BRIEFINGS IN FUNCTIONAL GENOMICS. VOL 9. NO 2. 166–177 doi:10.1093/bfgp/elq001

Genomic selection in plant breeding: from theory to practice

Jean-Luc Jannink, Aaron J. Lorenz and Hiroyoshi Iwata

Advance Access publication date 15 February 2010

Abstract

We intuitively believe that the dramatic drop in the cost of DNA marker information we have experienced should have immediate benefits in accelerating the delivery of crop varieties with improved yield, quality and biotic and abiotic stress tolerance. But these traits are complex and affected by many genes, each with small effect. Traditional marker-assisted selection has been ineffective for such traits. The introduction of genomic selection (GS), however, has shifted that paradigm. Rather than seeking to identify individual loci significantly associated with a trait, GS uses all marker data as predictors of performance and consequently delivers more accurate predictions. Selection can be based on GS predictions, potentially leading to more rapid and lower cost gains from breeding. The objectives of this article are to review essential aspects of GS and summarize the important take-home messages from recent theoretical, simulation and empirical studies. We then look forward and consider research needs surrounding methodological questions and the implications of GS for long-term selection.

Keywords: *breeding value prediction; marker-assisted selection; linkage disequilibrium; ridge regression; machine learning*

INTRODUCTION

It has been predicted for over two decades that molecular marker technology would reshape breeding programs and facilitate rapid gains from selection [1, 2]. Currently, however, marker-assisted selection (MAS) has failed to significantly improve polygenic traits [3, 4]. While MAS has been effective for the manipulation of large effect alleles with known association to a marker [5], it has been at an impasse when many alleles of small effect segregate and no substantial, reliable effects can be identified [6].

The weaknesses of traditional MAS come from the way MAS splits the task into two components, first identifying QTL and then estimating their effects. QTL identification methods can make MAS poorly suited to crop improvement: (i) biparental populations may be used that are not representative and in any event do not have the same level of allelic diversity and phase as the breeding program as a whole [7, 8]; (ii) the necessity of generating such populations is costly such that the populations may be small and therefore underpowered; (iii) validation of discoveries is then warranted, requiring additional effort; (iv) the separation of QTL identification from estimation means that estimated effects will be biased [9–11], and small-effect QTL will be missed entirely [12] as a result of using stringent significance thresholds.

Association mapping (AM) applied directly to breeding populations has been proposed to mitigate the lack of relevance of biparental populations in QTL identification [13] and QTL have been mapped in this way [14, 15]. This practice nevertheless retains the disadvantage of biased effect estimates and therefore poor prediction of line performance [12, 16].

Corresponding author. Jean-Luc Jannink, USDA-ARS, R.W. Holley Center for Agriculture and Health; and Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York, 14853, USA. E-mail: jeanluc.jannink@ars.usda.gov

Jean-Luc Jannink works with the USDA-ARS, R.W. Holley Center for Agriculture and Health, and also at the Department of Plant Breeding and Genetics, Cornell University. He develops methods using DNA markers to increase the efficiency of small grain improvement and facilitate the use of those methods in public small grain breeding programs.

Aaron J. Lorenz is a postdoctoral research associate at the USDA-ARS, R.W. Holley Center for Agriculture and Health studying association mapping and genomic selection for the U.S. Department of Agriculture Barley Coordinated Agricultural Project.

Hiroyoshi Iwata is with the National Agriculture and Food Science Research Organization, National Agricultural Research Center at Japan. He applies data mining methods to a number of biological phenomena, from genetic map construction, to analysis of multi-dimensional shapes, to prediction of performance in breeding.

The solution to this quandary lies not in seeking single markers associated with single large effects but in capitalizing on the developing capacity for scoring many markers at low cost. Add to this capacity novel statistical methods that enable the simultaneous estimation of all marker effects and you get genomic selection (GS;[16]). GS uses a 'training population' of individuals that have been both genotyped and phenotyped to develop a model that takes genotypic data from a 'candidate population' of untested individuals and produces genomic estimated breeding values (GEBVs). These GEBVs say nothing of the function of the underlying genes but they are the ideal selection criterion. In the plant breeding context, untested individuals would belong to a broader population defined as a crop market class or the breeding program as a whole. In simulation studies, GEBVs based solely on individuals' genotype have been remarkably accurate [16–18]. These accuracies have held up in empirical studies of dairy cattle [19–21], mice [22, 23] and in biparental populations of maize, barley and *Arabidopsis* [24]. Given decreasing genotyping costs and stagnant or increasing phenotyping costs, and the ability to select individuals much earlier in the breeding cycle, GS is revolutionizing both animal [19, 25] and plant [26, 27] breeding.

In this context, the objectives of this article are to review essential aspects of GS and summarize the important take-home messages from recent theoretical, simulation and empirical studies. We then look forward and consider research needs surrounding the questions of best prediction methods, most informative training population design, and implications of GS for long-term selection.

GENOMIC SELECTION METHODS

GS emerged out of a desire to exploit high density parallel genotyping technologies [16]. At such high densities, it was assumed that linkage phase between markers or haplotype blocks of markers and causal polymorphisms would be consistent across families so that population-wide estimates of marker effects would be meaningful [16]. The authors also decided to avoid marker selection in the development of a prediction model so that estimated marker effects would be unbiased. A consequence of that decision was that more predictor effects, p, need to be estimated than the number, n, of available observations. Furthermore, there may be a high degree of correlation or multicollinearity between the predictors.

In so-called 'large p, small n' problems, standard multiple linear regression cannot be used without variable selection, which conflicts with the original goal of avoiding marker selection. An important danger in the development of a prediction model is overfitting: an overfitted model can exaggerate minor fluctuations in the data and will generally have poor predictive ability. To overcome these problems, a variety of methods, e.g. best linear unbiased prediction [28], ridge regression [29], Bayesian regression [16], kernel regression [30, 31] and machine learning methods [32–34], have been proposed to develop prediction models for GS.

Meuwissen *et al*. [16] estimated the effects of two-marker haplotypes though since then it has become far more common to estimate effects of single markers directly [35]. To make marker effects estimable, Meuwissen *et al*. [16] proposed to model them as random effects and calculate their best linear unbiased predictors (BLUP). These random effects were drawn from a normal distribution, $N(0, \sigma_g^2)$, where σ_g^2 was obtained by dividing the (known) genetic variance by the number of effects [17]. This parameterization, where all effects are assumed to have equal variance, is also called ridge regression and was first proposed for MAS by Whittaker *et al*. [36] in the context of biparental crosses. Note that assuming all marker effects are drawn from the same distribution does not mean the effects are all equal but that they are all equally shrunken toward zero.

The assumption of even distribution of genetic causation was not satisfactory and Meuwissen *et al*. [16] sought to relax it using two Bayesian analyses. In the first analysis (dubbed BayesA), each effect i is drawn from a normal distribution with its own variance: $N(0, \sigma_{gi}^2)$. The variance parameters are in turn sampled from a scaled inverted chi-squared distribution. In the second (dubbed BayesB), a further probability π is given that the marker has no effect at all. A more complete accounting of these methods and their relationship to traditional quantitative genetic models is given in Gianola *et al*. [37].

In this era of high-throughput data collection, other disciplines are also confronted with large p small n problems, and various methods have been proposed for their solution. Reduced-dimension regression methods such as partial least squares regression [PLS; 38] and principal component regression [PCR; 39] are well-known statistical methods in

chemometrics that are useful when the researcher is faced with many variables whose relationships are ill-understood, and the object is merely to construct a good predictive model [40]. In both methods, latent variables are extracted as linear combinations of the predictors and are used for response prediction. In PCR, the latent variables are chosen to explain as much of the variation in the original predictors as possible. In PLS, the latent variables are chosen so that the relationship between the latent variables and response is as strong as possible. The number of latent variables, generally determined through cross-validation, is much lower than the number of predictors or observations, which avoids model overfitting and achieves stable estimation of regression coefficients (e.g. genetic effects of genome-wide markers), though lower prediction accuracies than for BayesB have thus far been observed [41].

Machine-learning methods, such as support vector machine [SVM; 42] and random forest [RF; 43], have been successfully applied to data under large p, small n conditions in various research fields. Both methods were originally developed to solve a classification problem, but have been extended to the domain of regression [44]. The basic idea of SVM regression is to map samples from the predictor space to a high-dimensional feature space via a nonlinear mapping function and to do linear regression in this latter space [45]. Random Forest is an ensemble predictor consisting of a collection of tree-structured predictors, where each tree in the ensemble is 'grown' on the basis of a bootstrapped sample of the training dataset. Each tree individually predicts the target response and the 'forest' (i.e. the ensembles of 'trees') predicts the target response as an average of individual tree predictions. Since both SVM and RF build a non-linear prediction model, they may be especially useful when the relationships between predictors and responses are nonlinear, as would occur if epistatic effects account for a significant amount of genetic variation of a target trait. Non-parametric regression methods that may also account for non-additive effects have also been proposed [30, 31, 46], and in some cases perform favorably [30].

SIMULATION RESULTS

Overall performance and analysis method

Meuwissen *et al.* [16] and Habier *et al.* [17] evaluated the accuracies of ridge regression and BayesB using similar approaches assuming additive gene action and a heritability of 0.5. Forward-simulation of the population was performed to reach mutation-drift equilibrium under conditions that generated about 50 segregating QTL. For both studies, however, the expected effective QTL number [in the sense of ref. 12] was low: only 6 and 13 for Meuwissen *et al.* and Habier *et al.* respectively. We believe both of these numbers are unrealistic to model polygenic traits. For a training population size of 1000, respective GS accuracies were 0.66 and 0.64 for ridge regression and 0.79 and 0.69 for BayesB. The greater overall accuracy and greater difference between ridge regression and BayesB in Meuwisen *et al.* [16] can probably be attributed to the larger variances generated by individual QTL in that study.

Zhong *et al.* [18] took a different approach to simulation: rather than generating marker data from an ideal population in mutation–recombination–drift equilibrium, they took actual marker data from a diverse set of 42 lines of two-row barley. This approach retains the effects on linkage disequilibrium of the more complex and realistic demographic history of a crop. Of the markers available, 1040 were retained as evenly distributed over the genetic map. Training populations were simulated by randomly mating the founders to generate 500 lines and assuming a trait heritability of 0.4 with 80 QTL. In this case, the accuracies for ridge regression and BayesB were 0.62 and 0.61, respectively. The superiority of BayesB over ridge regression found in previous simulations was thus reversed in this case.

Two main take-home messages can be derived from the accuracies obtained in these simulation studies. First, in all cases, GS provided accuracies greater than might be achieved on the basis of pedigree information alone. Thus, if the objective is to accelerate the breeding cycle by making selections prior to extensive phenotyping, GS is the solution. Second, the more complicated random effect distribution used by the BayesB method is only useful if markers pick up strong associations with QTL. Such strong associations will occur when QTL effects themselves are large [particularly as in ref. 16] and when the associated markers are in high-linkage disequilibrium with the QTL. The importance of strong association can be confirmed in simulations by putting the QTL allelic states in the marker dataset, providing so-called perfect markers to the analysis. This situation improves predictions from BayesB more than from ridge regression [18].

Marker type and density

Solberg *et al.* [35] used the simulation conditions of Meuwissen *et al*. [16] to evaluate the effect of marker density and of SSR-like multiallelic markers versus SNP-like biallelic markers. They found that similar accuracies were achieved in different populations if the marker density scaled with each population's effective size (N_e). Historic recombination between loci scales linearly with N_e so that maintaining a fixed amount of recombination between loci also requires marker density to scale linearly with N_e [47]. As might be expected, accuracy increased with density though gains for a fixed density increment decreased at high density. Even at the highest tested densities of 2*Ne* SSR markers per Morgan or 8*Ne* SNP markers per Morgan, accuracy had not reached a plateau. Comparing the two marker types, they found that for similar accuracies, the SNP markers required a density of 2 to 3 times that of the SSR. Finally, assembling pairs of adjacent markers into haplotype blocks tended to decrease accuracy relative to considering all markers separately [35]. In these previous simulations, GEBVs were predicted on progeny of the training population. If predictions for less-related individuals are needed, higher marker density is needed [47]. Both the number of markers and the training population size will need to scale with N_e and with the length of the recombination genetic map marker L. When predicting GEBVs of individuals that are not more closely related to the training population than second cousins, Meuwissen [47] found that $10^{\times}N_e$ markers per Morgan and a training population size of $N_e^{\times}L$ generated accuracies between 0.73 and 0.83. The accuracy of BayesB benefitted more from increased marker density than that of ridge regression.

GS in biparental populations

Thus far, we have only discussed GS in the context of population-wide linkage disequilibrium, where the population might be defined as an entire breed of cattle, a market class of a crop (e.g. hard red wheat), or perhaps a breeding program. Because plants can often produce very large full sibships (an F_2 population derived from a single F_1 by selfing is an example of such a sibship), however, there is also a tradition of QTL detection, MAS and GS within such sibships [i.e. in F2, recombinant inbred line, or doubled haploid populations; 24, 48–50]. These simulations have almost exclusively used ridge regression. Some interesting results are (i) very low marker densities, on the order of eight per Morgan, can deliver accuracies close to the maximum observed; (ii) using ridge regression, there was a marker density optimum above which accuracy declined [48]; (iii) accuracy assuming true marker variances were known was only marginally higher (0–8%) than assuming all marker variances were equal [48]; (iv) GS can out-perform phenotypic selection even when the biparental population is composed of very few (e.g. 35) individuals [50]. As an overall population improvement strategy, no study has contrasted performing GS within biparental crosses to performing it across a breeding program as a whole. The primary advantage we see to the former approach is its low marker density need. The primary disadvantages are (i) that it requires separate model training within each cross: it seems suboptimal not to analyze all crosses jointly (as would occur if GS were performed over the breeding program as a whole); and (ii) the first generation of progeny from a cross cannot be selected on the basis of prior information but needs to be phenotyped. This practice slows down the breeding cycle relative to program-wide GS.

Joint use of linkage disequilibrium and co-segregation

The need for high marker densities in GS may be reduced if the candidate population consists of progeny of the training population. In that case, an evenly spaced low-density subset of the markers typed on the training population can be used on the candidates, and scores for the full complement of markers can be inferred by co-segregation [51]. This approach has also been proposed for association mapping in humans [52] and plants [53]. Assuming parents were always typed at high density, loss of accuracy due to typing the candidates at a density of only one marker every 10 cM ranged between 4% and 6% [51]. This loss was compared to what might be incurred if a low-density marker panel was developed by selecting markers most strongly associated with the trait. The performance of this latter approach depended on the number of QTL simulated, with lost accuracy ranging from 1% to 3%. The slightly greater losses of the evenly spaced compared to the selected marker approach must be set against its greater ease of development and its potential universality across traits and populations or breeds [51]. In addition, the evenly spaced markers will become fixed more slowly than those

directly selected upon, increasing their long-term value.

THEORETICAL STUDIES

Theoretical studies have yielded important results on three topics: (i) the sources of GS accuracy; (ii) accuracy formulated as a function of QTL number and training population size; and (iii) impacts of GS on long-term response. GS models genetic variance in two ways [17]. As expected, it uses markers in strong LD with QTL by estimating associated marker allele effects. However, as somewhat of an unanticipated side effect of GS arithmetic, it also uses marker data to model genetic relationships between individuals in the training and prediction populations. Accuracy of breeding values then also depends on the strength of predicted individuals' relatedness to training individuals with phenotypes, much as it would when using pedigree information to perform prediction. The way genetic relationships enter into GS can be demonstrated most clearly by showing that the ridge regression model is equivalent to (provides the same predictions as) a mixed model analysis in which random individual effects co-vary according to a kinship matrix calculated using marker data [17, 54]. Exploring these two sources of GS accuracy, Habier *et al.* [17] showed that ridge regression is more effective at capturing genetic relationships because it fits more markers into the prediction model. In contrast, BayesB is more effective at capturing LD between markers and QTL. Because these marker–QTL linkages are tight, recombination does not cause them to decay rapidly, and accuracies from BayesB persist longer than those from ridge regression [17, 18]. Habier *et al.* [17] developed a regression approach to quantify the relative importance of the two sources, finding that under their simulation conditions 39% and 21% of GS accuracy was due to capturing genetic relationships for ridge regression versus BayesB. Similar equivalencies have been shown by Piepho [55], who compared GS to spatial analyses of field trials.

This research on the sources of GS accuracy has bearings on predicting overall accuracy and on the impacts for long-term selection. Analytical models of GS accuracy at the moment account solely for accuracy due to markers in strong LD with QTL [54, 56]. Daetwyler *et al.* [56] assumed additive and independent loci and modeled locus effects as fixed. They derived the correlation between predicted and true breeding value as

$$r_{g\hat{g}} = \sqrt{\frac{\gamma h^2}{\gamma h^2 + 1}}$$

where γ is the ratio of the number of phenotyped individuals, n_P, to the number of loci, n_G and h^2 is the entry-mean basis heritability for the trait.

Hayes *et al.* [54] increased the realism of the analysis by modeling locus effects as random and deriving an approximation for the effective number of independent chromosome segments, which indicates how many effects the GS model must estimate. While the predicted accuracies they developed look unwieldy, they can already begin to answer interesting questions. For example, if a program is constrained primarily by the number of field plots that it can evaluate, will accuracy be maximized by evaluating many unreplicated individuals (i.e. planting each plot to a unique individual), or can accuracy be increased by replicating individuals across plots (i.e. using several plots to evaluate one individual with lower error)? For both the Daetwyler *et al.* [56] and the Hayes *et al.* [54] analyses, one can show that GS accuracy should be maximized by a strategy of evaluating unreplicated individuals. This conclusion was also reached for maximizing QTL detection power [57]. These analytical predictions can be contrasted to simulations of the same phenomenon. Zhong *et al.* [18] simulated cases of a training population of 504 at a heritability of 0.4 versus a training population of 168 at a heritability of 0.67. This mimicked the relative heritabilities for one versus three independent repeated measures. Realized accuracies were 0.61 versus 0.62 for BayesB and 0.62 versus 0.66 for ridge regression. In other words, stochastic simulation gave the opposite result to what was expected from theory. The theory, however, considers only the component of accuracy due to LD between markers and QTL. When heritability increases, the component due to genetic relationships will gain in importance and, as observed, ridge regression should benefit more from that than BayesB.

The relative importance of LD between markers and QTL versus genetic relationships in determining accuracy also affects the loss of genetic diversity through GS, that is, the impact of GS on long-term gain. GS should maintain greater genetic diversity while increasing selection gains for the following reason [58]: in the absence of markers, more

accurate predictions of individual breeding value are possible by using information from relatives. This information not only increases selection gain but also increases the correlation between predicted breeding values for relatives. For example, in the absence of progeny testing, predicted breeding values for full sibs on the basis of family information are identical. In turn, the greater correlation in predicted breeding values between relatives causes more frequent co-selection of relatives and concomitant decline in genetic diversity. The key problem with information from relatives is that it contributes nothing to predicting the value of the specific alleles each progeny received from its parents, the so-called Mendelian segregation term [58]. GS mitigates this problem because those specific alleles are in LD with markers that have estimated effects. Compared to traditional BLUP evaluation, therefore, the correlation between predicted breeding values of relatives will be lower under GS. Given accuracy due to LD between markers and QTL, less loss of genetic diversity and greater long-term gains should be possible under GS [59].

This argument from theory again relies on the LD rather than the genetic relationship component of GS accuracy. Given the importance of the relative proportion of these two components to many aspects of GS, we have estimated the components under a wide simulated range of training population size, marker density, and genetic architecture conditions (Figure 1). From this brief exercise, it is clear that accuracy due to genetic relationships can represent from a small minority to a large majority proportion of the overall accuracy. Factors that we looked at that reduced that proportion were fewer QTL, higher marker density, larger training population size, and as expected, BayesB versus ridge regression (Figure 1). We note that the low marker density, low training population size setting that we used (400 markers and 400 individuals) is in the realm of what might be typical for small public sector plant breeding programs. Under those circumstances, the majority of GS is due to genetic relationship information and therefore the theoretical results given above may be off the mark.

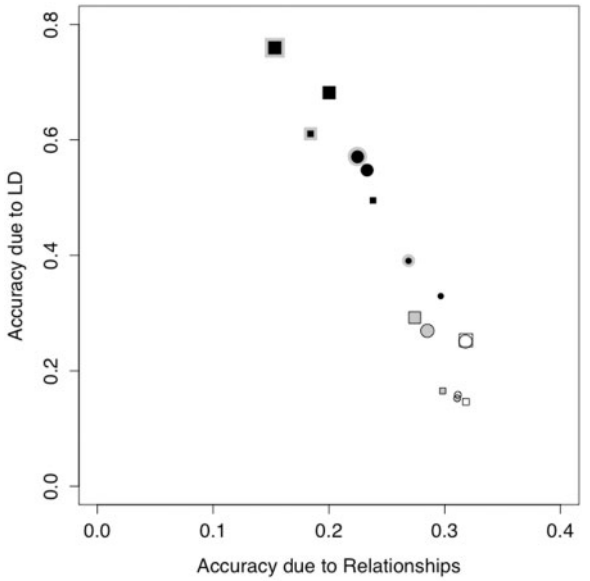

Figure 1: Decomposition of GS prediction accuracy using the method of Habier *et al*. [17]. On a genome comprised of seven chromosomes of 1.5 M each, individuals were generated using a coalescent assuming an effective population size of 100. Round and square symbols, ridge regression and Bayes-B, respectively. Symbols with gray (inside or around) and without, 40 QTL and 200 QTL, respectively. Black and non-black symbols, 4000 and 400 markers, respectively. Small and large symbols, training population size of 400 and 2000, respectively

EMPIRICAL STUDIES

Large-scale empirical studies are not yet available in the public sector for plants, but insight can be gained from livestock studies, particularly in dairy cattle. The largest single study was conducted by VanRaden *et al*. [21]. The training population contained over 3500 Holstein bulls with breeding values measured by progeny testing and genotyped with 38 416 SNP. They achieved accuracies of 0.44 to 0.79 for traits ranging in heritability from 0.04 to 0.50 (though note that the training bulls were characterized by progeny means of high accuracy). Decreasing marker number by 75% decreased the accuracy of net merit only from 0.53 to 0.50, while decreasing the training population size by 68% decreased that accuracy from 0.53 to 0.35. Increases in accuracy as a function of training population size were quite linear up to the maximum size available. Both ridge regression and a variant of BayesB [21] gave very similar accuracies.

A review of studies from three other dairy cattle GS experiments showed similar results [19]. The main observations from these studies are: (i) GS methods predicted breeding values better than did pedigree information alone, but less well than was expected based on simulations; (ii) GS methods

that assume many QTL evenly distributed over the genome (i.e. ridge regression) perform as well as methods that assume fewer QTL of varying effect (e.g. BayesB); (iii) decreasing marker numbers did not strongly affect GS accuracy; and (iv) GS accuracy increased linearly with training population size. One interpretation of these observations is that the infinitesimal model assumption, 'an infinite number of loci, all with infinitesimally small effects', is closer to being correct than an assumption of few QTL (where 'few' could mean dozens but not hundreds). Alternatively, there may be relatively few loci at which variants have a large effect on the phenotype, but these variants are at low frequency so that they each generate little variance. If loci carry several low frequency, high effect variants, a condition would arise where substantial genetic variance and high heritability would be possible, but where LD between markers and QTL would generally be low. The LD component of accuracy would therefore be constrained. This is one of the genetic architectures that is invoked in 'the case of the missing heritability' [60]. This case refers to instances of human association study where very little variation is explained by associated markers, even for traits with high heritability to which substantial effort at association has been applied (e.g. height studied in panels of 30 000 individuals). Recent extensive mapping efforts for flowering time in maize [61] provide support for the common gene hypothesis that the many variants that affect maize flowering time are clustered in a few common loci. This genetic architecture generates high heritability and resemblance between relatives but low association between QTL and markers: it would lead ridge regression to be more effective than BayesB.

GS has also been applied to data on a mouse population synthesized from eight inbred mice [22, 23]. Because of this narrow base, alleles that are polymorphic are expected to have minor allele frequencies strongly biased toward high values. QTL analysis of this population did not result in a 'case of missing heritability' [62]. Given this fact, it would be valuable to contrast ridge regression and BayesB analyses in this synthetic population. Such a contrast has not been performed. The Legarra *et al*. study used ridge regression while the Lee *et al*. study used an analysis similar to BayesB, but the two studies analyzed different traits. Both studies split the population into a training half and a validation half in two ways, either across families (different families ending up in the different halves) or within families (different individuals within families ending up in the different halves). For the split across families, only capturing LD between marker and QTL will be useful for prediction because the families were weakly related. In contrast, for the split within families, capturing genetic relatedness will also be useful. Interestingly, for traits of similar heritability, prediction across families was more accurate in the Lee *et al*. (BayesB-like analysis) study than in the Legarra *et al*. (ridge regression) study. Conversely, prediction within families was more accurate in the Legarra *et al*. than the Lee *et al*. study. We hypothesize that the high accuracy of the BayesB-like analysis across families was due in part to the unusual origin of this population.

FUTURE RESEARCH

Training population design

As envisioned in its purest form, GS will dramatically change the purpose of phenotyping in plant breeding [27]: phenotyping currently serves to determine which lines to select; under GS, phenotyping will serve primarily to train prediction models. While it is well known that GEBV accuracy increases as the size of the training population increases [21], to our knowledge no research has been conducted on training population design to develop accurate GEBV models while minimizing resources spent on phenotyping.

Maximizing marker variance, reducing collinearity between markers and uniformly sampling the genetic diversity of the breeding program are three possible objectives of training population design. Maximizing marker variance might be achieved by choosing individuals with divergent GEBVs. Simulations by Zhong *et al*. [18] suggested that, for certain GS models, collinearity reduced prediction accuracy. Collinearity between linked markers is reduced by recombination, suggesting that progeny that experienced a greater number of total recombination events should be phenotyped. This approach has been shown to be useful in QTL mapping [63]. Uniformly sampling a population's genetic diversity could be achieved by clustering based on multivariate distance statistics [64]. Such samples should improve estimates of the effects of rare alleles.

Near-infrared reflectance spectroscopy (NIRS) is analogous to GS as an application of multivariate

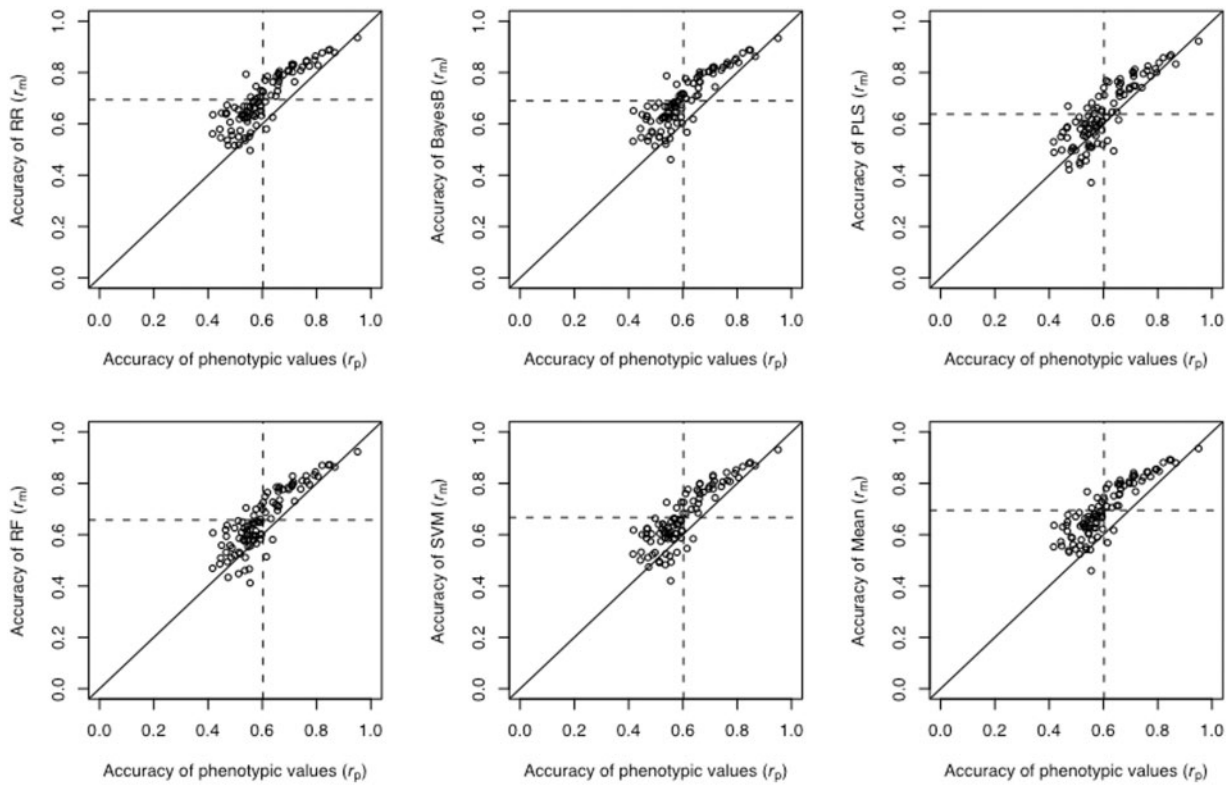

Figure 2: Accuracy of breeding value prediction for different GS methods compared to phenotypic selection. In each graph, a point is one simulation. Starting with 1325 SNP in barley, 80 SNP were removed to serve as additive-effect QTL. Expected trait heritability was 0.4 but varied between simulations because of covariance between QTL. Training population size was 400. Dashed lines correspond to mean accuracies for the GS method (horizontal) and the phenotype (vertical). The lower right-hand graph shows accuracy of the mean across five GS methods.

statistics to model development and prediction. Like GS, the advantage of NIRS is that a large set of variables is cheap to measure (NIR spectra) and can predict variables that are expensive to measure (wet chemistry measurements). NIRS has been intensively researched for decades [65]. Spectra (absorbance values at each of thousands of wavelengths) are collected on a large population of samples, and a subset of samples to be phenotyped is chosen. Prediction typically involves relating phenotypes to the spectra through PCR or PLS regression [66]. A common goal of selecting samples for phenotyping is to evenly span the range of spectral and phenotypic variation of the population, while minimizing the size of the set [65]. One multivariate distance metric often used for selecting samples that uniformly span the spectral diversity is the Mahalanobis distance [*H* distance; 65, 67]. The *H* distance accounts for collinearity between predictors in calculating their distance [68]. The *H* distance can also be used to define a population of samples similar enough that it could be predicted using a single training set and to identify outlier samples [67, 69]. In GS, a statistic similar to the *H* distance based on marker data could also relate training population diversity to model accuracy.

Routine use of NIRS involves a continual need to update the calibration as new variation in the phenotype is encountered. Several guidelines exist for deciding when a particular calibration can be used to predict new samples, and which samples should be added to the existing training population [70]. We envision similar guidelines for training population maintenance in GS. Because generations following a given selection event will contain only the alleles of the parents in each cycle of selection, it may be most efficient to update the training population by phenotyping the parents of each selection cycle. Empirical and simulation findings should resolve this question. Theory and practice in other areas such as

chemometrics could prove to be useful starting points.

Capitalizing on the strength of different methods

We have seen that different GS methods use substantially different approaches to address the large *p* small *n* problem. The methods may therefore capture different aspects of the marker genotype to phenotype map, and could complement each other. If such complementation occurs, a synthesis of methods might be superior to any single method. In the same way that Random Forest averages a number of predictors to achieve more accurate predictions, combining methods may be valuable. We have explored GS accuracy using a series of parametric and non-parametric methods (Figure 2), the details of which will be in a forthcoming publication. In general, the parametric methods (ridge regression and BayesB) outperformed the non-parametric methods (PLS, RF and SVM). Our most unexpected observation, however, was that a simple mean across all methods did best (Figure 2). Note that it was just barely superior to the best single method (ridge regression), but we find it surprising that by combining ridge regression with other methods that gave poorer accuracies, a meta-predictor can emerge that does best of all. Further theory needs to be explored to understand what signal is captured by the different methods to determine how to combine them to obtain maximum accuracy.

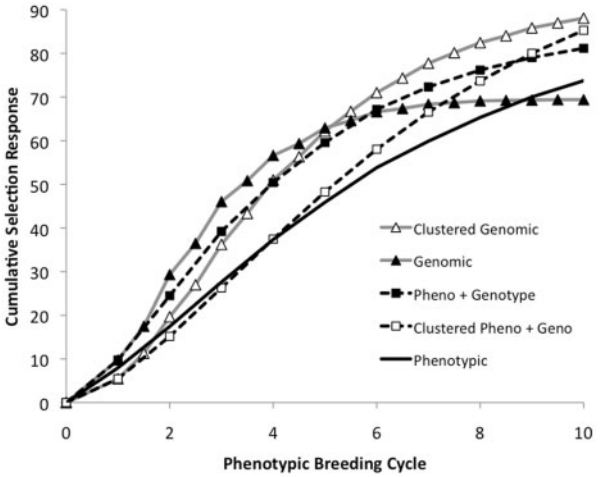

Figure 3: Simulated long-term selection response to different genomic selection approaches. In each generation, 200 candidates were evaluated and 20 selected. Marker data in the founder population came from the University of Minnesota barley breeding program. The trait was determined by 100 QTL with a heritability of 0.2. GEBVs were obtained from a BayesB analysis. Simple phenotypic selection was compared to either genomic selection performed after phenotyping each individual (Pheno + Genotype), or performed prior to phenotyping (Genomic), in which case it was assumed that the genomic selection breeding cycle time was half that of phenotypic selection and that the training population was updated every other breeding cycle. 'Clustered' methods of genomic selection indicate that marker data were used to group candidates into 20 clusters, and the best candidate in each cluster was selected. This practice reduced short-term but increased long-term response.

Managing short- and long-term selection gain

If QTL are in complete LD with markers, theory shows that GS should cause less inbreeding or loss of genetic diversity than selection on breeding values estimated using pedigree information [58]. Obviously, this condition does not hold and, in reality, GS can fall short of phenotypic selection: (i) GS will not 'discover' some QTL and these will drift rather than be subject to selection [71]; (ii) if marker and QTL are not in perfect LD, fixing a marker will not fix the QTL [72]; (iii) finally, as we have seen, GS does capture some relationship information increasing the likelihood of co-selection of relatives. For traditional pedigree-based selection, methods have been developed to select while constraining the rate of increase of relatedness in the population [73]. For GS, it seems sensible that we should also take advantage of marker data to manage inbreeding and optimize long-term selection gains. For example, Goddard [71] proposed varying the weight given to marker information as a function of the allele frequencies at each marker [19]. It would also be possible to use markers to mimic within-family selection, a practice that reduces the rate of inbreeding. We have done within-group selection by using marker data to cluster selection candidates and then selecting within clusters (Figure 3). Such selection reduces short-term but increases long-term gain. Figure 3 shows that beyond accelerating selection response, marker data offers wide possibilities for managing short- and long-term gains. Research into these possibilities has just begun [71, 74].

Key Points

- There is little doubt that GS will change plant breeding practices and efficiency, and that it is experiencing an intense period of scientific research activity.
- Even though GS is just beginning to be implemented, we believe its practice currently outpaces its theory. We argue there is a need for theory to (i) guide the design of training populations; (ii) predict the accuracy to be expected from GS as a function of training population size, genome-wide linkage disequilibrium and marker density; (iii) understand the reasons and contexts in which different methods perform best and how to optimally combine their predictions; and (iv) use marker information to manage genetic contributions from ancestors to maximize short-term gain without compromising long-term gain.
- Existing empirical studies make it clear that the underlying genetic architecture, as characterized at least by the number of QTL and the distributions of their allelic effects and frequencies, differentially affects performance of different GS methods. Future empirical work will therefore provide both fascinating insight into that architecture and important reality checks as genomic selection is implemented in plant breeding.

FUNDING

This work was supported by the United States Department of Agriculture, National Institute of Food and Agriculture (2009-85606-05701, 2009-65300-05661).

References

1. Stuber CW, Goodman MM, Moll RH. Improvement of yield and ear number resulting from selection at allozyme loci in a maize population. *Crop Sci* 1982;**22**:737–40.
2. Tanksley SD, Young ND, Paterson AH, *et al*. RFLP mapping in plant breeding: new tools for an old science. *Biotechnology* 1989;**7**:257–64.
3. Bernardo R. Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci* 2008; **48**:1649–64.
4. Xu Y, Crouch JH. Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci* 2008;**48**: 391–407.
5. Zhong S, Toubia-Rahme H, Steffenson BJ, *et al*. Molecular mapping and marker-assisted selection of genes for septoria speckled leaf blotch resistance in barley. *Phytopathology* 2006;**96**:993–9.
6. Moreau L, Charcosset A, Gallais A. Experimental evaluation of several cycles of marker-assisted selection in maize. *Euphytica* 2004;**137**:111–8.
7. Jannink J-L, Bink MCAM, Jansen RC. Using complex plant pedigrees to map valuable genes. *Trends Plant Sci* 2001;**6**:337–42.
8. Sneller CH, Mather DE, Crepieux S. Analytical approaches and population types for finding and utilizing QTL in complex plant populations. *Crop Sci* 2009;**49**:363–80.
9. Beavis WD. The power and deceit of QTL experiments: lessons from comparative QTL studies. In: Wilkinson DB (ed). *Proceedings of the 49th Annual Corn and Sorghum Research Conference*. Washington, DC: American Seed Trade Association, 1994:250–65.
10. Melchinger AE, Utz HF, Schon CC. Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics* 1998;**149**:383–403.
11. Schon CC, Utz HF, Groh S, *et al*. Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics* 2004;**167**:485–98.
12. Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 1990;**124**:743–56.
13. Rafalski JA. Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci* 2002;**162**: 329–33.
14. Crossa J, Burgueno J, Dreisigacker S, *et al*. Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. *Genetics* 2007;**177**:1889–913.
15. Kraakman ATW, Niks RE, Van den Berg PMMM, *et al*. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* 2004;**168**: 435–46.
16. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 2001;**157**:1819–29.
17. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 2007;**177**:2389–97.
18. Zhong S, Dekkers JCM, Fernando RL, *et al*. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* 2009;**182**:355–64.
19. Hayes BJ, Bowman PJ, Chamberlain AJ, *et al*. Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 2009;**92**:433–43.
20. Luan T, Woolliams JA, Lien S, *et al*. The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. *Genetics* 2009;**183**:1119–26.
21. VanRaden PM, Van Tassell CP, Wiggans GR, *et al*. Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci* 2009;**92**:16–24.
22. Lee SH, van der Werf JHJ, Hayes BJ, *et al*. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet* 2008;**4**:e1000231.
23. Legarra A, Robert-Granie C, Manfredi E, *et al*. Performance of genomic selection in mice. *Genetics* 2008; **180**:611–8.
24. Lorenzana R, Bernardo R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor Appl Genet* 2009;**120**:151–61.
25. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet* 2006;**123**:218–23.
26. Eathington SR, Crosbie TM, Edwards MD, *et al*. Molecular markers in a commercial breeding program. *Crop Sci* 2007; **47**:S154–63.
27. Heffner EL, Sorrells ME, Jannink J-L. Genomic selection for crop improvement. *Crop Sci* 2009;**49**:1–12.

28. Kolbehdari D, Schaeffer LR, Robinson JA. Estimation of genome-wide haplotype effects in half-sib designs. *J Anim Breed Genet* 2007;**124**:356–61.
29. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 2000;**42**: 80–6.
30. Bennewitz J, Solberg T, Meuwissen THE. Genomic breeding value estimation using nonparametric additive regression models. *Genet Select Evol* 2009;**41**:20.
31. Gianola D, van Kaam JBCHM. Reproducing Kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 2008;**178**: 2289–303.
32. Long N, Gianola D, Rosa GJM, *et al*. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J Anim Breed Genet* 2007;**124**:377–89.
33. Long N, Gianola D, Rosa G, *et al*. Comparison of classification methods for detecting associations between SNPs and chick mortality. *Genet Select Evol* 2009;**41**:18.
34. Wei Z, Wang K, Qu H-Q, *et al*. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet* 2009;**5**: e1000678.
35. Solberg TR, Sonesson AK, Woolliams JA, *et al*. Genomic selection using different marker types and densities. *J Anim Sci* 2008;**86**:2447–54.
36. Whittaker JC, Thompson R, Denham MC. Marker-assisted selection using ridge regression. *Genet Res* 2000;**75**:249–52.
37. Gianola D, de los Campos G, Hill WG, *et al*. Additive genetic variability and the Bayesian alphabet. *Genetics* 2009;**183**:347–63.
38. Wold H, Johnson NL, Kotz S. Partial least squares. *Encyclopedia of Statistical Science*. New York: Wiley, 1985:581–91.
39. Coxe KL, Johnson NL, Kotz S. Principal components regression analysis. *Encyclopedia of Statistical Science*. New York: Wiley, 1986:181–4.
40. Tobias RD. *An Introduction to Partial Least Squares Regression*. Cary, NC: SAS Institute, 1997.
41. Solberg TR, Sonesson AK, Woolliams JA, *et al*. Reducing dimensionality for prediction of genome-wide breeding values. *Genet Select Evol* 2009;**41** Article No.: 29.
42. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
43. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
44. Drucker H, Burges CJC, Kaufman L, *et al*. Support vector regression machine. *Adv Neural Info Proc Syst* 1997;**9**: 55–67.
45. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann, 2005.
46. Gonzalez-Recio O, Gianola D, Long N, *et al*. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* 2008;**178**:2305–13.
47. Meuwissen THE. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genet Select Evol* 2009;**41**:35.
48. Bernardo R, Yu J. Prospects for genome-wide selection for quantitative traits in maize. *Crop Sci* 2007;**47**:1082–90.
49. Bernardo R. Genomewide selection for rapid introgression of exotic germplasm in maize. *Crop Sci* 2009; **49**:419–25.
50. Wong C, Bernardo R. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor Appl Genet* 2008;**116**:815–24.
51. Habier D, Fernando RL, Dekkers JCM. Genomic selection using low-density marker panels. *Genetics* 2009; **182**:343–53.
52. Burdick JT, Chen W-M, Abecasis GR, *et al*. In silico method for inferring genotypes in pedigrees. *Nat Genet* 2006;**38**:1002–4.
53. Yu J, Holland JB, McMullen MD, *et al*. Genetic design and statistical power of nested association mapping in maize. *Genetics* 2008;**178**:539–51.
54. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Res* 2009;**91**:47–60.
55. Piepho HP. Ridge regression and extensions for genome-wide selection in maize. *Crop Sci* 2009;**49**:1165–76.
56. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* 2008;**3**:e3395.
57. Knapp SJ, Bridges WC. Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for unreplicated and replicated progeny. *Genetics* 1990;**126**:769–77.
58. Daetwyler HD, Villanueva B, Bijma P, *et al*. Inbreeding in genome-wide selection. *J Anim Breed Genet* 2007;**124**: 369–76.
59. Dekkers JCM. Prediction of response to marker-assisted and genomic selection using selection index theory. *J Anim Breed Genet* 2007;**124**:331–41.
60. Maher B. Personal genomes: the case of the missing heritability. *Nature* 2008;**456**:18–21.
61. Buckler ES, Holland JB, Bradbury PJ, *et al*. The genetic architecture of maize flowering time. *Science* 2009;**325**: 714–8.
62. Valdar W, Solberg LC, Gauguier D, *et al*. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet* 2006;**38**:879–87.
63. Jannink JL. Selective phenotyping to accurately map quantitative trait loci. *Crop Sci* 2005;**45**:901–8.
64. Mohammadi SA, Prasanna BM. Analysis of genetic diversity in crop plants - Salient statistical tools and considerations. *Crop Sci* 2003;**43**:1235–48.
65. Shenk JS, Workman JJ, Jr, Westerhaus MO. Application of NIR spectroscopy to agricultural products. In: Burns DA, Ciurczak EW (eds). *Handbook of Near-Infrared Analysis*. Boca Raton, LA: Taylor and Francis, 2001:419–74.
66. Workman JJ, Jr. Nir spectroscopy calibration basics. In: Burns DA, Ciurczak EW (eds). *Practical Spectroscopy Series*. Vol.13 *Handbook of Near-Infrared Analysis*, Xvii+681p. New York, New York, USA: Marcel Dekker, Inc.; Basel, Switzerland.Illus, 1992, 247–80.
67. Shenk JS, Westerhaus MO. Population definition, sample selection, and calibration procedures for near-infrared reflectance spectroscopy. *Crop Sci* 1991;**31**:469–74.
68. De Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* 2000;**50**:1–18.

69. Mark HL, Tunnell D. Qualitative near-infrared reflectance analysis using Mahalanobis distances. *Anal Chem* 1985;**57**: 1449–56.

70. Shenk JS, Fales SL, Westerhaus MO. Using near-infrared reflectance product library files to improve prediction accuracy and reduce calibration costs. *Crop Sci* 1993;**33**:578–81.

71. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 2009;**136**: 245–57.

72. Muir WM. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *J Anim Breed Genet* 2007;**124**:342–55.

73. Avendaño S, Woolliams JA, Villanueva B. Mendelian sampling terms as a selective advantage in optimum breeding schemes with restrictions on the rate of inbreeding. *Genet Res* 2004;**83**:55–64.

74. Li Y, Kadarmideen HN, Dekkers JCM. Selection on multiple QTL with control of gene diversity and inbreeding for long-term benefit. *J Anim Breed Genet* 2008;**125**:320–9.