Chapter 45
Genomic Selection in Plants: Empirical Results and Implications for Wheat Breeding

Mark E. Sorrells

The opinions expressed and arguments employed in this publication are the sole responsibility of the authors and do not necessarily reflect those of the OECD or of the governments of its Member countries.

The Special Session was sponsored by the OECD Co-operative Research Programme on Biological Resource Management for Sustainable Agricultural Systems, whose financial support made it possible for most of the invited speakers to participate in the Special Session.

Abstract Genotyping-by-sequencing technology is rapidly reducing marker costs and increasing genome coverage allowing the widespread use of molecular markers and methods in plant breeding. Marker assisted selection (MAS) and recurrent selection are based on the selection of statistically significant, marker-trait associations. However, MAS strategies are not well suited for complex traits controlled by many genes. Genomic selection (GS) incorporates genome-wide marker information in a breeding value prediction model, thereby minimizing biased marker effect estimates and capturing more of the variation due to small effect QTL. In GS, a training population related to the breeding germplasm is genotyped with genome-wide markers and phenotyped in a target set of environments. That data is used to train a prediction model that is used to estimate the breeding values of lines in a population using only the marker scores. Prediction models can incorporate performance over multiple environments and assess G x E effects to identify a highly predictive subset of environments. Because of reduced selection cycle time, annual genetic gain for GS is predicted to be two to threefold greater than for a conven-

M.E. Sorrells (✉)
Department of Plant Breeding & Genetics, Cornell University, Ithaca, NY 14853, USA
e-mail: mes12@cornell.edu

Y. Ogihara et al. (eds.), *Advances in Wheat Genetics: From Genome to Field*, DOI 10.1007/978-4-431-55675-6_45

tional phenotypic selection program. We have developed a new methodology for using genome-wide marker effects to group environments and identify outliers. In addition, environmental covariates can be identified that increase prediction accuracy and facilitate performance prediction in climate change scenarios. This new approach to crop improvement will facilitate a better understanding of the dynamic genome processes that generate and maintain new genetic variation.

Keywords Breeding methods • Genomic selection • Genotype by environment interaction • Marker-assisted selection • Molecular markers • Wheat

This chapter is a review of my presentation at the International Wheat Genetics Symposium held in Yokohama, Japan September 13, 2013.

Plant breeding is a predictive science. We are constantly trying to predict the performance of selected genotypes. Our evaluation methods are designed to predict traits such as grain yield, milling and baking quality, disease resistance. Prediction accuracy is important at every step in the breeding program and breeding methods are designed to improve accuracy of those predictions. Traditionally, breeding methods such as family-based selection, progeny testing, and more recently, the use of molecular markers has enabled marker-assisted selection, and genomic selection (GS). Novel breeding strategies are driven by technology and new knowledge.

Our molecular breeding goals include allele discovery, allele characterization and validation, and parent and progeny selection for superior alleles at multiple loci to generate transgressive segregation. We can increase the annual rate of genetic gain in several ways. The selection intensity can be increased or we can increase the heritability by improving the accuracy of selection or increasing the genetic variation. However the selection cycle time has the greatest impact on annual rate of gain so breeding methods that reduce the cycle time have the most impact.

Meuwissen, Hayes and Goddard first proposed genomic selection methodology in 2001 (Meuwissen et al. 2001). This methodology consists of two distinct populations, a training population that is genotyped with a large number of markers and phenotyped for important traits and a breeding population consisting of individuals that are genotyped but not phenotyped (Fig. 45.1). The training population consists of well-adapted breeding lines and varieties that are phenotyped in the target population of environments. Genome-wide markers are considered to be random effects and all marker effects on the phenotype are estimated simultaneously in a single model. One or more markers are assumed to be in linkage disequilibrium (LD) with each quantitative trait locus (QTL) affecting the trait of interest. We use a prediction model that attempts to capture the total additive genetic variance to estimate breeding value of individuals based on the sum of all marker effects. In the breeding population, genomic estimated breeding values (GEBVs) for each individual are obtained by summing the marker effects for that genotype. The prediction model can then be used to impose multiple generations of selection. The selected individuals can be recycled in the crossing program and/or evaluated in advanced replicated

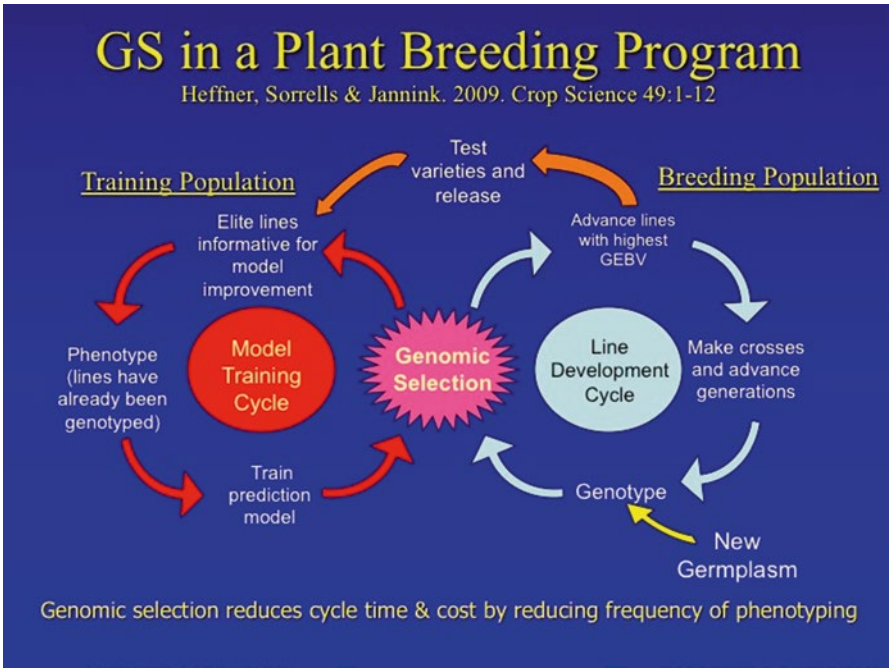

Fig. 45.1 Genomic selection training and breeding populations and their interactions in a plant breeding program

trials. Eventually the best lines are added to the training population and the prediction model is updated.

There are several factors that affect the accuracy of GEBVs. Because it is assumed that there is at least one marker in LD with each locus affecting the trait of interest, the level and distribution of LD between markers and QTL impact the accuracy of GEBVs. Using simulations, Meuwissen (2009) estimated that the minimum number of markers for across family predictions would be Ne*L where Ne is the effective population size and L is the genome size in Morgans. For example, wheat has a genome size of about 30 Morgans and if we assume an effective population size of 50, that would indicate that the minimum number of markers required would be 1,500. The size of the training population and its relationship to the breeding population are also important, and over time, re-training models is required. Meuwissen (2009) also estimated the minimum number of records for across family predictions would be 2*Ne*L. For wheat, that would be a population size of about 3,000. However, good accuracies have been reported for populations much smaller than that. The breeding population must be closely related to the training population for accuracy predictions. Population substructure in the training population can inflate accuracies and lead to inbreeding. Many small effect QTL or low LD favor Best Linear Unbiased prediction (BLUP) for capturing small effect QTL that may

not be in LD with a marker. More records are required for low heritability traits, just as for phenotypic selection.

Genomic selection was first developed and evaluated by the dairy industry. However plants have different constraints as well as some advantages. Obviously, mating schemes are quite different, even among different plant species. Animal parental values are mainly based on half-sib families with a sire in common among the progeny. Population size in animals is based on accumulating records over time for many families within a breed. For most crop species large populations can be easily generated either using biparental crosses or the progeny from multiple families. The ability to replicate is an important factor as well. Inbred lines, testcross hybrids, and clonally propagated crops can be replicated in time and space, while each animal is a unique genotype and heterozygous. Finally, genotype by environment interaction (GxE) is a more important issue in plant breeding than in animal breeding.

Next I would like to present some of the results of GS experiments from the Cornell wheat breeding program. Elliot Heffner was a Ph.D. student in my program when we initiated our GS research in 2007. He conducted experiments on GS in both biparental populations (Heffner et al. 2011b) and across multiple families in our breeding program (Heffner et al. 2011a). Jessica Rutkoski, a former Ph.D. student in our program has published a review on Genomic Selection for Adult Plant Stem Rust Resistance (Rutkoski et al. 2010), a study on GS for fusarium head blight resistance (Rutkoski et al. 2012), and methods for imputing missing data without ordered markers (Rutkoski et al. 2013).

Heffner et al. (2011b) used two doubled haploid biparental populations to evaluate GS for nine milling and baking quality traits tested over 3 years. The results averaged over both populations showed that the GS prediction models were 47 % more accurate than the multiple linear regression (MLR) model (Fig. 45.2). For the experiment involving multiple families, the training population consisted of 400 advanced breeding lines planted in an augmented field design in three locations over 3 years. It was genotyped using 1,500 polymorphic DArT markers and phenotyped for 13 agronomic traits. Prediction models included two multiple linear regression models, with or without the Kinship Matrix as a covariate, and four GS models, ridge regression, Bayes A, Bayes B, and Bayes C pi. The MLR model accuracy was similar with or without the Kinship matrix. GS accuracy was similar for all prediction models (~0.60). The GS prediction accuracy was 25 % greater than for MLR and phenotypic selection accuracy was 7 % greater than for GS. A comparison of eight of the individual traits reveals that the relative accuracy of GS compared to phenotypic selection is highest for the traits with the lowest heritability (Fig. 45.3). This highlights an important feature of GS. It is complementary to MAS because MAS is most effective for simply inherited, high heritability traits whereas GS is relatively more effective for low heritability traits. It is this complementarity that facilitates incorporation of GS into a molecular breeding program. To take advantage of the features of GS that increase annual genetic gain, a recurrent genomic selection program could be used to generate multiple cycles of selection per year (Fig. 45.4). In addition, once selections are inbred, whole-genome genotypes can be used to further select individuals with the highest GEBVs. For the

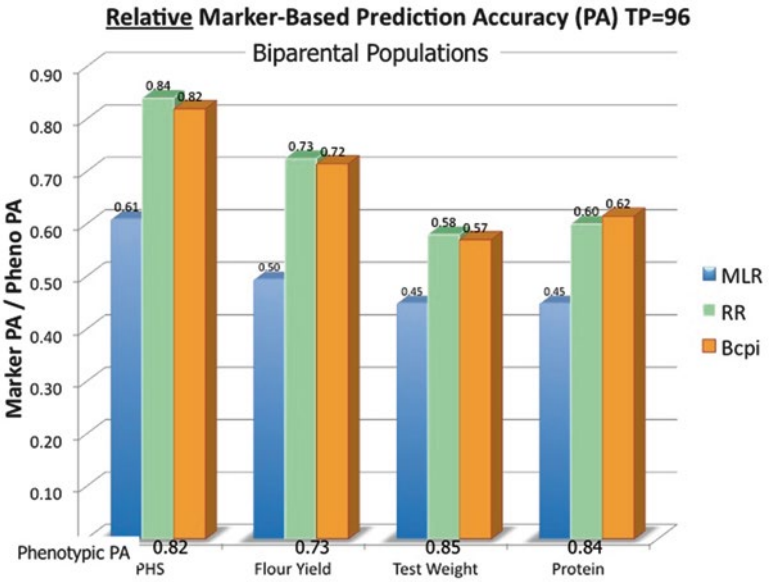

Fig. 45.2 Relative marker-based prediction accuracy (PA) for four quality traits in two biparental wheat populations. *MLR* multiple linear regression, *RR* ridge regression, *Bcpi* Bayes C pi

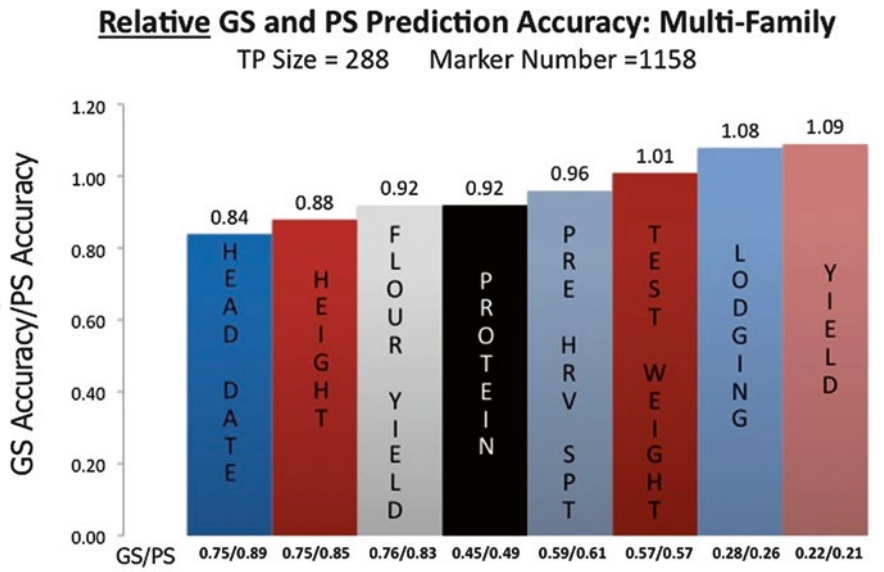

Fig. 45.3 Relative accuracy of GS compared to phenotypic selection (PS) for eight traits of varying heritability

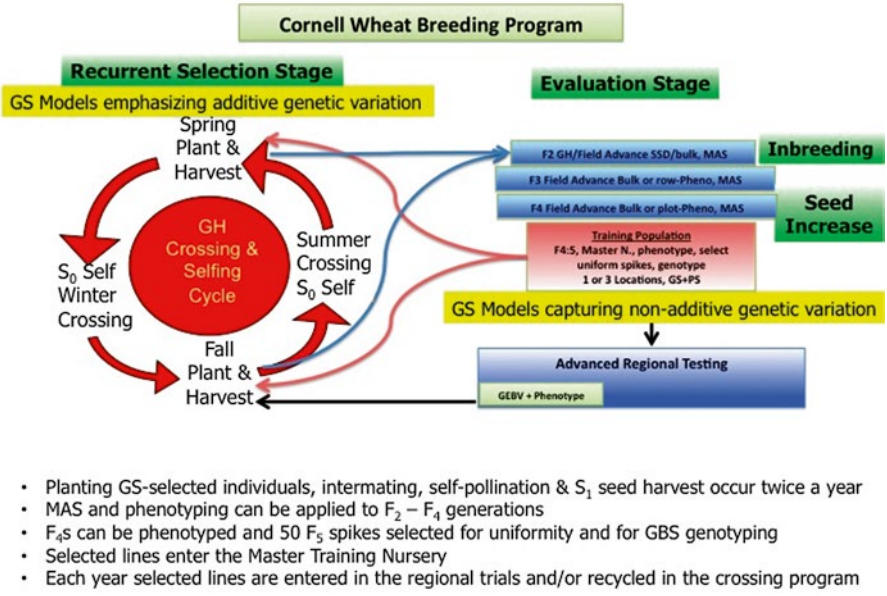

Fig. 45.4 Integration of GS in a wheat breeding program

recurrent genomic selection, GS models based on additive genetic effects would be appropriate, whereas for selecting purelines, models that incorporate non-additive effects may be more accurate.

There are several factors that need to be considered when initiating GS. Models for biparental populations are population specific but yield a higher accuracy. There is reduced epistasis and allele frequencies are balanced. Fewer markers and smaller training populations are required. Biparental populations are probably the only option for introgression of exotic germplasm. In contrast, multifamily GS allows prediction across a broader range of adapted germplasm, sampling of more environments, and larger training populations. In addition, cycle duration is reduced because model retraining is on-going.

Whole-genome genotyping opens up new opportunities for analyzing breeding trial data. For example, marker effects instead of genotypes can be used to increase GS prediction accuracy (Heslot et al. 2013). Advanced breeding trial data are typically unbalanced, i.e., all genotypes are not evaluated in all environments and this limits the kinds of analyses that can be used. However, if we use marker alleles instead, the data are balanced because all markers are evaluated in all environments. Heslot et al. (2013) used marker effects to identify outlier environments, classify relevant mega-environments, and to select an optimum subset of environments for GS prediction. Marker effects for each environment were calculated using the Bayesian LASSO GS model. Nearly 1,000 barley advanced lines were evaluated for grain yield in 58 European environments. The dataset was unbalanced with only 18

genotypes present in >50 % of the environments. Environment groups were based on Additive Main Effects & Multiplicative Interaction (AMMI) analysis, year. region, marker effects, and pairwise prediction accuracy between environments. Marker effects for all lines in each environment formed a balanced dataset for computing Euclidean distances between environments. Only clustering based on average reciprocal prediction accuracies significantly increased prediction accuracy. Clustering based on marker effects produced four clear subgroups that were not related to region or year, but also did not increase prediction accuracy. Reciprocal accuracies correlate with genetic correlations between environments based on a factor analytic model and were useful to measure genetic correlation without the numerical issues of factor analytic models. In a second experiment, Heslot et al. (2013) developed a protocol using the predictive ability of an environment for optimizing the composition of the training population. The procedure involves training a GS model in each environment and computing the mean accuracy for each training environment for predicting line performance in the other environments. They are then ranked and environments are removed one at a time starting from the least predictive.

The GS model is then trained and cross validated on the remaining training population and is referred to as the "Predictive set". The removed environments are referred to as the "Unpredictive set" and accuracy is predicted using the same GS model. Both accuracy measurements are used to decide the cut-off point for the optimum set. Using this procedure prediction accuracy rose from 0.54 to 0.61 with no change in heritability. Some outlier environments were included in the optimal model. Although it was not statistically significant, accuracy in the validation set increased from 0.279 to 0.292.

Probably the most important project we have ongoing is the merger of crop modeling methodologies integrating environmental covariates and crop modeling into the genomic selection framework to predict G*E (Heslot et al. 2014). Crop modeling has the goal of assessing the impacts of climate change on productivity and how crops adapt to climate change. Crop modelers calibrate crop models to give reliable predictions under baseline and future climate scenarios with the long-term goal of enhancing world food security and adaptation capacity. By combining crop modeling methods with whole-genome genotyping in a GS framework, we can predict G*E for unobserved environments, and thus, performance and stability based only on genotype. In addition, we can better understand the characteristics of the target population of environments and determine the genetic architecture controlling G*E. In this experiment, we used a dataset consisting of grain yield of 2,437 elite winter wheat lines grown in 44 environments over 6 years in France and genotyped with 1,287 SNP markers. Daily climatic weather data were used to characterize environments. In this study, we extended factorial regression to the GS context and developed a new machine learning approach to capture the non-parametric response of QTL to stresses. This approach was used along with a crop model to enable the use of daily weather data in prediction models. Those G*E predictions could be used to make breeding decisions for specific adaptation. Physiological integration of the environment data involved the use of a crop model (Sirius) to compute the

phenology and synchronize early, medium, and late maturing genotypes to the weather data. Stress covariates by developmental stage were derived by using knowledge about sensitivity of specific growth stages to abiotic stresses that were used as independent covariates in statistical genetic models for effect estimation and prediction.

Model description:

$$y_{ij} = \mu + \beta_j + u_i + \gamma_i f(x_j) + \varepsilon_{ij},$$

where Mu is the intercept, the Beta term is the environment effect, u is a genotype main effect, Gamma is the sensitivity of each genotype to a stress covariate x, that can be transformed by a function *f*(), and Epsilon is the model residual. Crop modeling allows us to leverage agronomy knowledge, reduce dimensionality and nonlinearity, and enables the use of existing breeding data. Performance is predicted as main effect plus G*E deviation and environment clustering based on predicted G*E. Because G*E can be predicted for any genotype in any environment, it is possible to use the table of G*E predictions to cluster environments and investigate the structure of the target population of environments. For this data set, environments were mainly grouped by year but also showed a North/South trend. If we plot the accuracy of the models with and without including G*E, we can get an overall view of the importance of this term (Fig. 45.5). Each dot represents an environment, and a dot above the line indicates higher accuracy with the model that included G*E. Overall, there was an 11.1 % increase in mean accuracy (P-value 0.02) and a 10.8 % decrease in the accuracy coefficient of variation. It is important to note that the largest gains occurred in environments where accuracies were low. These results are important because, if successful, we will be able to predict GxE for any genotype based only on genotype. This would allow the breeder to select genotypes that

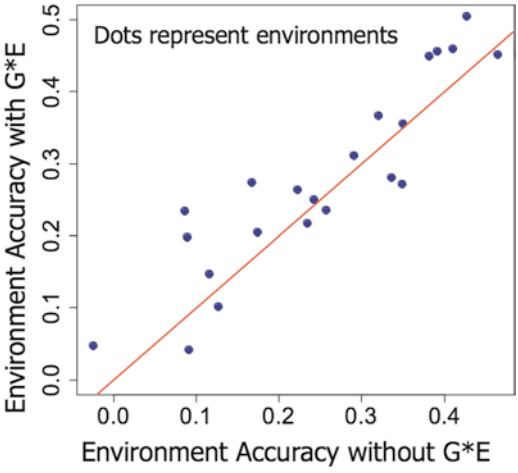

Fig. 45.5 Relative prediction accuracy of prediction models with and without modeling G*E

interact positively with a particular environment, genotypes that minimize GxE, or genotypes that interact positively with environmental factors that limit performance.

In summary, GS differs from MAS and Association Breeding in that the underlying genetic control and biological function is not necessarily known. GS preserves the creative nature of phenotypic selection to sometimes arrive at solutions outside the engineer's scope. Integrating environmental covariates and crop modeling into the genomic selection framework to predict G*E increases prediction accuracy and provides insight into the genetic architecture controlling G*E. The most important advantages are reductions in the length of the selection cycle resulting in greater genetic gain per year.

Acknowledgments I wish to acknowledge the many contributions of current and past students and post docs who have been instrumental in advancing my research program. In particular, my collaborations with Jean-Luc Jannink have been very productive and are much appreciated. Sources of financial support for the research presented include the Bill and Melinda Gates Foundation, U.S. Wheat and Barley Scab Initiative, USDA-NIFA-AFRI grants, award numbers 2009-65300-05661, 2011-68002-30029, and 2005-05130, Hatch project 149-449 and Limagrain Europe.

References

Heffner EL, Jannink JL, Sorrells ME (2011a) Genomic selection accuracy using multi-family prediction models in a wheat breeding program. Plant Genome 4:65–75

Heffner EL, Jannink JL, Iwata H et al (2011b) Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci 51:2597–2606

Heslot N, Jannink JL, Sorrells ME (2013) Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. Crop Sci 52:921–933

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127:463–480

Meuwissen TH (2009) Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Select Evol 41:35

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Rutkoski JE, Heffner EL, Sorrells ME (2010) Genomic selection for durable stem rust resistance in wheat. Euphytica 179:161–173

Rutkoski J, Benson J, Jia Y et al (2012) Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. Plant Genome 5:51–61

Rutkoski JE, Poland J, Jannink JL, Sorrells ME (2013) Imputation of unordered markers and the impact on genomic selection accuracy. G3 3:427–439