Accuracy Of Genome Wide Selection For Inbred Selection Based On Hybrid Performance

F. Pita[*], K. Robbins[*] and J.E. Backlund[*]

Introduction

Genome Wide Selection (GWS) is the use of a high density marker panel to predict the value of an individual in regard its additive genes. The rationality is that such large number of markers will be closely linked to all the Quantitative Trait Loci (QTL) affecting a given phenotypic trait. The summation of each marker effect in an individual is its Genomic Expected Breeding Value (GEBV). The rank of candidates to selection based on their GEBV can be used to choose individuals that have the best set of genes with additive effect for a given trait. Much research has been conducted to investigate the application of Genome Wide Selection using a bayesian framework in closed outbred populations. The number of individuals with both phenotype and marker information and the amount of linkage disequilibrium between markers and QTL have been shown as the main factors for the efficiency of GWS (Goddard and Hayes (2007)). Fernando and Stricker (2008) calculated the accuracy of predicting breeding values based only on marker information of direct descendents of 2,120 animals with phenotypes and scored for 3,000, 30,000, or 60,000 markers, and found those to be between 0.82 and 0.88. Many breeding designs, in special for plant species, aim to maximize heterosis in the commercial product by developing highly inbred breeding populations. The objective of this work was to evaluate, through computer simulation, the accuracy of GWS for inbred selection based on commercial hybrid performance.

Material and methods

Simulation. Two populations (A and B) were simulated with 100 individuals each and separated from each other by 200 generations. The genome of each individual had 10 chromosomes, different number of SNP markers, and 200 QTL affecting its phenotype. Both the SNP and QTL were randomly assigned throughout the genome, i.e., the distance between SNP and QTL varied. QTL effects were sampled from a Gamma distribution and their effects scaled to produce an additive genetic variance of 1.0. An environment effect was sampled from a normal distribution with mean zero and variance 1.0. The phenotype of the individuals was simulated as the summation of the 200 QTL effects plus an environment effect. Therefore, the heritability in a narrow sense (h^2) of the simulated trait was 0.50. Individuals in each population were randomly crossed with each other and their offspring were selfed for 8 generations to produce inbred populations A0 and B0. Individuals A0 were randomly crossed to produce F1's which in turn were selfed for 8 generations to produce inbred individuals in a process similar to the one used to develop inbred lines in plant

[*] DOW AGROSCIENCES, LLC - Indianapolis, Indiana USA

species. These inbred individuals were called A1. Using the same process, individuals A1 produced next generation of inbreds (A2) and individuals A2 produced another inbred generation (A3). In order to investigate the effect of the training dataset size and structure on GEBV accuracy, different number of individuals in distinct inbred generations (50, 100, or 200 individuals in generations A0 or A1) were crossed with individuals from population B0 (25 or 50 individuals) to produce hybrids with phenotypic data (Table 1). In order to investigate the effect of the marker panel density on GEBV accuracy, these hybrids were scored for a panel of SNP markers distributed throughout the simulated genome at an average 2.0 cM (centimorgan), 0.5 cM, or 0.25 cM. A validation dataset of 2,000 individuals without phenotypic records was produced in generations A1 and A3.

Genome Wide Selection predictions. GEBV were obtained using both phenotypic and marker information simulated in the training datasets using Bayes B (Meuwissen, Hayes and Goddard (2001)). The Markov Chain Monte Carlo algorithm implemented in Bayes B was run for 10,000 iterations. The efficiency of GWS for the different investigated scenarios was evaluated by calculating the accuracy of the GEBV defined as the correlation between calculated GEBV and the simulated GEBV for those individuals in the validation datasets. The accuracy results reported in this work are the average of 10 repetitions of each combination of training dataset size, marker panel density, and generation of the validation dataset.

Results and discussion

Table 1 shows the average GEBV accuracy over 10 repetitions for different investigated scenarios. When the training dataset was generated using parents from A0 generation, GEBV accuracy for the second generation of inbreds (A1) was between 0.70 and 0.92 with higher accuracies (9.1% higher in average) observed with larger number of A0 parents (100 versus 50). This may be related with better opportunity to estimate marker effects in population A using a more representative sample of individuals. On the other hand, the number of A0 parents that originated the training dataset had little effect on the GEBV accuracy for generation A3 (2.5% higher accuracy with 100 vs. 50 A0 parents and 0.25 cM between markers). Regarding the marker density effect, an average 19.05% and 142.3% higher GEBV accuracy was observed for generation A1 and A3, respectively, when marker density increased from one marker every 2 cM to one marker every 0.25cM. While the benefit of higher marker density in generation A1 is not evident, the GEBV accuracy in the later generation (A3) was low with small marker density and reasonable accuracies were only obtaining with the higher marker density (average accuracy of 0.83). It is because having markers closer to the QTL reduces the chance of recombination between them when a new generation is produced, and this benefit is higher in later generations. When the training dataset was produced from the cross between inbreds in generation A1 and B0 inbreds, the accuracy of GEBV for A1 contemporary inbreds was not higher than when the training dataset was obtained with A0 generation parents, in special when only 50 parents from population A were used. Therefore re-training the GWS model is not beneficial. This may due to the fact that most of the recombinations between inbreds and during the selfing process are ineffective and therefore few A1 individuals have portions of DNA of the recom-

Table 1: Accuracy of Genomic Expected Breeding Values for different scenarios investigates.

Training data[1]	Validation data (2,000 individuals)	SNP density[2]	Average accuracy (std)
50 A0 x 50 B0	A1	2.0	0.70 (0.06)
		0.5	0.79 (0.05)
		0.25	0.83 (0.05)
	A3	2.0	0.47 (0.24)
		0.5	0.60 (0.16)
		0.25	0.82 (0.06)
100 A0 x 25 B0	A1	2.0	0.77 (0.07)
		0.5	0.84 (0.06)
		0.25	0.92 (0.03)
	A3	2.0	0.27 (0.16)
		0.5	0.64 (0.15)
		0.25	0.84 (0.05)
50 A1 x 25 B0	A1	2.0	0.46 (0.09)
		0.5	0.63 (0.07)
		0.25	0.68 (0.11)
	A3	2.0	0.49 (0.22)
		0.5	0.63 (0.14)
		0.25	0.66 (0.16)
100 A1 x 25 B0	A1	2.0	0.63 (0.11)
		0.5	0.73 (0.09)
		0.25	0.85 (0.04)
	A3	2.0	0.63 (0.08)
		0.5	0.78 (0.08)
		0.25	0.90 (0.06)
200 A1 x 25 B0	A1	2.0	0.66 (0.09)
		0.5	0.89 (0.05)
		0.25	0.93 (0.04)
	A3	2.0	0.70 (0.13)
		0.5	0.87 (0.05)
		0.25	0.93 (0.05)

[1] number individuals in each generation

[2] average distance between markers (cM)

binant parental type. When the GWS training uses A1 parents, phenotypic and marker data from these few ones will produce more conservative marker effects due to different marker-QTL phase compared to most A1 parents. As a result, the GEBV accuracy will be smaller for most contemporaries who have the non-recombinant parental type. Higher number of A1 parents in the training dataset and higher marker density would avoid this issue.

Conclusion

These results show that Genome Wide Selection can be an effective tool for genetic improvement of highly inbred populations where breeding designs aim to explore heterosis. Structure and size of the training dataset and marker density are determinant for GWS success, even though the number of markers necessary to carry out GWS in these populations is smaller than those typically necessary for outbred populations.

References

Fernando, R.L., Stricker, C. *J. Anim. Sci.* 86: E- Suppl. 2. 103. 2008

Goddard, M.E, Hayes, B.J. *J. Anim. Breed. Genet.* 124:323-330. 2007

Meuwissen, T.H.E., Hayes, B.J, Goddard, M.E. *Genetics*. 157:1819-1829. 2001