Genomewide prediction accuracy within 1000 biparental maize populations

Genomewide selection was first brought up by Meuwissen (Meuwissen et al., 2001) and was reported in a simulated experiment to have much higher accuracy of prediction for the breeding value than traditional least square method, which only fits the significant markers. The expected correlation ($r_{MG}$) between predicted genotypic values and true genotypic values has been derived by Daetwyler et al. as a function of the training population size ($N$), trait heritability ($h^2$), and effective number of quantitative trait loci (QTL) or chromosome segments affecting the trait ($M_e$) (Daetwyler et al., 2008). The concept of genomewide selection has been tested in many animal-breeding experiments. The training population size in these animal breeding experiments ranged from 650-4500 and all animals were genotyped for approximately 50000 markers. The accuracies of genomewide selection for these experiments ranged from 0.4 to 0.82 (Hayes et al., 2009; VanRaden et al., 2009; Harris et al., 2009). However, prediction accuracy relies heavily on the relationship between training and validation population. Prediction accuracy ranged from 0.1 to 0.3 when SNP effects from one cattle breed to predict the breeding value of another cattle breed (Harris et al., 2009). Goddard et al reported that when predict Jersey breed using marker effects estimated from Holstein breed, the prediction accuracy was very low (from -.06 to 0.23) and vice versa (Hayes et al., 2009). Bernardo et al. began to introduce genomewide selection into plant breeding in 2007 by simulating a biparental population and the response to selection with genomewide selection was consistently higher than the response to selection with marker assisted recurrent selection (Bernardo & Yu, 2007). Empirical data in genomewide selection began to be published in plants in 2009 (Lorenzana & Bernardo, 2009) with seven different biparental populations in three plant species; The prediction accuracy with the best marker-training population design in those different trait-population combinations were mostly in the range of 0.5 to 0.8. Subsequent empirical studies on genomewide selection have been published in plants and most of those studies were designed with the training population and validation population derived from the same diverse panel of inbred lines (Asoro et al., 2011; Heffner et al., 2011; Crossa et al., 2010; Poland et al., 2012; Rutkoski et al., 2012; Lorenz et al., 2012). Eeuwijk used a training population of 1700 maize hybrid to predict the ear height of a set of 288 new combination of hybrids and get a prediction accuracy of 0.69 (Eeuwijk et al., 2010). Albrecht analyzed the prediction accuracy of 1380 maize DH lines from 36 crosses but analyzed the data in the same way as if the training population and validation population were from a diverse inbred panel and no inference was made toward the prediction of a specific cross (Albrecht et al., 2011). Windhausen used the test cross performance of a diverse maize inbred lines to predict the test cross performance of five
biparental F2 derived lines and the prediction accuracy was almost zero (Windhausen et al., 2012).

The published biparental data were with only seven populations (Lorenzana & Bernardo, 2009) and cannot be easily extrapolated to a wide range of populations in true breeding programs. Most of the empirical studies with a diverse panel of inbred lines are not designed in the way that the results can be directly used in a routine breeding program. We have 1000 biparental maize populations from Monsanto’s breeding program, which made the evaluation of genomewide selection in a large breeding program available. My first research project was to evaluate the accuracy of genomewide selection within 1000 biparental maize breeding populations and compare the observed accuracy with the expected \( r_{MG} \). Genomewide prediction was by ridge regression-best linear unbiased prediction and the observed \( r_{MG} \) was obtained by dividing the correlation between marker-predicted values and phenotypic values by the square root of \( h^2 \). We also modified Daetwyler et al.’s equation for expected \( r_{MG} \) to account for recombination between a QTL and a marker. Across the 1000 populations, the mean and range (in parentheses) of observed \( r_{MG} \) was 0.24 (–0.32, 0.77) for grain yield, 0.41 (–0.23, 0.90) for grain moisture, and 0.38 (–0.28, 0.95) for test weight. The \( h^2 \) and N had the largest influence on observed \( r_{MG} \). The expected \( r_{MG} \) from Daetwyler et al. was much higher than the observed \( r_{MG} \). In contrast, the observed \( r_{MG} \) values were centered around the expected \( r_{MG} \) based on imperfect linkage between a QTL and marker, although the spread of the observed \( r_{MG} \) was still very large. We conclude that predicting the accuracy of genomewide predictions is difficult.

The second part of my research project will be studying how to use closely related crosses to predict the performance of progenies in a new cross, so that selection can be done within that new cross without the need of phenotyping. Population will be divided into different subpopulations by clustering analysis on their genetic relationship matrix. Subpopulation with different number of crosses and different level of relatedness to the target cross will be set up. The effect of different subpopulation settings on the prediction accuracy and on the genetic diversity among the selected crosses will be evaluated.


