RESEARCH

Genomic Selection in Plant Breeding: A Comparison of Models

Nicolas Heslot, Hsiao-Pei Yang, Mark E. Sorrells, and Jean-Luc Jannink*

ABSTRACT

Simulation and empirical studies of genomic selection (GS) show accuracies sufficient to generate rapid genetic gains. However, with the increased popularity of GS approaches, numerous models have been proposed and no comparative analysis is available to identify the most promising ones. Using eight wheat (*Triticum aestivum* L.), barley (*Hordeum vulgare* L.), *Arabidopsis thaliana* (L.) Heynh., and maize (*Zea mays* L.) datasets, the predictive ability of currently available GS models along with several machine learning methods was evaluated by comparing accuracies, the genomic estimated breeding values (GEBVs), and the marker effects for each model. While a similar level of accuracy was observed for many models, the level of overfitting varied widely as did the computation time and the distribution of marker effect estimates. Our comparisons suggested that GS in plant breeding programs could be based on a reduced set of models such as the Bayesian Lasso, weighted Bayesian shrinkage regression (wBSR, a fast version of BayesB), and random forest (RF) (a machine learning method that could capture nonadditive effects). Linear combinations of different models were tested as well as bagging and boosting methods, but they did not improve accuracy. This study also showed large differences in accuracy between subpopulations within a dataset that could not always be explained by differences in phenotypic variance and size. The broad diversity of empirical datasets tested here adds evidence that GS could increase genetic gain per unit of time and cost.

N. Heslot and M.E. Sorrells, Dep. of Plant Breeding and Genetics, Cornell Univ., Ithaca, NY 14853; H.-P. Yang and J.-L. Jannink, USDA-ARS, R.W. Holley Center for Agriculture and Health, Cornell Univ., Ithaca, NY 14853. N. Heslot, Limagrain Europe, ZAC Portes de Riom, Riom, France, 63200. Received 1 June 2011. *Corresponding author (jeanluc.jannink@ars.usda.gov).

Abbreviations: BL, Bayesian Lasso; BRR, Bayesian ridge regression; CAP, Coordinated Agricultural Project; E-Bayes, empirical Bayes; EM, expectation maximization; *Fst*, the *F* statistics quantifying the differences in allele frequency among subpopulations; GEBV, genomic estimated breeding value; GS, genomic selection; LD, linkage disequilibrium; MCMC, Markov chain Monte Carlo; NNET, neural network; PCA, principal component analysis; QTL, quantitative trait loci/locus; RF, random forest; RKHS, reproducing kernel Hilbert space; RR-BLUP, random regression best linear unbiased predictor; SVM, support vector machine; SVR, support vector regression; wBSR, weighted Bayesian shrinkage regression.

The genomic selection (GS) concept encompasses a broad range of methods. Their common feature is the ability to estimate breeding values for quantitative traits based on whole genome genotypes through the simultaneous estimation of marker effects in a single step. This concept was first proposed by Meuwissen et al. (2001) along with several new statistical models. Since then, further models have been proposed. Simulations and empirical studies have demonstrated that GS can greatly accelerate the breeding cycle, maintain genetic diversity within the breeding programs, and increase genetic gain beyond what is possible with phenotypic selection or quantitative trait loci (QTL) approaches. Nevertheless, it is important to identify the best methods and statistical procedures for using high-throughput molecular marker technologies and previously available phenotypic records to

Published in Crop Sci. 52:146–160 (2012).
doi: 10.2135/cropsci2011.06.0297
Published online 17 Nov. 2011.

accelerate genetic gains per unit of time and cost. Several recent reviews are available on GS in plant breeding, in particular Heffner et al. (2009), Jannink et al. (2010), Xu and Hu (2010), and Lorenz et al. (2011).

There are few extensive studies of the comparative predictive ability of the proposed models in plants or in animals. Lorenzana and Bernardo (2009) showed that, in the case of biparental populations, the predictive ability of the models they tested (ridge regression and empirical Bayes [E-Bayes] [Xu, 2007]) was fairly similar. Heffner et al. (2011) compared several models for predictive ability in a multiparental wheat (*Triticum aestivum* L.) population. Crossa et al. (2010) focused on Bayesian Lasso (BL) and reproducing kernel Hilbert space (RKHS) models to evaluate GS for wheat and maize (*Zea mays* L.) improvement. Our objective in this study was to thoroughly compare all the models published to date, along with several machine learning procedures not previously evaluated for GS, using the same evaluation methods on several species, traits, and datasets. In addition, none of the above cited model comparison studies measured the level of overfitting in each model, which is also an important factor to quantify. It should also be emphasized that for a given level of accuracy, models use different assumptions on QTL effect distributions resulting in different marker effect distributions. Therefore, even if they have the same predictive ability, two models will likely give different genomic estimated breeding values (GEBVs) and exert different selection pressures along the genome. Our goal was to identify the most promising models, provide some recommendations for the implementation of GS approaches in breeding programs, and obtain empirical evidence of model similarities and dissimilarities.

MATERIALS AND METHODS

Phenotypic and Genotypic Data

Eight datasets of different origins were used (Table 1) including two published datasets previously used to test GS models for *Arabidopsis thaliana* (L.) Heynh. (Bay × Sha [Bay-0 × Shahdara]) (Lorenzana and Bernardo, 2009) and wheat (Wheat CIMMYT) (Crossa et al., 2010). The Wheat Cornell dataset used is a subset of the dataset used in Heffner et al. (2011). The Barley Coordinated Agricultural Project (CAP) dataset was from the Barley Coordinated Agricultural Project (2011). All other datasets were provided by Limagrain Europe (Riom, France). For both maize datasets, phenotype data were obtained from a testcross to a Limagrain Europe inbred.

Several types of markers were used. Diversity array technology markers (DArT) markers are dominant and single nucleotide polymorphism (SNP) and simple sequence repeat (SSR) are codominant. Missing marker data were imputed as the mean of the nonmissing data at the level of each marker.

Models Tested

Eleven GS models were used to estimate the genetic value of individuals.

Random regression best linear unbiased predictor (RR-BLUP), also named ridge regression, was used with either a grid search over the shrinkage parameter λ or an estimation of the level of shrinkage using a mixed model approach with the "emma" R package (Kang et al., 2008). We also used the Bayesian ridge regression (BRR) as implemented in the R package "BLR" (Pérez et al., 2010). The model is of the form

$$Y = \mu + \mathbf{X}\boldsymbol{\beta} + \varepsilon,$$

where Y is the trait value, μ is the population mean, $\mathbf{X}$ is the marker design matrix, $\boldsymbol{\beta}$ is the vector of marker effects, and the error term, ε, is assumed to be normally distributed with mean and variance equal to 0 and σ_2. The estimator of $\boldsymbol{\beta}$ is $(X`X + \lambda I)^{-1}X`y$. This estimator can be expressed as

$$\arg\min_{\beta}\left(\frac{\|y - X\beta\|_2^2}{2\sigma^2} + \lambda\|\beta\|_2^2\right),$$

with the notation $\|.\|_2$ used for the L_2 norm or Euclidean norm $\|\beta\|_2 = \left(\Sigma_i \beta_i^2\right)^{1/2}$. The $\arg\min_{\beta}$ notation refers to the determination of coefficients β minimizing the expression inside the brackets. Random regression best linear unbiased predictor assumes all markers have a common variance (Meuwissen et al., 2001) and therefore shrinks equally for each marker effect. Bayesian ridge regression makes the same assumptions as RR-BLUP but the level of shrinkage is estimated with a Bayesian hierarchical model.

In the case of the BL (Park and Casella, 2008; Yi and Xu, 2008; de los Campos et al., 2009), the shrinkage is marker specific and dependent on a regularization parameter λ. The estimator of β is

$$\arg\min_{\beta}\left(\frac{\|y - X\beta\|_2^2}{2\sigma^2} + \lambda\|\beta\|_1\right),$$

thus illustrating the similarities between the BL and RR-BLUP. In both cases, $\|y - X\beta\|_2^2$ is a sum of squares penalty while the remaining term is a penalty function promoting sparseness. In the BL, this function is based on the L_1 norm (also named Taxicab or Manhattan norm)

$$\|\beta\| = \sum_i |\beta_i|$$

while in RR-BLUP it is based on the L_2 norm as described above. The BL produces stronger shrinkage of regression coefficients that are close to zero and less shrinkage of those with large absolute values, leading to a sparse model, whereas RR-BLUP shrinks more strongly the regression coefficients with a large value. For the ridge regression, there are several possibilities to determine λ as described above. In the Bayesian version of the Lasso, each marker effect β_j is assigned a normal prior of mean 0 and variance σ_j^2. Each σ_j^2 follows an independent exponential prior with parameter $\lambda^2/2$. A Gamma prior is further assigned to λ. It is important to note a fundamental difference between the BL and the lasso: The Bayesian version does not select variables by assigning coefficients to 0 as does the non-Bayesian version. For both BRR and BL we used the default prior parameters provided in Pérez et al. (2010) with 60,000 iterations and the first 10,000 iterations were discarded as burn-in.

The elastic net (Zou and Hastie, 2005) relies on a combination of both the L_1 norm (lasso) and L_2 norm penalties (ridge regression). The estimator of β is

$$(1+\lambda_2)\arg\min_{\beta}\left(\frac{\|y - X\beta\|_2^2}{2\sigma^2} + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1\right),$$

Table 1. Dataset origins and details.

Name	Crop	Type	Traits	Number of genotypes	Phenotypic data	Number and type of marker	Origin
Barley 1	Spring barley	Panel†, elite breeding lines	Yield	761	Three years, eight trial locations per year on average, across mainland Europe, nine replicates per genotype on average	338 SNP‡	Limagrain Europe (Riom, France)
Barley CAP§	Spring barley	Panel with structure	Betaglucan content	911	Three years, unbalanced data from five locations per year. Lines per trial ranged from 22 to 96.	2146 SNP	Barley CAP project
Bay × Sha (Bay-0 × Shahdara)	*Arabidopsis thaliana* (L.) Heynh.	Biparental population from two ecotypes, Bay-0 and Shadara	Flowering time under short day length, dry matter under nonlimiting or limiting conditions	422	Data available from the Study of the Natural Variation of *Arabidopsis thaliana* website (INRA, 2007)	69 SSR¶	INRA# (France) (Loudet et al., 2002; Lorenzana and Bernardo, 2009)
Panel maize	Elite maize	Panel, elite lines (one heterotic group)	Yield and moisture content	332	BLUE†† of GCA‡‡ for both traits, tested in 2009 in northern Italy.	355 SNP	Limagrain Europe
Diallel maize	Elite maize	Partial diallel, elite lines (one heterotic group) (six parents and five crosses)	Yield and moisture content	370	BLUE of GCA for both traits, tested in 2009 in northern Italy.	319 SNP	Limagrain Europe
Wheat CIMMYT	Spring wheat	Panel	Yield measured in four different target environments	599	Environments were grouped into four target sets (E1–E4)	1279 DArT§§	CIMMYT (Crossa et al., 2010)
Wheat Cornell	Winter wheat	Panel with family structure	Yield and heading date	374	One year (2009), two locations in New York state and two replicates per genotypes	1158 DArT	Cornell University (Heffner et al., 2011)
Wheat diallel	Winter wheat	Partial diallel, elite lines (eight crosses and five parents)	Yield, plant height, and thousand kernel weight	551	Three years, two to five trial locations per year in France, 18 replicates per genotype on average	319 SNP	Limagrain Europe

†By "panel" we mean a group of mostly unrelated lines.
‡SNP, single nucleotide polymorphism.
§CAP, Coordinated Agricultural Project.
¶SSR, simple sequence repeat.
#INRA, Institut National de la Recherche Agronomique (Paris, France).
††BLUE, best linear unbiased estimator.
‡‡GCA, general combining ability.
§§DArT, diversity array technology markers (Triticarte Pty. Ltd. Canberra, Australia).

with λ_1 and λ_2 shrinkage parameters. This double regularization generates a sparse model through the L_1 norm penalty and the L_2 part removes the limitation on the number of selected variables, encourages grouping effects, and stabilizes the L_1 path. This model was implemented using the R package glmnet (Friedman et al., 2008). This implementation performs a coordinate descent search for the lasso parameter. In addition we performed a grid search via cross-validation for the other shrinkage parameter controlling the relative amount of L_1 and L_2 penalties. The model with the minimum MSE was selected.

BayesB and its relative, BayesA, (Meuwissen et al., 2001) relax the assumption of common variance across marker effects made by RR-BLUP. The prior for marker effect j is a mixture distribution with β_j equal to 0 with probability π and, with probability $1 - \pi$, β_j is sampled from a normal distribution with mean 0 and variance σ^2_j. Finally, σ^2_j is sampled from a scaled inverse χ^2 with degrees of freedom ν and scale S^2. In the case of the original BayesB publication, π was set to 0.95. The BayesB model reduces to BayesA for $\pi = 0$. However, the computational demand of those original models limits their implementation even though simulation studies (Habier et al., 2007) stress their advantages over RR-BLUP. For this reason, BayesB was tested only on the smaller datasets (Barley 1 [Limagrain Europe, Riom, France], Bay × Sha, Diallel maize [Limagrain Europe], and Panel maize [Limagrain Europe]) with two chains of 10,000 iterations and 1000 for burn-in.

The weighted Bayesian shrinkage regression (wBSR) method (Hayashi and Iwata, 2010) is an expectation maximization (EM) algorithm for the BayesB model (Meuwissen et al., 2001). Preliminary testing of this model revealed that the initial convergence parameter set up for the algorithm was not adequate for some datasets and would generate unstable results. The authors of this model provided an updated version that allows the user to set the convergence parameter. The high computational efficiency of this algorithm allows a complete grid search to be performed on the prior parameters. The prior parameters searched were ν, the degree of freedom, and S^2, the scale parameter of the scaled inverse χ^2 distribution of the marker effect variance prior and π. Six hundred triplets of

prior parameters were tested for each dataset using a 10-fold cross-validation. The range of the grid search for the scale prior parameter was chosen following Gianola et al. (2009).

BayesCπ (Lorenz et al., 2010) assumes a common marker effect variance for all markers with nonzero effects, but rather than using a fixed π, it estimates π. The model was fitted with a single chain of 10,000 iterations, the first 1000 being discarded as burn-in. For Bayesian models the Markov chain Monte Carlo (MCMC) algorithm was used to obtain the posterior parameters and visually checked for convergence using the R package "coda" (Plummer et al., 2006).

Empirical Bayes (E-Bayes) (Xu, 2007) is a differential parameter shrinkage method for an oversaturated regression model. The original model was intended to incorporate linear combinations of all additive and pairwise epistatic effects among markers. As in BayesA, the prior for each β_j is assumed to follow a normal distribution with mean 0 and variance σ_j^2. The marker variance parameter σ_j^2 is further assumed to be inverse χ^2 distributed with degree of freedom τ and scale parameter ω. However, the E-Bayes algorithm does not require MCMC samplings to estimate the variance parameters σ_j^2. Instead a maximization algorithm is used to reduce computation time. The full model including additive and all pairwise epistatic effects contained too many effects. We therefore tested this model only with additive effects for all datasets. We optimized the model prediction by grid-searching multiple combinations of parameters (τ and ω) for each dataset. The parameter space tested ranged from −2 to −0.5 for τ and from 0 to 0.1 for ω.

The RKHS approach first uses a kernel function to convert the marker dataset into a set of distances between pairs of observations that results in a square matrix to be used in a linear model. Because RKHS regression does not assume linearity it might better capture nonadditive effects. The model can be formulated as

$$Y = W\boldsymbol{\mu} + \mathbf{K}_h\alpha + \varepsilon,$$

where $\boldsymbol{\mu}$ is a vector of fixed effects and ε is a vector of random residuals. The parameters α and ε are assumed to have independent prior distributions $\alpha \sim N(0, \mathbf{K}_h\sigma_\alpha^2)$ and $\varepsilon \sim N(0, I\sigma_\varepsilon^2)$, respectively. Matrix $\mathbf{K}_h$ depends on a reproducing kernel function with a smoothing parameter h, which measures the "genomic distance" between genotypes and can be interpreted as a correlation matrix. Parameter h controls the rate of decay of the correlation between genotypes.

Given h the RKHS regression is the same as a standard mixed-effects linear model. The mutual exchange of information between α-coefficients due to the nontrivial correlation structure induced by $\mathbf{K}_h$ is similar to the exchange of information between relatives induced by the genetic additive relationship matrix in the classical additive genetic model.

The kernel function we tested is a Gaussian kernel,

$$\mathbf{K}_h(x_i, x_j) = \exp(-hd_{ij}) = \exp(-\theta d_{ij}/k),$$

where d_{ij} is a marker-based distance between two individuals i and j. To decide which parameter combinations to use for optimal predictions for RKHS regression, we tested two distance methods built in R (R Development Core Team, 2010) to compute genetic distance: the squared Euclidean and the squared Manhattan distances. We also tested different values of θ and k. The θ values tested ranged from 0.1 to 10. The k value tested include (i) d_{median}, the sample median of d_{ij}, (ii) $d^2_{\max}$, the maximum value of squared distance of d_{ij}, and (iii) m, the number of markers genotyped. Only the Barley CAP dataset was used to optimize parameters for RKHS regression. For all other datasets, we used the same parameter combination that was optimized for the Barley CAP dataset

$$\theta/k = 2/d_{median}$$

with d_{ij} based on the Manhattan distance to predict trait values.

Machine-learning methods such as random forest (RF) regression (Breiman, 2001), support vector regression (SVR) (Drucker et al., 1997), and artificial neural networks (Gardner and Dorling, 1998) have been widely used in research and industrial settings. They could also be useful in the prediction of breeding values (Moser et al., 2009; González-Recio and Forni, 2011) and identification of causal polymorphisms (Bureau et al., 2005). Since those methods are nonparametric and the underlying theory behind them is quite different from the linear model for GS approaches described above, they may be able to capture different relationships between markers and phenotypes.

A RF is a collection of classification or regression trees grown on bootstrap samples of observations using a random subset of predictors to define the best split at each node. Different variables are used at each split in different trees. The RF prediction for an observation is computed by averaging the predictions over trees for which the given observation was not used to build the tree. This model was implemented using the R package "RandomForest" (Liaw & Wiener, 2002). We used the default setting of the function except for the number of trees, which was set to 1000, and the minimum size of terminal node as 50, as suggested by preliminary testing. We used the tuning function provided to optimize the number of variables randomly sampled at each split for each trait. This model will be referred to as support vector machine (SVM).

Support vector regression (Smola and Schölkopf, 2004) uses linear models to implement nonlinear regression by mapping the input space (the marker dataset) to a feature space of a different dimension (lower in the case of GS) using a nonlinear kernel function followed by linear regression in this feature space. The SVR simultaneously minimizes an objective function that accounts for both model complexity and the error in the training data. This model was implemented using the R package "e1071" (Dimitriadou et al., 2011). A linear kernel was used along with an epsilon-insensitive loss function. This means that during the model fitting, all the error up to the epsilon level is simply discarded from the model. A tuning function was used to optimize the level of epsilon and the cost parameter that weights the relative contribution of error and model complexity to the objective function.

The artificial neural network is a very broad class of models inspired by the structure and functions of biological neural networks. It has been demonstrated that the multilayer perceptron, a particular case of neural network, can be trained to approximate virtually any smooth, measurable function (Hornik, 1989). The multilayer perceptron is a system of simple interconnected neurons or nodes.

Each node sums its inputs multiplied by weights w_{ij}, linking nodes i and j and adding a node specific constant, the bias α_j. The output is then produced by applying an activation function

f_i. This activation function can be linear or nonlinear. The system is defined by an input layer with one neuron per input variable x_i (here for each of the N markers), which sends the data to intermediate hidden layers, and by an output layer made of one neuron per output variable y_k that receives input from the last hidden layer. In our case the output layer is made of a single neuron to output the GEBVs. Figure 1 gives a general example of a neural network.

This graphical example of a network is equivalent to the following function from input to output, with the subscript k indexing the output variables and U the number of nodes in the hidden layer.

$$y_k = f_k\left[\alpha_k + \sum_j^U w_{jk} f_j\left(\alpha_j + \sum_i^N w_{ij} x_i\right)\right]$$

The fitting of a neural network is thus controlled by the number of hidden layers, the number of neurons per hidden layer, the activation function, and the weights of each connection. The training procedure of such a neural network implies the determination of the individual weights and several techniques are possible. The general goal is to find a combination of weights that will result in the smallest error, by looking for a minimum point in a multidimensional error surface. This task can be done with a back-propagation algorithm, which uses a gradient descent approach to identify a minimum on the error surface. An additional parameter of the model fitting is the weight decay that penalizes large weights and thus large input to neurons. Note that the extreme case of a regression neural network with no hidden layers will be equivalent to a ridge regression. A description of the category of neural network described here can be found in chapter five of Ripley (1996). We chose to focus on a simple form of neural network called "single hidden layer feed-forward perceptron" in which the system has three layers, with only one hidden layer, as shown in Fig. 1. Feed-forward means that the nodes can be numbered so that all connections go one way from a lower node to one with a higher number. We chose to use a linear activation function. Although this algorithm has the capacity to handle highly nonlinear systems, with our choice of activation function, the system will only be linear. The model was implemented using the R libraries "nnet" (Venables and Ripley, 2002) and "e1071" (Dimitriadou et al., 2011). The model was optimized for the number of neurons in the hidden layer and the weight decay parameter. The number of iterations to fit each neural network was set to 200. In every case (for non-cross-validated and cross-validated predictions), the model was run 10 times and the computed GEBV averaged, to take into account dependency on initial weight parameters. This model will be referred to as neural network (NNET)

When it was readily possible to include covariates in the models, namely for the ridge regression methods, for BayesCπ, and for wBSR, the same analyses described above were performed with and without a covariate accounting for the population structure. The calculation of the covariate used is described below.

Prediction Accuracy and Cross-Validation

The predictive ability of the models was assessed using the Pearson correlation coefficient between the observations and the cross-validated GEBVs and will be referred to as the accuracy. Some publications define accuracy as the correlation between the GEBVs and observed phenotypic values divided by the square root of the heritability (Dekkers, 2007; Lorenzana and Bernardo, 2009). Using such a definition of accuracy introduces an additional error due to the heritability computation. In addition, the adjustment would be identical across all methods and would not contribute to differentiating them.

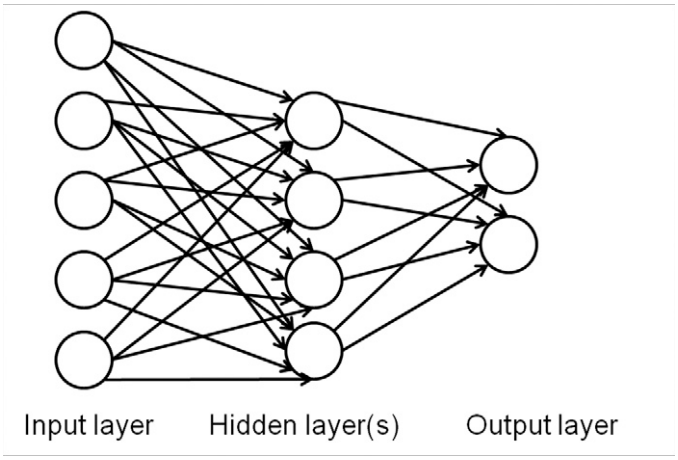

Figure 1. A generic feed-forward neural network with a single hidden layer.

To compute the accuracy, we used a 10-fold cross-validation. Each phenotypic dataset was randomly divided into 10 equal parts. Then the GEBVs for each fold were predicted by training the model on the nine remaining folds. The accuracy was computed in one step on the whole vector of predicted values. To take into account the identified population structure of our datasets, a stratified sampling was used in each of the identified subpopulations to ensure that each fold was representative of the entire dataset composition. For the diallels, we treated the different crosses as subpopulations. In the other cases, we used the R package mclust (Fraley and Raftery, 2002) to identify subpopulations by hierarchical clustering using a parameterized Gaussian mixture models. The Bayesian information criteria was used to identify the optimal number of subpopulations as well as the optimal clustering model to use.

To ensure an accurate comparison of models, the same cross-validation folds were used for each model. The non-cross-validated correlation was calculated as the correlation between the GEBVs obtained by using the whole dataset as a training population and the observed values on the training population. The difference between this non-cross-validated correlation and the accuracy was used as a measure of overfitting. For each dataset–trait combination we compared the GEBV estimates and the marker effect estimates between models. Marker effect estimates used in this comparison were computed for each model and dataset–trait combination using the whole dataset, as this would be the standard procedure in a GS application (use of all the data available to train the model). Significance of the differences in accuracies obtained with different models was tested using a binomial test for each pair of models using the accuracies obtained with each of the 18 traits as observations and considering the sign of the difference between accuracies; the null hypothesis is that the difference is not significant and then that the sign of the difference follows a Bernoulli distribution with parameter 0.5.

Accuracy within Subpopulations

The predictive ability of each of the tested GS models was also considered at the level of each subpopulation or cross. The homogeneity of variance among subpopulations was assessed with a Fligner–Killeen test (Conover et al., 1981) that is robust to nonnormality of the data. In addition, a significance test for the difference in accuracy among subpopulations was based on a randomization method as follows. The null hypothesis is that the genetic parameters came from a single statistical population. Define Var(X) as the variance of the accuracy measured across subpopulations. For each randomization k, randomly assign individuals to a population and calculate Var(X_k). The probability of observing Var(X) by chance alone is

$$p\text{-value} = \frac{1 + \text{number of randomization for which } \mathrm{Var}(X_k) \geq \mathrm{Var}(X)}{1 + n}$$

with n number of randomizations (Manly, 1991). We used 10,000 permutations to compute the p-values.

The relevance of subpopulation structure was investigated with the pairwise F statistics quantifying the differences in allele frequency among subpopulations (*Fst*), estimated with a jackknifed estimator, as well as with a test of significance of the subpopulation structure on the differentiation using the R package hierfstat (Goudet, 2005) with 1000 permutations. For all hypothesis testing, the Bonferroni correction for multiple testing was used.

Model Similarity

The similarities between models were also investigated through the use of clustering methods. For each of the 18 dataset–trait combinations, a matrix of Euclidean distances between GS models was calculated based on the cross-validated GEBVs. The GEBVs for each model were standardized to zero mean and unit variance before distance computation. Those distance matrices were then averaged (equal contribution of each dataset–trait combination) and used as an input for hierarchical clustering using the Ward criterion (i.e., based on the increase of variance of the cluster being merged during the tree building process). The tree built by equal contribution of each dataset (as opposed to dataset–trait combination) gave the same tree topologies. The similarities were also analyzed by considering separately the traits whose genetic architecture was known to be characterized by some major effect loci, such as plant height in wheat, flowering time in the biparental *Arabidopsis thaliana* (L.) Heynh. population, and the betaglucan content in barley (*Hordeum vulgare* L.). The average excess kurtosis of the marker effect distributions obtained from ridge regression, wBSR, the BL, and BayesCπ as described below was also used to confirm this separation.

The similarities identified between models were further investigated by analysis of the marker effect distribution for each model, using the excess kurtosis that is a measure of the "peakedness" of the distribution. (A normal distribution has an excess kurtosis of 0.) Higher kurtosis means that more of the variance is the result of few extremely deviant marker effect estimates. We hypothesized that for some models, high kurtosis could be linked with the high multicolinearity of the data. This hypothesis was tested using a nonparametric correlation test based on Spearman's rho between the observed kurtosis and the number of lines, number of markers, and a statistic measuring the number of uncorrelated variables in the model. To construct this latter statistic, we used the number of eigenvectors necessary to capture 95% of the variance on a principal component analysis (PCA) of the marker dataset as an indicator of the number of uncorrelated variables.

Model Combinations

Considering the large diversity of GS models, instead of identifying a single best performing model, it could be advisable to build predictors based on a combination of models to increase the prediction accuracy. Various procedures to combine models were tested using the cross-validated GEBVs. To avoid introducing overfitting in the combined predictor, in all cases it was constructed using the same cross-validation folds used to train the individual models, by building a combined predictor for each fold using the nine remaining folds as a training set. This procedure allowed us to make a direct comparison between the accuracies of the individual models and of the combined predictors. We tested the simple averaging of two to four models. We also used a simple least squares approach by regressing the cross-validated predictors on the observed data on each of the training sets. To obtain a more parsimonious model, we also used a backward stepwise model selection with the Aikake information criterion starting from the complete regression model described above. The R package MASS (Venables and Ripley, 2002) was used to carry out this procedure. The approach of stacked regression described by Breiman (1996a) was also used to build a combined predictor. This method is similar to the least square method described above, but with the regression coefficients constrained to be positive to account for the high colinearity between the predictors. Breiman (1996a) reported a decrease of 10% in the prediction error with this approach. This modified least square approach was implemented using the optim function in R with a box constraint (Byrd et al., 1994) on the sign of the regression coefficients. Finally, in the light of the difference in prediction accuracies across subpopulations, a modified stacked regression method was tested giving equal weight to each subpopulation in the least square equation instead of giving an equal weight to each individual.

In addition to combining different models we sought to improve the accuracy of single models using a technique known as bagging in the machine learning literature (Breiman, 1996b). This approach is the basis of the RF algorithm. For a given GS model it consists of generating training sets from the original dataset by sampling with replacement, with the size of the training set being equal to the original training set. The bagged predictor is then constructed by averaging the predictors obtained on the different training sets. Breiman (1996b) showed that bagging effectively improved prediction accuracy of an unstable learning algorithm where a small perturbation in the training set can cause significant changes in the predictions. In an attempt to increase accuracy at the subpopulation level, we tested both a uniform sampling on the dataset and an equal sampling at the subpopulation level to obtain more balanced training sets.

Finally, we tried an approach called boosting (Drucker, 1997), or AdaBoost in the machine learning literature, that was reported to be at least equivalent and in most cases superior to bagging in reducing the prediction error. In this approach, the model is trained repeatedly on the same sample. After each iteration, a measure of prediction error is computed for each individual. In the following iteration, the individuals with the highest

error are given more weight in the training of the model. Over iterations, patterns that are more difficult to predict are given more importance and different machines are better in different parts of the observation space. The different predictors are combined using the weighted median such that those predictors with a reduced error are given more importance. This technique has been initially developed for classification purposes but has received an extension to regression problems. There are different versions of the boosting algorithm for regression; in this study we used AdaBoost.R2 (Drucker, 1997). Even though this algorithm is not reported to be the best one, it has the advantage of not requiring the set up of additional parameters relative to the error function used to update the weights given to each genotype in the training population (Shrestha and Solomatine, 2006). A detailed presentation of the algorithm used can be found in Shrestha and Solomatine (2006). The original paper proposed several functions to compute the error, linear, squared, and exponential. All three functions were tested. One potential drawback from this approach is a sensitivity to noise and outliers as the reweighting is proportional to the prediction error.

To test both bagging and boosting, we used the BL with the same parameters as described above for the BL alone and in the same cross-validation setting as for the other models. Thus, for each fold, the nine remaining folds were used as a training set. As for the single models, we also computed the non-cross-validated correlation as an additional measure of overfitting.

All statistical procedures were executed using R (R Development Core Team, 2010). The executable for wBSR was obtained from the authors Hayashi and Iwata (2010).

RESULTS

Ridge Regression Models

The accuracies of the three different ridge regression methods were quite similar. Across the traits, the average of the accuracies was 0.55 for the BRR and 0.56 for ridge regression with grid search and the ridge regression using a mixed model to estimate the shrinkage factor. Across all traits, the median of the correlations between the cross-validated GEBVs obtained using the three different ridge regression methods was above 0.96. The comparison of marker effects for those ridge regression approaches also demonstrated their high similarities: the correlation between marker effects was above 0.99 for all traits except for betaglucan (correlation of 0.78). The kurtosis of the marker effect distribution was in the same range for each method with an average excess kurtosis of 1.41 for the grid search case and 1.55 for the two other methods. This means that the marker distribution was on average slightly less "peaked" for the grid search version of the ridge regression. In addition, the non-cross-validated correlations were similar between those methods for all traits. The computation time was considerably lower for the ridge regression using a mixed model: the grid search ridge regression took approximately half of the computation time of the Bayesian ridge and the ridge regression with a mixed model took only one third of the time required to do the ridge regression with a grid search.

BayesB and Weighted Bayesian Shrinkage Regression

The results of the grid search to optimize the prior parameters of wBSR revealed a wide range of accuracies from 0.48 for the average accuracy across traits with the best performing common combinations of prior parameters (Pi = 0.25, Nu = 9, S^2 = 0.05) to an average of 0.56 for the best traitwise performing combinations. The Pi parameter (prior probability that a marker will have a null effect) and the scale parameter of the prior are the most important parameters according to the complete grid search made on the different traits (600 sets of prior parameters tested). The best values of the scale parameters were between 0.001 and 10 with most of them close to the value of 0.043 used by Meuwissen et al. (2001). The best Pi value ranged from 0.01 to 0.99. Figure 2 presents heat maps made by averaging across traits the accuracies obtained with the grid search. Since no best set of parameters used a scale parameter greater than 0.1, the heat maps exclude all grid points with a scale parameter above 0.1. The black dots in Fig. 2 indicate the best set of prior parameters identified for each trait. It is interesting to notice that those dots are relatively scattered across the heat maps, visualizing the fact that no single parameter setting was best for all traits.

Overall, and on average, the set of prior parameters of BayesB (Pi = 0.9, Nu = 4.36, S^2 = 0.01) was approximately a good value. The average accuracy for the best set of prior parameters for each trait was 0.56 compared to 0.45 using the original prior parameter set of BayesB in wBSR. It was not computationally feasible to compute BayesB for the Wheat CIMMYT, Barley CAP, and Wheat Cornell datasets. For other datasets, the average accuracy for BayesB was 0.52 compared to 0.53 for wBSR with the prior parameter set of BayesB and 0.71 for the best set of prior parameters for each trait, as identified by the grid search. The difference between the non-cross-validated correlation and the accuracy, taken as a measure of overfitting (a lower value indicates less overfitting), was also favorable to wBSR: 0.08 for the best set of prior parameters for each trait for wBSR, 0.12 for wBSR with the prior parameter set of BayesB, and 0.18 for BayesB itself.

The sensitivity of wBSR to the marker order was tested for all datasets by randomizing the marker order in the design matrix. In all cases, the correlation between cross-validated GEBVs using different marker orders was above 0.98. For some datasets, however (Wheat diallel [Limagrain Europe] and Wheat Cornell), the correlation between marker effects themselves was only around 0.6. Averaging of marker effects from several runs with different marker order did not improve wBSR accuracy. This result suggested that the algorithm always captures the same signal, but if several markers are in high linkage disequilibrium (LD) with a QTL, the algorithm tended to pick the first marker entered in the model. Given that the focus of GS is on the GEBVs,

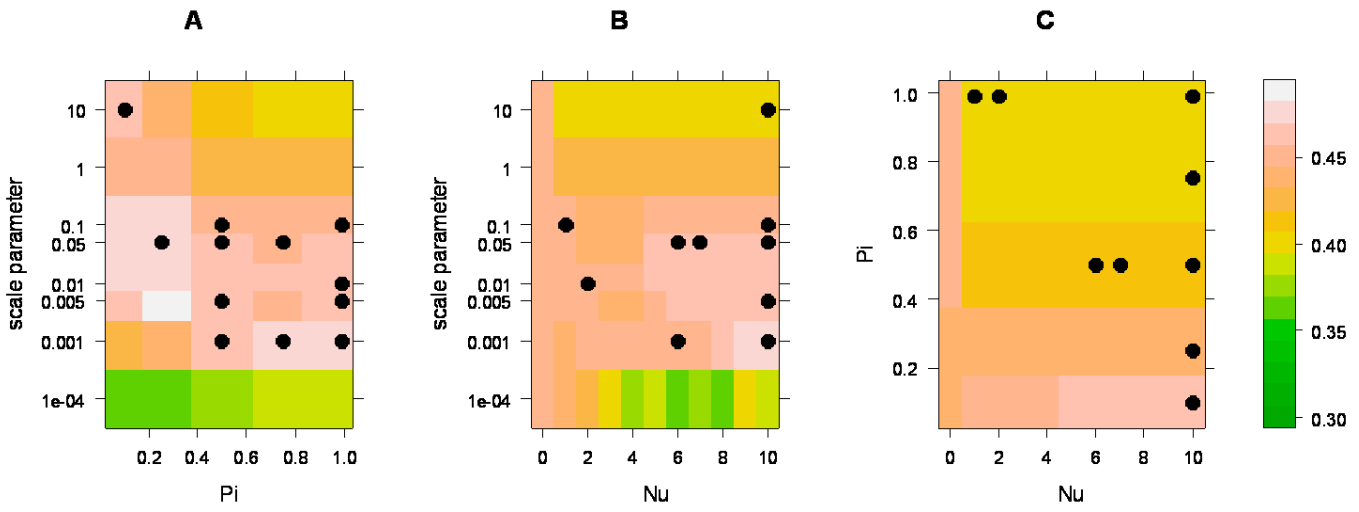

Figure 2. Heat maps summarizing accuracies for the grid search on weighted Bayesian shrinkage regression (wBSR) prior parameters. The accuracies were centered across the 600 triplets of parameters for each trait and averaged. The axes of the heat map correspond to the prior parameters for the marker effect variance. Nu, the degree of freedom, and the scale parameter are the parameters of the scaled inverse chi square distribution of the marker effect variance prior and Pi the prior proportion of loci with a null variance. Each cell of the heat maps is an average across the nonplotted parameter values. The scale parameter is plotted on a log scale (the set of prior parameter of BayesB is Pi = 0.9, Nu = 4.36, S^2 = 0.01). The black dots indicate for each trait the best set of prior parameters identified.

this is probably not an issue for the use of this algorithm for GS purposes. Similar findings were reported in the use of VBay, which is an EM algorithm equivalent to BayesCπ developed for genome-wide association studies approaches in humans (Logsdon et al., 2010). This finding prevented further direct comparison of marker effects between wBSR and other models. Distributions of marker effects could be compared, however. The average excess kurtosis of the marker effects distribution for BayesB was 38.2 compared to 19.42 for wBSR with BayesB prior and 8.76 for wBSR with an optimized prior. The correlation between GEBVs from BayesB and wBSR with the prior parameter set of BayesB was high and ranged from 0.77 to 0.95 except for the moisture trait in the Panel maize dataset where it was only 0.62.

Empirical Bayes Grid Search

The results of the grid search to optimize the prior parameter of E-Bayes revealed a wide range of accuracies from 0.46 for the average accuracy across traits with the best performing common combinations of prior parameters ($\tau = -0.5$, $\grave{u} = -0.5$) to an average of 0.54 for the trait wise best performing combinations. Figure 3 presents a heat map made by averaging across traits the accuracies obtained with the grid search. On average, the set of original set of prior parameters (i.e., $\tau = 2$ and $\omega = 2$ in the bottom left of the heat map) was close to the optimum, with an average accuracy of 0.46. This set of prior parameters corresponds to a flat (noninformative) prior.

Comparison of Accuracies and Overfitting

Table 2 presents the accuracy obtained for each trait and model tested. For the sake of clarity and considering the similarity of the three ridge regression models, only the ridge regression using a mixed model to estimate the shrinkage parameter is given here. Hereafter, "wBSR" and "E-Bayes" denote the optimized versions, with a different set of prior parameters for each trait. The last three lines of the table give the average accuracy, the average non-cross-validated correlation (providing a measure of the overfitting level for each model), and the average MSE.

Most models reached a very similar accuracy for a given trait. However, RKHS tended to outperform the other models in terms of accuracy. Support vector machine performed poorly on these datasets, even though the model was optimized for each trait for the cost and epsilon parameter SVM was the only method significantly different from all the other for the accuracy ($p < 0.05$) with Bonferronni correction for multiple testing. The others pairs of methods significantly different from each other ($p < 0.05$), with Bonferronni correction for multiple testing, were wBSR from the elastic net, RKHS from E-Bayes, elastic net, and the neural network. The performance of the elastic net model was slightly below that of ridge regression and the BL. The relative percentage of lasso penalty ranged from 0.7 to 1 (pure lasso) with an average of 0.92 across traits. Meanwhile, if we consider the difference between the non-cross-validated correlation and the accuracy as a measure of overfitting, E-Bayes, RKHS, SVM, and NNET are clearly overfitting much more than the other models. The MSE was computed on the scaled phenotypic data and cross-validated GEBVs centered and scaled by the phenotypic variance. This scaling ensured that the traits with a higher phenotypic variance were not weighted more heavily. Most models were rather similar but SVM and NNET performed

poorly in terms of MSE. As the data were centered and scaled with the phenotypic variance before MSE computation, it did not measure the bias between the cross-validated GEBVs and the phenotypic data but only the error and the difference in the level of shrinkage between models. As the average accuracy of NNET is similar to the best performing model, we can attribute the higher MSE on average to a higher variance of the cross-validated GEBVs, even when scaled by the phenotypic variance.

Comparison of Cross-Validated Genomic Estimated Breeding Values

The comparison of cross-validated GEBVs between models allowed estimation of the similarities and dissimilarities of the models. To determine whether model similarities were affected by genetic architecture, we analyzed separately traits that were believed to be influenced by major loci versus traits that were unlikely to be affected by major loci as described in the material and methods. The dendrogram topologies were extremely similar for these two categories, indicating that, at least at the crude level explored here, genetic architecture did not affect model similarity.

Figure 4 presents the hierarchical clustering tree obtained through the averaging of the distance matrix across all traits. Those results clearly showed the similarities in terms of GEBVs between the linear models represented

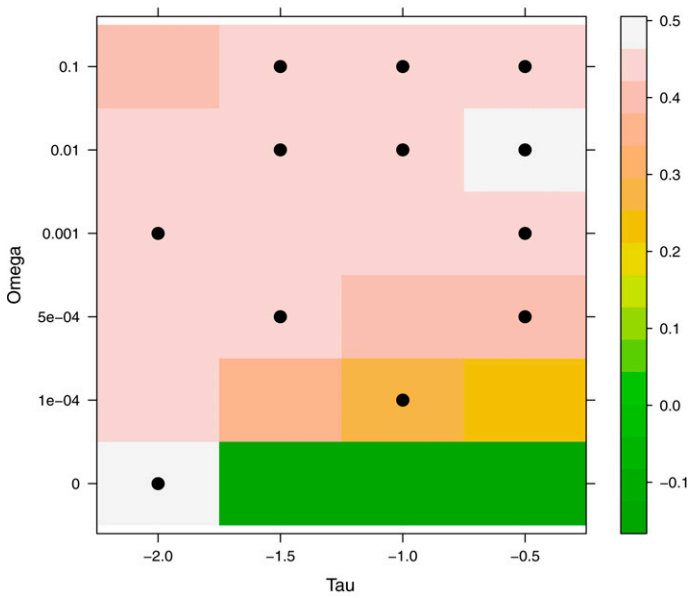

Figure 3. Heat map summary of the grid search on empirical Bayes (E-Bayes) prior parameters. The accuracies were centered across the 24 pairs of parameters for each trait and averaged. The axes of the heat map correspond to the prior parameters for the marker effect variance. Tau is the degree of freedom and Omega is the scale parameter of the marker effect variance prior. Each cell of the heat maps is an average across the nonplotted parameter values. (The original set of prior parameter is Tau = –2, Omega = 0). The omega parameter is plotted on a log scale. The black dots indicate for each trait the best set of prior parameters identified.

Table 2. Accuracy for each trait and model, average non-cross-validated correlation for each model, and average MSE for each model.

Dataset†	Trait‡	RR-BLUP§	BL	Elastic net	wBSR	BayesCπ	E-Bayes	RKHS	SVM	RF	NNET
Barley 1	Yield	0.53	0.55	0.52	0.53	0.53	0.53	0.6	0.43	0.56	0.51
Barley CAP	Betaglucan	0.57	0.57	0.57	0.57	0.57	0.57	0.6	0.35	0.55	0.54
Bay × Sha (Bay-0 × Shahdara)	FLOSD	0.82	0.82	0.83	0.83	0.82	0.82	0.83	0.8	0.85	0.82
	DM10	0.63	0.63	0.63	0.64	0.63	0.63	0.64	0.56	0.57	0.56
	DM3	0.4	0.39	0.40	0.4	0.39	0.4	0.41	0.33	0.38	0.35
Panel maize	Moisture	0.75	0.75	0.75	0.76	0.75	0.73	0.79	0.45	0.73	0.73
	Yield	0.63	0.63	0.61	0.63	0.63	0.59	0.64	0.32	0.6	0.59
Diallel maize	Moisture	0.74	0.74	0.72	0.73	0.74	0.73	0.75	0.56	0.61	0.72
	Yield	0.52	0.52	0.49	0.51	0.52	0.51	0.5	0.29	0.49	0.48
Wheat CIMMYT	YLD1	0.51	0.5	0.46	0.48	0.51	0.49	0.59	0.36	0.52	0.54
	YLD2	0.5	0.49	0.45	0.5	0.5	0.46	0.52	0.36	0.43	0.51
	YLD4	0.38	0.37	0.35	0.36	0.38	0.36	0.43	0.32	0.38	0.43
	YLD5	0.44	0.47	0.42	0.47	0.44	0.39	0.52	0.27	0.46	0.44
Wheat Cornell	Yield	0.36	0.35	0.37	0.37	0.34	0.26	0.28	0.22	0.36	0.36
	Height	0.45	0.44	0.41	0.44	0.44	0.41	0.55	0.37	0.46	0.45
Wheat diallel	Height	0.64	0.66	0.68	0.67	0.66	0.67	0.73	0.51	0.62	0.67
	TKW	0.6	0.57	0.59	0.6	0.59	0.59	0.68	0.41	0.54	0.65
	Yield	0.53	0.52	0.51	0.52	0.53	0.51	0.58	0.39	0.52	0.57
Average accuracy (cross-validated)		0.56	0.56	0.54	0.56	0.55	0.54	0.59	0.41	0.54	0.55
Average non-cross-validated correlation		0.77	0.79	0.75	0.77	0.77	0.93	0.99	0.89	0.76	0.85
Average MSE		0.67	0.67	0.69	0.68	0.68	0.76	0.64	1.36	0.72	10.54

†Barley 1, Limagrain Europe, Riom, France; Barley CAP (Barley Coordinated Agricultural Project, 2011); Bay Sha (Loudet et al. 2002); Panel maize, Limagrain Europe; Diallel maize, Limagrain Europe; Wheat CIMMYT (Crossa et al., 2010); Wheat Cornell (Heffner et al., 2011); Wheat diallel, Limagrain Europe.

‡Betaglucan, betaglucan content; FLOSD, flowering time in short days; DM10, dry matter in nonlimiting N conditions; DM3, dry matter in limiting N conditions; YLD1 to YLD5 refers to the yield traits reported in Crossa et al. (2010); TKW, thousand kernel weight.

§RR-BLUP, random regression best linear unbiased predictor; BL, Bayesian Lasso; wBSR, weighted Bayesian shrinkage regression; E-Bayes, empirical Bayes; RKHS, reproducing kernel Hilbert space; SVM, support vector machine; RF, random forest; NNET, neural network.

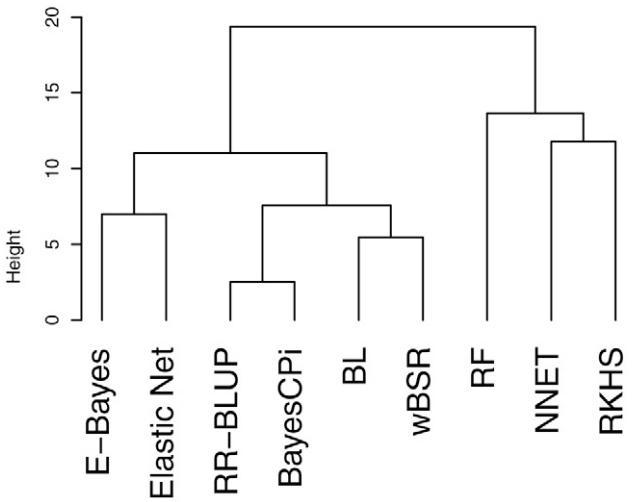

Figure 4. Hierarchical clustering of genomic selection (GS) models based on cross-validated genomic estimated breeding values (GEBVs), the height on the *y* axis refers to the value of the criterion associated with a particular agglomeration of models. RR-BLUP, random regression best linear unbiased predictor; BL, Bayesian Lasso; wBSR, weighted Bayesian shrinkage regression; RF, random forest; NNET, neural network; RKHS, reproducing kernel Hilbert space.

by ridge regression and the hierarchical Bayesian methods and the distinctness of nonparametric methods such as RF, neural network, and RKHS regression. Note too that while the nonparametric methods cluster with each other, they are all quite different, with deep divisions in the clustering between each method. Support vector machine was not used in this analysis as its poor prediction performance would have clustering difficult to interpret. This analysis also showed the strong similarity between ridge regression and BayesCπ. The similarities between wBSR and the BL is also interesting as wBSR was grid searched for the optimal prior parameters but the BL was not. It is also interesting to note that the elastic net clustered with E-Bayes despite being a combination of lasso and ridge regression penalty.

Comparison of Marker Effect Distributions

The distribution of the excess kurtosis of marker effects for each GS model that estimates marker effects was studied across the 18 dataset–trait combinations. The ridge regression marker effect distribution rarely departs from a normal distribution excess kurtosis (0), whereas other models such as the BL, wBSR and to some extent BayesCπ displayed significant differences in the marker effect distribution according to the trait. This suggests that Bayesian learning was taking place for the BL, wBSR, and to some extent for BayesCπ. Empirical Bayes and the elastic net performed differently, which is consistent with the clustering results and were characterized by an extremely high kurtosis. This is consistent with the variable selector properties of the elastic net.

The relationships between the excess kurtosis of different model marker effect distributions were investigated (Fig. 5). It is important to note that although the linear correlation was presented here, the excess kurtosis was not a linear function of the marker distribution. This figure revealed a striking behavior of E-Bayes and elastic net compared to the other models. We expected a significant correlation across trait–dataset combinations between kurtosis of different models. High and significant correlations were observed for ridge regression, BayesCπ BL, and wBSR but not for E-Bayes or the elastic net. As E-Bayes seemed to be characterized by more overfitting than the other models, we investigated the impact of the number of lines, number of markers, and a measure of the number of uncorrelated variables in the marker dataset on the number of PCA axes needed to capture 95% of the variance. For all models except E-Bayes and elastic net, the Spearman correlation between these variables and the excess kurtosis was not significant (p-values > 0.4). For E-Bayes the correlations were significant, with p-values of 0.011, 0.04, and 0.009 for the number of lines, the number of markers, and the number of uncorrelated variables, respectively. A multiple regression with these variables captured 38% of the variance of the E-Bayes marker effect distribution excess kurtosis. These correlations constitute evidence that E-Bayes did not handle highly multidimensional data well and tended to capture more noise than the other models. For the elastic net, the correlations were also significant with p-values of 0.0001, 0.09, and 0.02 for the number of lines, the number of markers, and the number of uncorrelated variables, respectively. A multiple regression with these variables captured 72% of the variance of the elastic net marker effect distribution excess kurtosis. This can be related to the formulation of the lasso that can retain only as many variables as observations. However, the elastic net performed correctly in terms of accuracy whereas E-Bayes did not.

Prediction in Each Subpopulation

Our clustering approach revealed no genetic structure in the Bay × Sha and Panel maize datasets.

Genomic selection accuracy was strongly affected by subpopulation (Table 3). This observation was true across all GS models: all models tended to be better for some subpopulations than for others with some trait–dataset combinations showing extremely high differences in accuracy. For example, in the Wheat Cornell yield dataset we distinguished six subpopulations, one of which had an accuracy of 0.7 and two that had an accuracy of 0.0. For the other traits and datasets, the differences in accuracy were less striking but in most cases the accuracy varied by at least twofold between the best and worst predicted subpopulations, even for the diallel design. Across all models and traits the standard deviation of accuracy between subpopulations ranged from 0.05 to 0.3.

We considered three hypotheses to explain the subpopulation effect on GS accuracy. First, subpopulations with

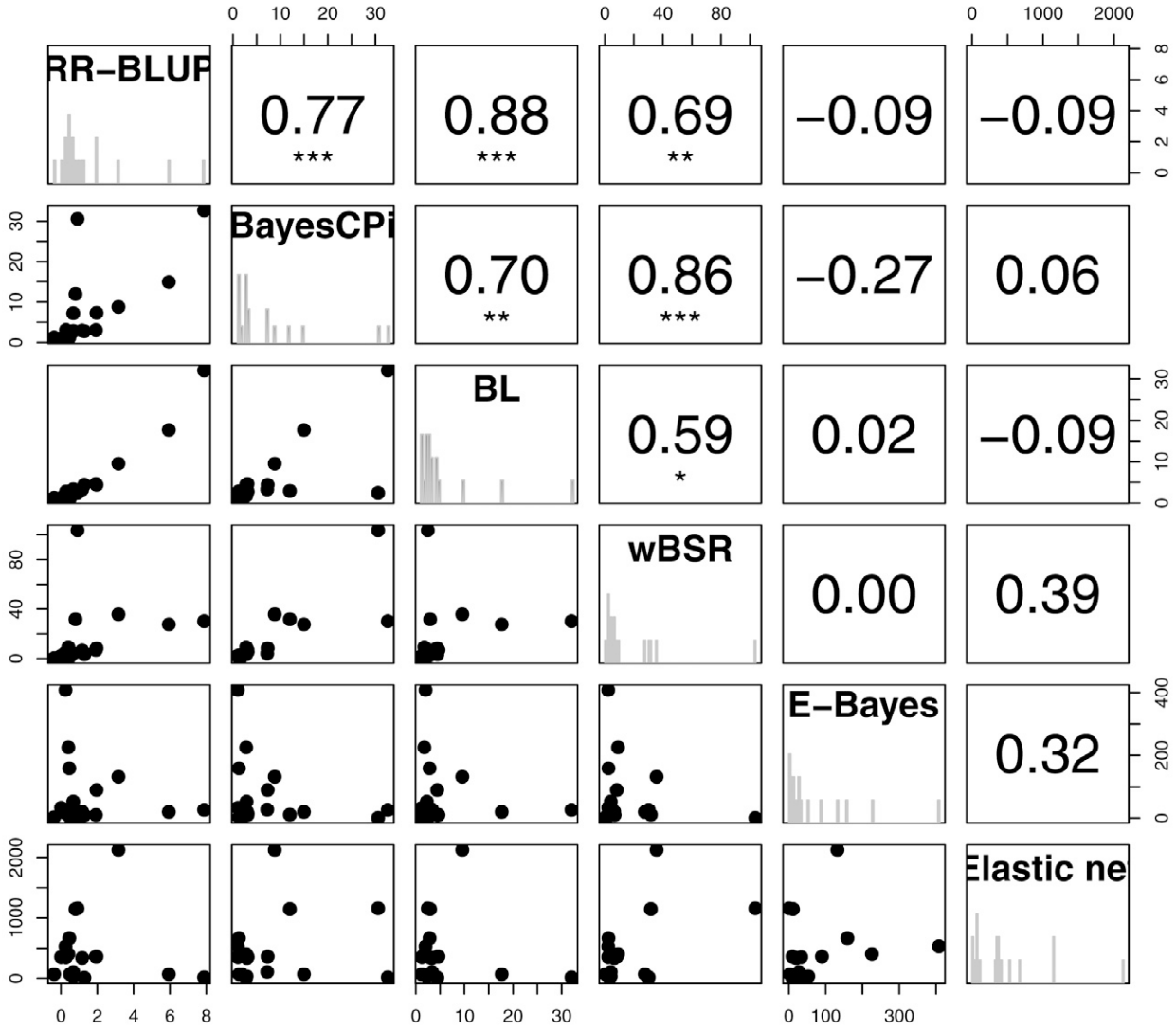

Figure 5. For each model, histogram of the marker effect distribution excess kurtosis on the diagonal, scatter plots comparing two models below the diagonal (each point represents one trait–dataset combination), and Spearman correlation between models above the diagonal with the significance level of the correlation based on Spearman's rho. (*Significant at the 0.05 probability level; **Significant at the 0.01 probability level; ***Significant at the 0.001 probability level. RR-BLUP, random regression best linear unbiased predictor; BL, Bayesian Lasso; wBSR, weighted Bayesian shrinkage regression.

Table 3. Summary of the subpopulation results. The accuracies reported here are from the Bayesian Lasso.

Dataset†	Trait‡	Number of groups	Smallest group size	Biggest group size	Minimum accuracy	Maximum accuracy	Total accuracy	*p*-value of subpopulation effect on accuracy	Fligner-Killeen test for phenotypic variance homogeneity
Barley CAP	Betaglucan	6	61	285	0.44	0.64	0.57	1	0.69
Wheat CIMMYT	YLD1	7	38	161	0.32	0.53	0.5	1	1
	YLD2				0.4	0.59	0.49	1	1
	YLD4				0.26	0.44	0.37	1	1
	YLD5				0.29	0.57	0.47	1	1
Wheat Cornell	Yield	6	38	95	–0.11	0.7	0.35	1.08×10^{-2}	0.08
	Height				0.21	0.45	0.44	1	2.73×10^{-5}
Wheat diallel	Height	8	32	82	0.11	0.69	0.66	2.40×10^{-3}	2.82×10^{-14}
	TKW				0.13	0.62	0.6	3.84×10^{-2}	0.49
	Yield				0.01	0.55	0.52	2.40×10^{-2}	1.09×10^{-2}
Diallel maize	Moisture	5	48	114	0.38	0.8	0.74	1.20×10^{-3}	0.53
	Yield				0.18	0.63	0.51	0.34	1

†Barley CAP (Barley Coordinated Agricultural Project, 2011); Wheat CIMMYT (Crossa et al., 2010); Wheat Cornell (Heffner et al., 2011); Wheat diallel, Limagrain Europe, Riom, France.
‡Betaglucan, betaglucan content; YLD1 to YLD5 refers to the yield traits reported in Crossa et al. (2010), Bonferroni correction for multiple testing; TKW, thousand kernel weight.

higher phenotypic variance might also have higher genetic variance and strongly influence the models, which would in turn lead them to have higher accuracy. Second, subpopulations that were larger would have more individuals in the training population and would have higher accuracy. Third, subpopulations with higher average pairwise *Fst* values would be more genetically unrelated to the population as a whole and therefore have lower accuracy. We tested the three hypotheses simultaneously using multiple regression of subpopulation accuracy on the three variables within each dataset–trait combination but no variable showed a consistently significant or suggestive relationship with accuracy using a Bonferroni correction for multiple testing. However, the two traits that displayed a significant difference in accuracy despite a nonsignificant heterogeneity of variance, the moisture content trait for the Diallel maize dataset and the thousand kernel weight trait for the Wheat diallel dataset, had significant *p*-values for the effect of the *Fst* before Bonferroni correction (0.06 and 0.08 respectively).

Use of Structure Covariate

The fraction of phenotypic variance explained by the structure covariates we used was below 5% for all traits considered and models except for Barley CAP with wBSR, in which it reached 14%, and the Diallel maize with BayesCπ, in which it reached 18 and 16% for moisture content and grain yield, respectively. The use of a structure covariate did not improve the accuracy inside each subpopulation compared to a model without a covariate (data not shown). This result suggested that the GS models are able to capture the subpopulation structure information most of the time and that this information can contribute to a significant portion of the accuracy.

Combinations of Models

The various differences observed between GS models suggest that complementarities exist between them that could be used to improve accuracy. Nevertheless, in most cases, combining different models did not result in a gain in accuracy. The only gains observed were for flowering time in Bay × Sha where the accuracy went from 0.85 with the best single method to 0.9 with a combination and the Wheat Cornell yield dataset where the accuracy went from 0.36 to 0.39. The accuracy at the subpopulation level was not improved, even with the modified version of the stacked regression estimator designed to favor the accuracy at the subpopulation level.

Bagging and Boosting

The results of bagging applied to the BL did not bring any additional accuracy gain. The average accuracy reached with the BL alone was 0.56, while with bagging the average accuracy dropped to 0.28 and the non-cross-validated correlation increased from 0.78 to 0.84. However, important differences in the impact of bagging were observed between the structured and nonstructured datasets. Bagging minimally reduced accuracy in nonstructured datasets (to 0.49 from 0.56) whereas in structured datasets bagging reduced the average accuracy to 0.01. The use of an alternative bagging strategy to bootstrap samples equally in each of the subpopulations, however, did not bring any improvement of the accuracy.

The boosting of the BL did not bring an improvement of the accuracy for either the median or the average approach used to combine the different predictors or for any of the loss functions tried. The average accuracy was below 0.1 for the linear and exponential loss functions and was equal to 0.2 for the squared loss function. Here again, the nonstructured datasets were better predicted than the structured ones. For none of the traits considered did bagging or boosting bring an increase in accuracy.

DISCUSSION

One of the key results of this study was that despite similar average accuracies between most of the models tested, there were major differences between them in terms of cross-validated GEBVs and marker effects.

Choice of a Genomic Selection Model for Plant Breeding

An optimal GS method should provide the highest accuracy possible, limit overfitting on the training dataset, and be based as much as possible on marker-QTL LD rather than on kinship (Habier et al., 2007). Moreover, such methods must be easy to implement, reliable across a wide range of traits and datasets, and computationally efficient. To be implemented, it should be possible to run the models overnight for the datasets we used. Model sparsity has been advocated as a key criterion for method selection. Sparsity can be achieved in two ways. First, it could arise from the elimination of markers with small effects from the model. This way is not favorable because, for polygenic traits, small or partial-effect markers do explain some true genetic variance not captured by large-effect markers. Second, it could arise from the capacity of the method to ensure that markers in strong LD with large QTL can capture their full effect rather than allowing the effect to be distributed over a number of markers. This way is favorable and might be observed by measuring excess kurtosis. These two ways are not mutually exclusive.

These general guidelines would lead to the recommendation to use RR-BLUP with a mixed model, the BL for its versatility, and wBSR. The elastic net performed well and it produced extremely sparse models, much more so than its Bayesian counterparts. From a breeding point of view, a nonsparse model could be more favorable: with more markers being selected by the model, more time will be required to reach fixation. In addition, Legarra et al. (2010) stressed that conditional expectations are optimal

for selection (Gianola and Fernando, 1986). Conditional expectations of the GEBVs based on the observations maximize expected selection response based on truncation selection via maximization of the correlation between predictor and predictand. These can be obtained through the BL but not with the regular Lasso or elastic net.

The use of BayesCπ cannot be recommended considering the extremely high similarities with RR-BLUP and the increased computation time. The high overfitting observed with E-Bayes as well as the observation that excess kurtosis was driven by marker colinearity suggested that this model should not be used in its current form. The high overfitting observed with the neural network approach would suggest that this model should not be used for GS at this time. In addition, the high computing requirement, mainly because of the model optimization step for training the neural network, also precluded its use for GS.

The case of RKHS regression is more difficult because, even though the model was overfitting, its accuracy was higher, indicating that the model was capturing both more genetic signal and more noise than the other models. This problem might be addressed by the use of different kernels and distance functions. An advantage of RKHS regression is that it is performed on the individual rather than the marker space. Thus, it is feasible to implement RKHS regression even if the dataset segregates for millions of markers, as would be the case in species where LD decays rapidly.

Despite overall good results and a reasonable computing time, the RF should be used with caution considering that this is a new method for GS. However, the apparent distinctness of this method and its potential to capture nonadditive effects, compared to the more classical approaches, should encourage more development.

Kurtosis of Marker Effects

The variation of the excess kurtosis is of importance as it signals if a model was able to adjust the marker effect distribution to the distribution of the QTL effects. For a given level of accuracy, it seems reasonable to favor a model whose marker effects are closer to the distribution of the QTL effects. A variable excess kurtosis is also an indirect indicator of the basis of the accuracy of a given model. If excess kurtosis varies with the traits, it suggests that an important part of the accuracy is based on LD marker-QTL association rather than on kinship.

Need for Further Analysis on the Basis of Accuracy for the Best Models Selected

These results would need to be confirmed using an approach similar to Habier et al. (2007) to identify the basis of predictive ability of models. If the predictive ability of a given model is based mainly on kinship, it will decrease much faster than if the predictive ability of the model is based on LD between markers and QTL. In addition, if a model is based on kinship rather than marker-QTL LD, the increase in inbreeding due to the application of a GS scheme using such model will be much faster. The simulation results of Habier et al. (2007) suggested that the accuracy of Bayesian methods such as BayesB (Meuwissen et al., 2001) would be based more on marker-QTL LD than on kinship, while that of ridge regression (RR-BLUP) is based mainly on kinship. The results of simulations (Long et al., 2011) suggest that most of the BL accuracy is due to LD marker-QTL.

Population Substructure

The large difference in accuracy observed in some cases between subpopulations in this study deserves additional analysis to uncover the basis for those differences. This cannot be explained by an uneven sampling of the cross-validation folds because the sampling approaches we used ensured that each fold was representative of the total dataset composition. All models were similar in terms of their differences in accuracy among the subpopulations. However, on a trait-by-trait basis, not all models performed the same in the subpopulations. The small number of dataset–trait combinations considered precluded broader conclusions on this observation. Clearly more investigation is needed to uncover the basis for those differences in accuracy. We observed that subpopulation accuracy differences were trait dependent for the same marker dataset (e.g., yield versus height for the Wheat Cornell dataset). Furthermore, the distribution of the polymorphism information content values, minor allele frequencies, and marker and individual call rates were roughly similar in the subpopulations. Together, these observations suggest that subpopulation accuracy differences could be caused by differences in the genetic determination of the trait in each subpopulation as well as by differences in phenotypic variance. For the Wheat Cornell dataset, the *Fst* values between the well predicted subpopulations and the poorly predicted subpopulations were somewhat higher than the other pairwise *Fst* values, suggesting that they were more differentiated, but such elevated *Fst* values were not found in other traits (data not shown). Excluding the poorly predicted subpopulations from the training set did not affect the cross-validated accuracy in the remaining part of the dataset (data not shown). In addition, using only the best predicted subpopulation did not result in accuracy as high as with the complete dataset for those populations. Thus, even data from subpopulations that are poorly predicted contribute beneficially to prediction accuracies. With the data available, it was not possible to clearly distinguish what part of the gain in accuracy was associated with the increase in the training population size and what part was due to the use of a more diverse training dataset with more recombination events than in any single subpopulation.

For only two traits (moisture content for the Diallel maize dataset and thousand kernel weight for Wheat

diallel dataset), the difference in phenotypic variance between subpopulations was not significant while the difference in accuracy was significant. For both traits, the multiple regression approach only allowed us to suggest a negative correlation between *Fst* and the accuracies.

Overall, these observations suggest that the difference in accuracy cannot be explained only by a difference in phenotypic variance in subpopulations but rather by difference in genetic architecture between subpopulations. This difference in genetic architecture is also more likely to exist when two given subpopulations are more unrelated than others, which would account for the observed *Fst* pattern. The nonsignificance of the multiple regression approaches precluded drawing a strong conclusion on the origin of those differences in accuracy.

Given that we do not understand the basis for differences among subpopulation accuracies, it seems necessary when implementing a GS model to focus not only on the overall accuracy but also on the accuracy at the level of each subpopulation as an additional check of model accuracy. This result has potentially wide ranging implications. For example, differences in accuracy could affect the rate of inbreeding generated by GS. If only some of the subpopulations of a breeding program are predicted well by a GS model, it may lead to preferential selection from those subpopulations and to the loss of diversity represented by the other subpopulations as most of the candidates lines identified by the GS model and confirmed in the field will originate from those well predicted subpopulations. We do not argue that differences in accuracy among subpopulations should preclude GS in structured populations and restrict it to biparental populations: loss of genetic diversity is also a risk in biparental GS and could even be greater because of the already reduced genetic diversity within any given cross.

Combination of Models

We were disappointed by the efficiency of the combination of predictor approaches that we tested. The lack of gain in accuracy is interesting in and of itself as it suggests that all models tested capture the same signal but in different ways, as shown by the differences between GEBVs. As discussed above, the differences in the capture of the signal, for example through kinship or marker-QTL LD, have important implications. This is additional evidence supporting the use of a few models based more on marker-QTL LD than on kinship as both will capture the same signal but in ways that may have different consequences for successful breeding.

As our study was unable to identify an all-purpose model or combination of models, we would suggest that for implementation in breeding programs the BL or wBSR with a grid search should be used. Results from RF seem promising but need more study with simulated datasets to better understand the genetic basis of the accuracy with this model (kinship or marker QTL LD). As this model does not produce marker effects it was not possible to investigate that point by studying the variation of the excess kurtosis of the marker effects distribution across traits. In addition, RKHS could potentially capture nonadditive relationships but the predictions obtained would not be GEBVs.

Bagging and Boosting

The lack of gain of accuracy by the use of bagging and boosting is also interesting as an indication that by a simple use of the BL model we already reach a plateau in terms of accuracy. For the boosting, this indicates that the poorly predicted individuals do not carry any additional genetic signal that can be effectively captured by the BL. Breiman (1996b) and Drucker (1997) acknowledge that neither bagging nor boosting can transform a poor predictor into a good one in all cases. In addition, those approaches were mainly developed for so called "weak learners," that is, predictors that are only weakly correlated with the true value. We are not sure that this definition applies to the BL. However, this definition is quite arbitrary. In addition, recent work on various boosting algorithms applied to classification (Long and Servedio, 2009) demonstrated that convex potential boosting algorithms such as AdaBoost are sensitive to noise in real datasets. Thus, our study can only conclude that bagging and boosting of the BL are not useful for GS. However, this approach could be useful to enhance the predictive ability of simpler models as reported by González-Recio et al. (2010).

Acknowledgments

The authors thank P. Flament, S. Chauvet, and all the Limagrain Europe biostatistics team for their helpful suggestions. The USDA-NIFA-AFRI provided grant support (award numbers 2009-65300-05661 and 2011-68002-30029). Additional funding for this research was provided by USDA-NIFA National Research Initiative CAP grant No. 2005-05130 and by Hatch 149-402. Part of this work was carried out by using the resources of the Computational Biology Service Unit at Cornell University, which is partially funded by Microsoft Corporation.

References

Barley Coordinated Agriculture Project. 2011. Introduction to project. Available at http://www.barleycap.org (verified 26 Oct. 2011). Univ. of Minnesota, St. Paul, MN.

Breiman, L. 1996a. Stacked regressions. Mach. Learn. 24:49–64.

Breiman, L. 1996b. Bagging predictors. Mach. Learn. 24:123–140.

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32. doi:10.1023/A:1010933404324

Bureau, A., J. Dupuis, K. Falls, K.L. Lunetta, B. Hayward, T.P. Keith, and P. Van Eerdewegh. 2005. Identifying SNPs predictive of phenotype using random forests. Genet. Epidemiol. 28:171–182. doi:10.1002/gepi.20041

Byrd, R.H., P. Lu, J. Nocedal, and C. Zhu. 1994. A limited-memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. 16:1190–1208. doi:10.1137/0916069

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K.A. Weigel, and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385. doi:10.1534/genetics.109.101501

Conover, W.J., M.E. Johnson, and M.M. Johnson. 1981. A comparative study of

tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics 23:351–361. doi:10.2307/1268225

Crossa, J., G. de Los Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V.N. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. doi:10.1534/genetics.110.118521

Dekkers, J.C.M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. 124:331–341. doi:10.1111/j.1439-0388.2007.00701.x

Dimitriadou E., K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. 2011. e1071: Misc functions of the department of statistics (e1071), TU Wien. R package version 1.6. Available at http://CRAN.R-project.org/package=e1071 (verified 26 Oct. 2011). R Foundation for Statistical Computing, Vienna, Austria.

Drucker, H. 1997. Improving regressors using boosting techniques. p. 107–115. *In* D.H. Fisher Jr (ed.) Proc. 14th Int. Conf. Machine Learning, Nashville, TN. 8–12 July 1997. Morgan Kaufmann, San Mateo, CA.

Drucker, H., C.J.C. Burges, L. Kaufman, A.J. Smola, and V. Vapnik. 1997. Support vector regression machines. Adv. Neural Information Processing Syst. 9:155–161.

Fraley, C., and A.E. Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 97:611–631. doi:10.1198/016214502760047131

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1):1.

Gardner, M.W., and S.R. Dorling. 1998. Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. Atmos. Environ. 32:2627–2636. doi:10.1016/S1352-2310(97)00447-0

Gianola, D., G. de los Campos, W.G. Hill, E. Manfredi, and R.L. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363. doi:10.1534/genetics.109.103952

Gianola, D., and R.L. Fernando. 1986. Bayesian methods in animal breeding theory. J. Anim. Sci. 63:217–244.

González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet. Sel. Evol. 43:7. doi:10.1186/1297-9686-43-7

González-Recio, O., K.A. Weigel, D. Gianola, H. Naya, and G.J.M. Rosa. 2010. L2-boosting algorithm applied to high-dimensional problems in genomic selection. Genet. Res. 92:227–237. doi:10.1017/S0016672310000261

Goudet, J. 2005. A package for R to compute and test hierarchical F -statistics. Molecular Ecology Notes 5:184–186.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

Hayashi, T., and H. Iwata. 2010. EM algorithm for Bayesian estimation of genomic breeding values. BMC Genet. 11:1–9. doi:10.1186/1471-2156-11-3

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen. 4:1–11. doi:10.3835/plantgenome2011.12.0001

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1–12. doi:10.2135/cropsci2008.08.0512

Hornik, K. 1989. Multilayer feedforward networks are universal approximators. Neural Netw. 2:359–366. doi:10.1016/0893-6080(89)90020-8

Institut National de la Recherche Agronomique (INRA). 2007. Web Service VNAT. Study of the natural variation of *Arabidopsis thaliana*. Available at http://dbsgap.versailles.inra.fr/vnat/ (verified 24 Oct. 2011). INRA, Paris, France.

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. Brief. Funct. Genomics Proteomics 9:166–177. doi:10.1093/bfgp/elq001

Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723. doi:10.1534/genetics.107.080101

Legarra, A., C. Robert-Granié, P. Croiseau, F. Guillaume, and S. Fritz. 2010. Improved lasso for genomic selection. Genet. Res. 93:77–87. doi:10.1017/S0016672310000534

Liaw, A., and M. Wiener. 2002. Classification and regression by random forest. R News 2(3):18–22.

Logsdon, B.A., G.E. Hoffman, and J.G. Mezey. 2010. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. BMC Bioinform. 11:58. doi:10.1186/1471-2105-11-58

Long, N., D. Gianola, G.J.M. Rosa, and K.A. Weigel. 2011. Long-term impacts of genome-enabled selection. J. Appl. Genet. 52(4):467.

Long, P.M., and R.A. Servedio. 2009. Random classification noise defeats all convex potential boosters. Mach. Learn. 78:287–304. doi:10.1007/s10994-009-5165-z

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, and J.-L. Jannink. 2011. Genomic selection in plant breeding: Knowledge and prospects. Adv. Agron. 110:77–123. doi:10.1016/B978-0-12-385531-2.00002-5

Lorenz, A.J., M.T. Hamblin, and J.-L. Jannink. 2010. Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. PLoS ONE 5:e14079. doi:10.1371/journal.pone.0014079

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120:151–161. doi:10.1007/s00122-009-1166-3

Loudet, O., S. Chaillou, C. Camilleri, D. Bouchez, and F. Daniel-Vedele. 2002. Bay-0 × Shahdara recombinant inbred line population: A powerful tool for the genetic dissection of complex traits in *Arabidopsis*. Theor. Appl. Genet. 104:1173–1184. doi:10.1007/s00122-001-0825-9

Manly, B.F.J. 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall/CRC, London, UK.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Moser, G., B. Tier, R.E. Crump, M.S. Khatkar, and H.W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41:56. doi:10.1186/1297-9686-41-56

Park, T., and G. Casella. 2008. The Bayesian Lasso. J. Am. Stat. Assoc. 103:681–686. doi:10.1198/016214508000000337

Pérez, P., G. de los Campos, J. Crossa, and D. Gianola. 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Gen. 3:106. doi:10.3835/plantgenome2010.04.0005

Plummer M., N. Best, K. Cowles, and K. Vines. 2006. CODA: Convergence diagnosis and output analysis for MCMC. R News 6:7–11.

R Development Core Team. 2010. R: A language and environment for statistical computing. Available at http://www.r-project.org (verified 18 Oct. 2011). R Foundation for Statistical Computing, Vienna, Austria.

Ripley, B.D. 1996. Pattern recognition and neural networks. Cambridge Univ. Press, Cambridge, UK.

Shrestha, D.L., and D.P. Solomatine. 2006. Experiments with AdaBoost.RT, an improved boosting scheme for regression. Neural Comput. 18:1678–1710. doi:10.1162/neco.2006.18.7.1678

Smola, A.J., and B. Schölkopf. 2004. A tutorial on support vector regression. Stat. Comput. 14:199–222. doi:10.1023/B:STCO.0000035301.49549.88

Venables, W.N., and B.D. Ripley. 2002. Modern applied statistics with S. Fourth edition. Springer, New York, NY.

Xu, S. 2007. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63:513–521. doi:10.1111/j.1541-0420.2006.00711.x

Xu, S., and Z. Hu. 2010. Methods of plant breeding in the genome era. Genet. Res. 92:423–441. doi:10.1017/S0016672310000583

Yi, N., and S. Xu. 2008. Bayesian LASSO for quantitative trait loci mapping. Genetics 179:1045–1055. doi:10.1534/genetics.107.085589

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. B. Stat. Methodol. 67:301–320.