Utilizing Evolutionary Conservation Information to Improve Prediction Accuracy in Genomic Selection

Jinliang Yang[1], Sofiane Mezmouk[1,2], Rita Mumm[3], and Jeffrey Ross-Ibarra[1]

[1] *Department of Plant Sciences, University of California, Davis, CA 95616, USA*
[2] *Current address: KWS SAAT AG, Grimsehlstr. 31, 37555 Einbeck, Germany*
[3] *Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

ABSTRACT

Genomic selection (GS) has gained popularity recently as the availability of genome-wide markers has increased. Current methods for GS weigh all the available SNPs equally in model training, without considering their functional differences. Genetic variations detected at evolutionary conserved sites tend to be deleterious and, thus, may be more informative for GS. To utilize this kind of information as a prior into the GS model, we proposed a method to put more weight on evolutionarily constrained sites. As a proof-of-concept, a half diallel population based on 12 diverse inbred lines was used, from which seven phenotypic traits were collected. Some of these traits show high levels of heterosis. After sequencing the 12 founder lines, about 14 million SNPs were discovered and the SNPs were used to identify 502,913 haplotype blocks shared through identity by descent (IBD). A five fold cross-validation experiment was conducted using the summary statistics of the SNP conservation scores, which were computed by evaluating sequences similarity of multiple divergent species, in the IBD blocks as explanatory variables. The results show that the prediction accuracies are significantly better than shuffled data with randomly assigned conservation scores. This study demonstrates the importance of incorporating evolutionary information in GS and its potential use in plant breeding.