Genomics-based prediction of performance of quantitative traits involving epistasis using a nonparametric method

Xiaochun Sun1, Ping Ma2, Rita H. Mumm1*

1 Department of Crop Sciences and the Illinois Plant Breeding Center, 2 Department of Statistics; University of Illinois at Urbana-Champaign; *email: ritamumm@illinois.edu

Genomic selection (GS) methods have proven useful in estimating breeding value based on dense, genome-wide molecular marker information to guide choice of parents and predict phenotype (Albrecht et al. 2011; Bernardo and Yu 2007; Jannink et al. 2010; Zhong et al. 2009). Most of the GS methods are implemented with parametric models; however, these lack the ability to represent and integrate epistasis properly due to limited understanding the underlying biological process of epistasis and difficulties in statistically decomposing epistatic variance. Furthermore, issues of high dimensionality and multicollinearity inherent to dense, genome-wide marker data must be overcome to increase accuracy and enhance predictive ability. We, therefore, have proposed a new nonparametric method, pRKHS, which combines the features of supervised principal component analysis (SPCA) (Bair et al. 2006) and reproducing kernel Hilbert spaces (RKHS) regression (Gu 2002; Wahba 1990), with versions for traits with no/low epistasis, pRKHS-NE, to high epistasis, pRKHS-E. pRKHS-NE accounts for only additive effects, and pRKHS-E includes additive-by-additive interaction effects as well as additive effects in the model. Instead of assigning a specific relationship to explain the underlying epistatic process, these approaches map genotype-to-phenotype in a nonparametric way, requiring few genetic assumptions. With SPCA, principal components are computed from a reduced marker matrix after 'non-significant' markers (markers with zero or near-zero effect) have been filtered out. The criterion of 'significance' is determined by cross-validation. Employing RKHS regression, these principal components are then included in the smoothing spline ANOVA model as independent variables to fit the data. The new method was evaluated in comparison with current popular methods for practicing GS, specifically RR-BLUP (Whittaker et al. 2000), BayesA and BayesB (Meuwissen et al. 2001), using both simulated and real data. Results demonstrate that pRKHS delivers greater predictive ability, particularly when epistasis impacts trait expression. Beyond prediction, the new method can also be used to facilitate inferences about the extent to which epistasis influences trait expression.