better capture the shared genetic architecture of correlated traits, we propose a multivariate GWAS framework using the Genome-wide Efficient Mixed Model Association algorithm, which jointly models multiple phenotypes while accounting for population structure and kinship. This approach enhances statistical power and facilitates the detection of pleiotropic loci influencing multiple agronomic traits. Machine learning models are integrated to complement GWAS for marker prioritization and precise detection. Random Forest is used to estimate marker importance, Least Absolute Shrinkage and Selection Operator (LASSO) enables sparse feature selection, and eXtreme Gradient Boosting (XGBoost) captures complex nonlinear relationships and interactions among loci that may not be detected through conventional GWAS. The proposed integrative framework, combining GWAS, multivariate modeling, and machine learning, can elucidate the genetic architecture of biomass yield and forage quality in alfalfa. The results are expected to enhance marker discovery, underlying putative genes and functions to accelerate the development of superior cultivars through informative marker-assisted breeding.

Keywords: Alfalfa, Genome Wide Association Studies (GWAS), Forage Biomass Yield, Forage Nutritional Quality, Multivariate GWAS, Machine Learning enhanced GWAS