Comparison Between Linear and Non-parametric Regression Models for Genome-Enabled Prediction in Wheat

Paulino Pérez-Rodríguez,*,[1] Daniel Gianola,[†] Juan Manuel González-Camacho,* José Crossa,[‡] Yann Manès,[‡] and Susanne Dreisigacker[‡]

*Colegio de Postgraduados, Montecillo, Texcoco 56230, México, [†]Departments of Animal Sciences, Dairy Science, and Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin 53706, and [‡]Biometrics and Statistics Unit and Global Wheat Program, International Maize and Wheat Improvement Center (CIMMYT), 06600 Mexico, D.F., México

ABSTRACT In genome-enabled prediction, parametric, semi-parametric, and non-parametric regression models have been used. This study assessed the predictive ability of linear and non-linear models using dense molecular markers. The linear models were linear on marker effects and included the Bayesian LASSO, Bayesian ridge regression, Bayes A, and Bayes B. The non-linear models (this refers to non-linearity on markers) were reproducing kernel Hilbert space (RKHS) regression, Bayesian regularized neural networks (BRNN), and radial basis function neural networks (RBFNN). These statistical models were compared using 306 elite wheat lines from CIMMYT genotyped with 1717 diversity array technology (DArT) markers and two traits, days to heading (DTH) and grain yield (GY), measured in each of 12 environments. It was found that the three non-linear models had better overall prediction accuracy than the linear regression specification. Results showed a consistent superiority of RKHS and RBFNN over the Bayesian LASSO, Bayesian ridge regression, Bayes A, and Bayes B models.

KEYWORDS
GenPred
Shared data resources

Genome-enabled prediction of complex traits based on marker data are becoming important in plant and animal breeding, personalized medicine, and evolutionary biology (Meuwissen *et al.* 2001; Bernardo and Yu 2007; de los Campos *et al.* 2009, 2010; Crossa *et al.* 2010, 2011; Ober *et al.* 2012). In the standard, infinitesimal, pedigree-based model of quantitative genetics, the family structure of a population is reflected in some expected resemblance between relatives. The latter is measured as an expected covariance matrix among individuals and is used to predict genetic values (*e.g.* Crossa *et al.* 2006; Burgueño *et al.* 2007, 2011). Whereas pedigree-based models do not account for Mendelian segregation and the expected covariance matrix is constructed using assumptions that do not hold (*e.g.* absence of selection and mutation and random mating), the marker-based models allow tracing Mendelian segregation at several positions of the genome and observing realized (as opposed to expected) covariances. This enhances the potential for improving the accuracy of estimates of genetic values, thus increasing the genetic progress attainable when these predictions are used for selection purposes in lieu of pedigree-based predictions. Recently, de los Campos *et al.* (2009, 2010) and Crossa *et al.* (2010, 2011) used Bayesian estimates from genomic parametric and semi-parametric regressions, and they found that models that incorporate pedigree and markers simultaneously had better prediction accuracy for several traits in wheat and maize than models based only on pedigree or only on markers.

The standard linear genetic model represents the phenotypic response of the i^{th} individual (y_i) as the sum of a genetic value, g_i, and of a model residual, ε_i, such that the linear model for n individuals ($i = 1, \ldots, n$) is represented as $y_i = g_i + \varepsilon_i$. However, building predictive models for complex traits using a large number of molecular markers (p) with a set of lines comprising individuals (n) with $p \gg n$ is challenging because individual marker effects are not likelihood-identified. In this case, marker effects can be estimated via penalized parametric or semi-parametric methods or their Bayesian counterparts, rather than via ordinary least squares. This reduces

doi: 10.1534/g3.112.003665
Manuscript received July 9, 2012; accepted for publication October 5, 2012

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.112.003665/-/DC1
[1] Corresponding author: Colegio de Postgraduados, Montecillo, Texcoco 56230, México. E-mail: perpdgo@gmail.com

the mean-squared error of estimates; it also increases prediction accuracy of out-of-sample cases and prevents over-fitting (de los Campos *et al.* 2010). In addition to the well-known Bayes A and B linear regression models originally proposed by Meuwissen *et al.* (2001) for incorporating marker effects into g_i, there are several penalized parametric regression methods for estimating marker effects, such as ridge regression, the least absolute shrinkage and selection operator (LASSO), and the elastic net (Hastie *et al.* 2009). The Bayesian counterparts of these models have proved to be useful because appropriate priors can be assigned to the regularization parameter(s), and uncertainty in the estimations and predictions can be measured directly by applying the Bayesian paradigm.

Regression methods assume a linear relationship between phenotype and genotype, and they typically account for additive allelic effects only; however, evidence of epistatic effects on plant traits is vast and well documented (*e.g.* Holland 2001, 2008). In wheat, for instance, detailed analyses have revealed a complex circuitry of epistatic interactions in the regulation of heading time involving different vernalization genes, day-length sensitivity genes, and earliness *per se* genes, as well as the environment (Laurie *et al.* 1995; Cockram *et al.* 2007). Epistatic effects have also been found to be an important component of the genetic basis of plant height and bread-making quality traits (Zhang *et al.* 2008; Conti *et al.* 2011). It is becoming common to study gene × gene interactions by using a paradigm of networks that includes aggregating gene × gene interaction that exists even in the absence of main effects (McKinney and Pajewski 2012). Interactions between alleles at two or more loci could theoretically be represented in a linear model via use of appropriate contrasts. However, this does not scale when the number of markers (p) is large, as the number of 2-locus, 3-locus, *etc.*, interactions is mind boggling.

An alternative approach to the standard parametric modeling of complex interactions is provided by non-linear, semi-parametric methods, such as kernel-based models (*e.g.* Gianola *et al.* 2006; Gianola and van Kaam 2008) or artificial neural networks (NN) (Okut *et al.* 2011; Gianola *et al.* 2011), under the assumption that such procedures can capture signals from high-order interactions. The potential of these methods, however, depends on the kernel chosen and on the neural network architecture. In a recent study, Heslot *et al.* (2012) compared the predictive accuracy of several genome-enabled prediction models, including reproducing kernel Hilbert space (RKHS) and NN, using barley and wheat data; the authors found that the non-linear models gave a modest but consistent predictive superiority (as measured by correlations between predictions and realizations) over the linear models. In particular, the RKHS model had a better predictive ability than that obtained using the parametric regressions.

The use of RKHS for predicting complex traits was first proposed by Gianola *et al.* (2006) and Gianola and van Kaam (2008). de los Campos *et al.* (2010) further developed the theoretical basis of RHKS with "kernel averaging" (simultaneous use of various kernels in the model) and showed its good prediction accuracy. Other empirical studies in plants have corroborated the increase in prediction accuracy of kernel methods (*e.g.* Crossa *et al.* 2010, 2011; de los Campos *et al.* 2010; Heslot *et al.* 2012). Recently, Long *et al.* (2010), using chicken data, and González-Camacho *et al.* (2012), using maize data, showed that NN methods provided prediction accuracy comparable to that obtained using the RKHS method. In NN, the bases functions (adaptive "covariates") are inferred from the data, which gives the NN great potential and flexibility for capturing complex interactions between input variables (Hastie *et al.* 2009). In particular, Bayesian regularized neural networks (BRNN) and radial basis function neural networks (RBFNN) have features that make them attractive for use in genomic selection (GS).

In this study, we examined the predictive ability of various linear and non-linear models, including the Bayes A and B linear regression models of Meuwissen *et al.* (2001); the Bayesian LASSO, as in Park and Casella (2008) and de los Campos *et al.* (2009); RKHS, using the "kernel averaging" strategy proposed by de los Campos *et al.* (2010); the RBFNN, proposed and used by González-Camacho *et al.* (2012); and the BRNN, as described by Neal (1996) and used in the context of GS by Gianola *et al.* (2011). The predictive ability of these models was compared using a cross-validation scheme applied to a wheat data set from CIMMYT's Global Wheat Program.

MATERIALS AND METHODS

Experimental data

The data set included 306 elite wheat lines, 263 lines that are candidates for the 29th Semi-Arid Wheat Screening Nursery (SAWSN), and 43 lines from the 18th Semi-Arid Wheat Yield Trial (SAWYT) from CIMMYT's Global Wheat Program. These lines were genotyped with 1717 diversity array technology (DArT) markers generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au). Two traits were analyzed: grain yield (GY) and days to heading (DTH) (see Supporting Information, File S1).

The traits were measured in a total of 12 different environments (1–12) (Table 1): GY in environments 1–7 and DTH in environments 1–5 and 8–12 (10 in all). Different agronomic practices were used. Yield trials were planted in 2009 and 2010 using prepared beds and flat plots under controlled drought or irrigated conditions. Yield data from experiments in 2010 were replicated, whereas data from trials in 2009 were adjusted means from an alpha lattice incomplete block design with adjustment for spatial variability in the direction of rows and columns using the autoregressive model fitted in both directions.

Data used to train the models for GY and DTH in 2009 were the best linear unbiased estimator (BLUE) after spatial analysis, whereas the BLUE data for 2010 were obtained after performing analyses in each of the 12 environments and combined. The experimental designs in each location consisted of alpha lattice incomplete block designs of different sizes, with two replicates each.

Broad-sense heritability at individual environments was calculated as $h^2 = \sigma_g^2/(\sigma_g^2 + \frac{\sigma_e^2}{nreps})$, where σ_g^2 and σ_e^2 are the genotype and error variance components, respectively, and *nreps* is the number of replicates. For the combined analyses across environments, broad-sense heritability was calculated as $h^2 = \sigma_g^2/\left(\sigma_g^2 + \frac{\sigma_{ge}^2}{nenv} + \frac{\sigma_e^2}{nenv \times nreps)}\right)$, where the term σ_{ge}^2 is the genotype × environment interaction variance component, and *nenv* is the number of environments included in the analysis.

Statistical models

One method for incorporating markers is to define g_i as a parametric linear regression on marker covariates x_{ij} with form $g_i = \sum_{j=1}^{p} x_{ij}\beta_j$, such that $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$ (j = 1,2,. . .,p markers); here, β_j is the partial regression of y_i on the j^{th} marker covariate (Meuwissen *et al.* 2001). Extending the model to allow for an intercept

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \tag{1}$$

We adopted Gaussian assumptions for model residuals; specifically, the joint distribution of model residuals in Equation 1 was

Table 1 Twelve environments representing combinations of diverse agronomic management (drought or full irrigation, sowing in standard, bed, or flat systems), sites in Mexico, and years for two traits, grain yield (GY) and days to heading (DTH), with their broad-sense heritability (h^2) measured in 2010

Environment Code	Agronomic Management	Site in Mexico	Year	Trait Measured	h^2 (GY)	h^2 (DTH)
1	Drought-bed	Cd. Obregon	2009	GY, DTH	—	—
2	Drought-bed	Cd. Obregon	2010	GY, DTH	0.833	0.991
3	Drought-flat	Cd. Obregon	2010	GY, DTH	0.465	0.984
4	Full irrigation-bed	Cd. Obregon	2009	GY, DTH	—	—
5	Full irrigation-bed	Cd. Obregon	2010	GY, DTH	0.832	0.086
6	Heat-bed	Cd. Obregon	2010	GY	0.876	—
7	Full irrigation-flat melga	Cd. Obregon	2010	GY	0.876	—
8	Standard	Toluca	2009	DTH	—	—
9	Standard	El Batan	2009	DTH	—	—
10	Small observation plot	Cd. Obregon	2009	DTH	—	—
11	Small observation plot	Cd. Obregon	2010	DTH	—	0.950
12	Standard	Agua Fria	2010	DTH	—	0.990

assumed normal with mean zero and variance σ_e^2. The likelihood function is

$$p\left(\mathbf{y}|\mu,\mathbf{g},\sigma_e^2\right)=\prod_{i=1}^{n}N\left(y_i|\mu+\sum_{j=1}^{p}x_{ij}\beta_j,\sigma_e^2\right) \qquad (2)$$

where $N\left(y_i|\mu+\sum_{j=1}^{p}x_{ij}\beta_j,\sigma_e^2\right)$ is a normal density for random variable y_i centered at $\mu+\sum_{j=1}^{p}x_{ij}\beta_j$ and with variance σ_e^2. Depending on how priors on the marker effects are assigned, different Bayesian linear regression models result.

Linear models: Bayesian ridge regression, Bayesian LASSO, Bayes A, and Bayes B

A standard penalized regression method is ridge regression (Hoerl and Kennard 1970); its Bayesian counterpart, Bayesian ridge regression (BRR), uses a prior density of marker effects, $p(\beta_j|\boldsymbol{\omega})$, that is, Gaussian, centered at zero and with variance common to all the markers, that is, $p(\beta_j|\sigma_\beta^2)=N(\beta_j|0,\sigma_\beta^2)$, where σ_β^2 is a prior-variance of marker effects. Marker effects are assumed independent and identically distributed *a priori*. We assigned scaled inverse chi distributions $\chi^{-2}(df_.,s_.)$ to the variance parameters σ_e^2 and σ_β^2. The prior degrees of freedom parameters were set to $df_.=4$ and $s_.=1$. It can be shown that the posterior mean of marker effects is the best linear unbiased predictor (BLUP) of marker effects, so Bayesian ridge regression is often referred to as RR-BLUP (de los Campos *et al.* 2012).

The Bayesian LASSO, Bayes A, and Bayes B relax the assumption of common prior variance to all marker effects. The relationship among these three models is as follows: Bayes B can be considered as the most general of the three, in the sense that Bayes A and Bayesian ridge regression can be viewed as special cases of Bayes B. This is because Bayes A is obtained from Bayes B by setting $\pi=0$ (the proportion of markers with null effects), and Bayesian ridge regression is obtained from Bayes B by setting $\pi=0$ and assuming that all the markers have the same variance.

Bayes B uses a mixture distribution with a mass at zero, such that the (conditional) prior distribution of marker effects is given by

$$\beta_j|\sigma_j^2,\pi=\begin{cases}0 & \text{with probability } \pi\\ N(0,\sigma_j^2) & \text{with probability 1-}\pi\end{cases} \qquad (3)$$

The prior assigned to $\sigma_j^2,\ j=1,....,p$ is the same for all markers, *i.e.* a scaled inverted chi squared distribution $\chi^{-2}(df_\beta,s_\beta)$, where df_β are the degrees of freedom and s_β is a scaling parameter. Bayes B becomes Bayes A by setting π = 0.

In the case of Bayes B, we took $\pi=0.95$, $df_\beta=4$, and $s_\beta=\tilde{\sigma}_a^2(df_\beta-2)/df_\beta$ with $\tilde{\sigma}_a^2=\tilde{\sigma}_S^2/\left[(1-\pi)\sum_{j=1}^{p}2q_j(1-q_j)\right]$, where q_j is the allele frequency for marker j and $\tilde{\sigma}_S^2$ is the additive genetic variance explained by markers [see Habier *et al.* (2011) and Resende *et al.* (2012) for more details]. In the case of σ_e^2, we assigned a flat prior as in Wang *et al.* (1994).

The Bayesian LASSO assigns a double exponential (DE) distribution to all marker effects (conditionally on a regularization parameter λ), centered at zero and with marker-specific variance, that is, $p(\beta_j|\lambda,\sigma_e)=DE\left(\beta_j|0,\frac{\lambda}{\sigma_e^2}\right)$. The DE distribution does not conjugate with the Gaussian likelihood, but it can be represented as a mixture of scaled normal densities, which allows easy implementation of the model (Park and Casella 2008; de los Campos *et al.* 2009). The priors used were exactly the same as those used in González-Camacho *et al.* (2012).

The models used in this study, the Bayesian ridge regression, Bayesian LASSO (BL), Bayes A, and Bayes B, are explained in detail in several articles; for example, Bayes A and Bayes B are described in Meuwissen *et al.* (2001), Habier *et al.* (2011), and Resende *et al.* (2012), and an account of BL is given in de los Campos *et al.* (2009, 2012), Crossa *et al.* (2010, 2011), Perez *et al.* (2010), and González-Camacho *et al.* (2012).

Non-linear models: RBFNN, BRNN, and RKHS

In this section, we describe the basic structure of the non-linear single hidden layer feed-forward neural network (SLNN) with two of its variants, the radial basis function neural network and the Bayesian regularized neural network. We also give a brief explanation of RKHS with the averaging kernel method at the end of this section.

Single hidden layer feed-forward neural network: In a single-layer feed-forward (SLNN), the non-linear activation functions in the hidden layer enable a NN to have universal approximation ability, giving it great potential and flexibility in terms of capturing complex patterns. The structure of the SLNN is depicted in Figure 1, which illustrates the structure of the method for a phenotypic continuous response. This NN can be thought of as a two-step regression (*e.g.* Hastie *et al.* 2009). In the first step, in the non-linear hidden layer, S data-derived basis functions (k = 1, 2,. . ., S neurons), $\{z_i^{[k]}\}$, are inferred, and in the second step, in the linear output layer, the response is regressed on the basis functions (inferred in the hidden layer). The inner product between the input vector and the weight

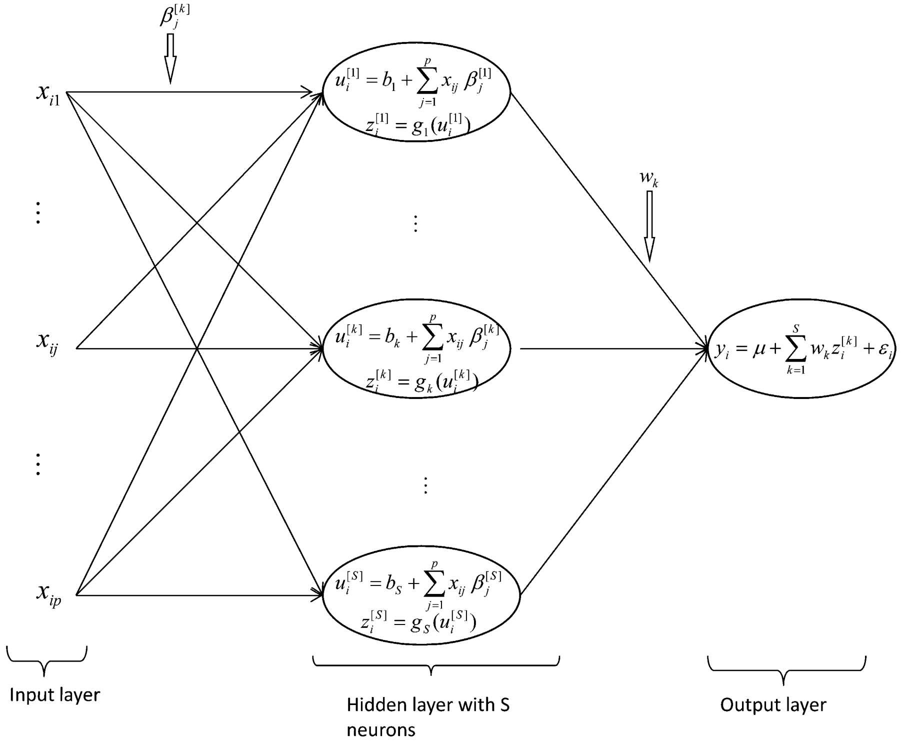
$\beta_j^{[k]}$
x_{i1}
x_{ij}
x_{ip}
$u_i^{[1]} = b_1 + \sum_{j=1}^{p} x_{ij}\,\beta_j^{[1]}$
$z_i^{[1]} = g_1(u_i^{[1]})$
$u_i^{[k]} = b_k + \sum_{j=1}^{p} x_{ij}\,\beta_j^{[k]}$
$z_i^{[k]} = g_k(u_i^{[k]})$
$u_i^{[S]} = b_S + \sum_{j=1}^{p} x_{ij}\,\beta_j^{[S]}$
$z_i^{[S]} = g_S(u_i^{[S]})$
w_k
$y_i = \mu + \sum_{k=1}^{S} w_k z_i^{[k]} + \varepsilon_i$
Input layer
Hidden layer with S neurons
Output layer

Figure 1 Structure of a single-layer feed-forward neural network (SLNN) adapted from González-Camacho *et al.* (2012). In the hidden layer, input variables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ (j = 1,. . .,p markers) are combined for each neuron (k=1,. . .,S neurons) using a linear function, $u_i^{[k]} = b_k + \sum_{j=1}^{p} x_{ij}\,\beta_j^{[k]}$, and subsequently transformed using a non-linear activation function, yielding a set of inferred scores, $z_i^{[k]} = g_k(u_i^{[k]})$. These scores are used in the output layer as basis functions to regress the response using the linear activation function on the data-derived predictors $y_i = \mu + \sum_{k=1}^{S} w_k z_i^{[k]} + \varepsilon_i$.

vector ($\boldsymbol{\beta}^{[k]}$) of each neuron of the hidden layer, plus a bias (intercept b_k), is performed, that is, $u_i^{[k]} = b_k + \sum_{j=1}^{p} x_{ij}\,\beta_j^{[k]}$, ($j$ = 1,. . .,p markers); this is then transformed using a non-linear activation function $g_k(u_i^{[k]})$. One obtains $z_i^{[k]} = g_k\left(b_k + \sum_{j=1}^{p} x_{ij}\,\beta_j^{[k]}\right)$, where b_k is an intercept and $(\beta_1{}^{[1]}, \ldots, \beta_p{}^{[1]}\,;\ldots, \beta_1{}^{[S]}, \ldots, \beta_p{}^{[S]})'$ is a vector of regression coefficients or "weights" of each neuron k in the hidden layer. The $g_k(.)$ is the activation function, which maps the inputs into the real line in the closed interval $[-1{,}1]$; for example, $g_k(x) = \dfrac{\exp(2x) - 1}{\exp(2x) + 1}$ is known as the tangent hyperbolic function. Finally, in the linear output layer, phenotypes are regressed on the data-derived features, $\{z_i^{[k]}\}$, according to

$$y_i = \mu + \sum_{k=1}^{S} w_k z_i^{[k]} + \varepsilon_i = \mu + \sum_{k=1}^{S} w_k\, g_k\left(b_k + \sum_{j=1}^{p} x_{ij}\,\beta_j^{[k]}\right) + \varepsilon_i. \tag{4}$$

Radial basis function neural network: The RBFNN was first proposed by Broomhead and Lowe (1988) and Poggio and Girosi (1990). Figure 2 shows the architecture of a single hidden layer RBFNN with S non-linear neurons. Each non-linear neuron in the hidden layer has a Gaussian radial basis function (RBF) defined as $z_i^{[k]} = \exp[-h_k\,\|\mathbf{x}_i - \mathbf{c}_k\|^2]$, where $\|\mathbf{x}_i - \mathbf{c}_k\|$ is the Euclidean norm between the input vector $\mathbf{x}_i$ and the center vector $\mathbf{c}_k$ and h_k is the bandwidth of the Gaussian RBF. Subsequently, in the linear output layer, phenotypes are regressed on the data-derived features, $\{z_i^{[k]}\}$, according to $y_i = \mu + \sum_{k=1}^{S} w_k z_i^{[k]} + \varepsilon_i$, where ε_i is a model residual.

Estimating the parameters of the RBFNN: The vector of weights $\boldsymbol{\omega} = \{w_1, \ldots, w_S\}$ of the linear output layer is obtained using the ordinary least-squares fit that minimizes the mean squared differences between the $\hat{y}_i$ (from RBFNN) and the observed responses y_i in the training set, provided that the Gaussian RBFs for centers $\mathbf{c}_k$ and h_k of the hidden layer are defined. The centers are selected using an orthogonalization least-squares learning algorithm, as described by Chen *et al.* (1991) and implemented in Matlab 2010b. The centers are added iteratively such that each new selected center is orthogonal to the others. The selected centers maximize the decrease in the mean-squared error of the RBFNN, and the algorithm stops when the number of centers (neurons) added to the RBFNN attains a desired precision (*goal* error) or when the number of centers is equal to the number of input vectors, that is, when $S = n$. The bandwidth h_k of the Gaussian RBF is defined in terms of a design parameter of the net *spread*, that is, $h_k = \left(\dfrac{0.8326}{spread}\right)^2$ for each Gaussian RBF of the hidden layer. To select the best RBFNN, a grid for training the net was generated, containing different values of *spread* and different precision values (*goal* error). The initial value of the *spread* was the median of the Euclidean distances between each pair of input vectors ($\mathbf{x}_i$), and an initial value of 0.02 for the *goal* error was considered. The parameter *spread* allows adjusting the form of the Gaussian RBF such that it is sufficiently large to respond to overlapping regions of the input space but not so big that it might induce the Gaussian RBF to have a similar response.

Bayesian regularized neural networks: The difference between SLNN and BRNN is in the function to be minimized (see the penalized function below); therefore, the basic structure of a BRNN can be represented in Figure 1 as well. The SLNN described above is flexible enough to approximate any non-linear function; this great flexibility allows NN to capture complex interactions among predictor

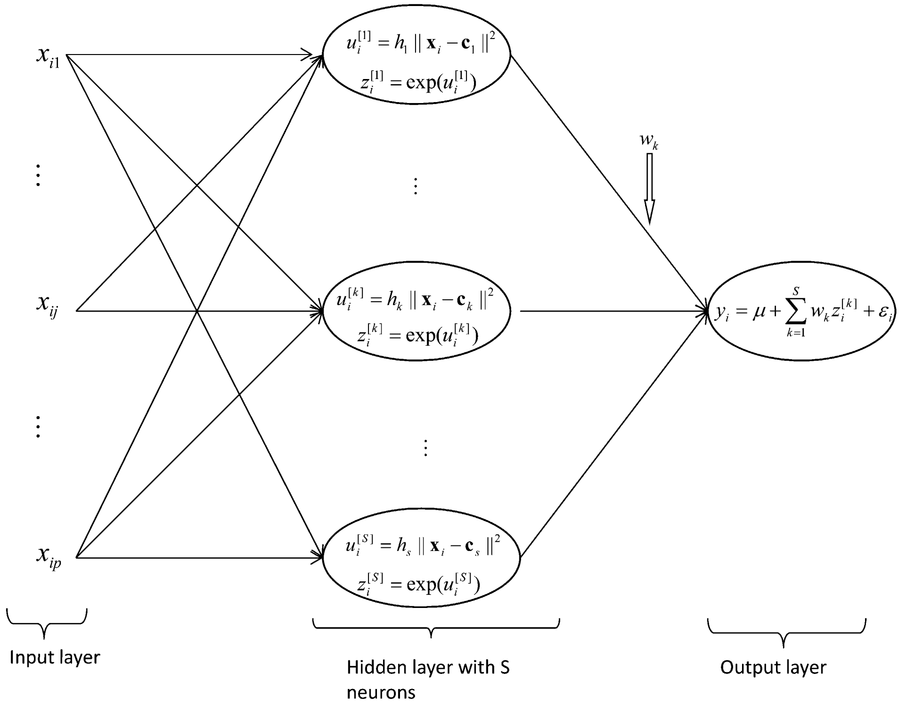
x_{i1}
$\vdots$
x_{ij}
$\vdots$
x_{ip}
$u_i^{[1]} = h_1 \| \mathbf{x}_i - \mathbf{c}_1 \|^2$
$z_i^{[1]} = \exp(u_i^{[1]})$
$u_i^{[k]} = h_k \| \mathbf{x}_i - \mathbf{c}_k \|^2$
$z_i^{[k]} = \exp(u_i^{[k]})$
$u_i^{[S]} = h_s \| \mathbf{x}_i - \mathbf{c}_s \|^2$
$z_i^{[S]} = \exp(u_i^{[S]})$
w_k
$y_i = \mu + \sum_{k=1}^{S} w_k z_i^{[k]} + \varepsilon_i$
Input layer
Hidden layer with S neurons
Output layer

Figure 2 Structure of a radial basis function neural network adapted from González-Camacho *et al.* (2012). In the hidden layer, information from input variables $(x_{i1}, ..., x_{ip})$ (j = 1,. . .,p markers) is first summarized by means of the Euclidean distance between each of the input vectors $\{\mathbf{x}_i\}$ with respect to S (data-inferred) (k=1,. . .,S neurons) centers $\{\mathbf{c}_k\}$, that is, $u_i^{[k]} = h_k \,||\mathbf{x}_i - \mathbf{c}_k||^2$. These distances are then transformed using the Gaussian function $z_i^{[k]} = \exp(-u_i^{[k]})$. These scores are used in the output layer as basis functions for the linear regression $y_i = \mu + \sum_{k=1}^{S} w_k z_i^{[k]} + \varepsilon_i$.

variables (Hastie *et al.* 2009). However, this flexibility also leads to two important issues: (1) as the number of neurons increases, the number of parameters to be estimated also increases; and (2) as the number of parameters rises, the risk of over-fitting also increases. It is common practice to use penalized methods via Bayesian methods to prevent or palliate over-fitting.

MacKay (1992, 1994) developed a framework for obtaining estimates of all the parameters in a feed-forward single neural network by using an empirical Bayes approach. Let $\boldsymbol{\theta} = (w_1,. . .,w_S;\ b_1,. . .,b_S\ ;\ \beta_1^{[1]}, . . ., \beta_p^{[1]}\ ;. . ., \beta_1^{[S]}, . . ., \beta_p^{[S]}, \mu)'$ be the vector containing all the weights, biases, and connection strengths. The author showed that the estimation problem can be solved in two steps, followed by iteration:

(1) Obtain the conditional posterior modes of the elements in $\boldsymbol{\theta}$ assuming that the variance components σ_e^2 and σ_θ^2 are known and that the prior distribution for the all the elements in $\boldsymbol{\theta}$ is given by $p(\boldsymbol{\theta}|\sigma_\theta^2) = MN(\mathbf{0}, \sigma_\theta^2 I)$. It is important to note that this approach assigns the same prior to all elements of $\boldsymbol{\theta}$, even though this may not always be the best thing to do. The density of the conditional (given the variance parameters) posterior distribution of the elements of $\boldsymbol{\theta}$, according to Bayes' theorem, is given by

$$p(\boldsymbol{\theta}|y, \sigma_e^2, \sigma_\theta^2) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, \sigma_e^2)p(\boldsymbol{\theta}|\sigma_\theta^2)}{p(\mathbf{y}|\sigma_e^2, \sigma_\theta^2)} \tag{5}$$

The conditional modes can be obtained by maximizing Equation 5 over $\boldsymbol{\theta}$. However, the problem is equivalent to minimizing the following penalized sum of squares [see Gianola *et al.* (2011) for more details]

$$F(\boldsymbol{\theta}) = \beta \sum_{i=1}^{n} e_i^2 + \alpha \sum_{j=1}^{m} \theta_j^2$$

where $\beta = 1/(2\sigma_e^2)$, $\alpha = 1/(2\sigma_\theta^2)$, e_i is the difference between observed and predicted phenotypes for the fitted model, and θ_j $(j = 1, ..., m)$ is the j^{th} element of vector $\boldsymbol{\theta}$.

(2) Update σ_e^2 and σ_θ^2. The updating formulas are obtained by maximizing an approximation to the marginal likelihood of the data $p(\mathbf{y}|\sigma_e^2, \sigma_\theta^2)$ (the "evidence") given by the denominator of Equation 5.

(3) Iterate between (1) and (2) until convergence.

The original algorithm developed by MacKay was further improved by Foresee and Hagan (1997) and adopted by Gianola *et al.* (2011) in the context of genome and pedigree-enabled prediction. The algorithm is equivalent to estimation via maximum penalized likelihood estimation when "weight decay" is used, but it has the advantage of providing a way of setting the extent of "weight decay" through the variance component σ_θ^2. Neal (1996) pointed out that the procedure of MacKay (1992, 1994) can be further generalized. For example, there is no need to approximate probabilities via Gaussian assumptions; furthermore, it is possible to estimate the entire posterior distributions of all the elements in $\boldsymbol{\theta}$, not only their (conditional) posterior modes. Next, we briefly review Neal's approach to solving the problem; a comprehensive revision can be found in Lampinen and Vehtari (2001).

Prior distributions:

a) Variance component of the residuals: Neal (1996) used a conjugate inverse Gamma distribution as a prior for the variance associated with the residual, ε_i, given in Equation 4, that is, $\sigma_e^2 \sim$ Inv-Gamma(s_e, df_e), where s_e and df_e are the scale and degrees of freedom parameters, respectively. These parameters can be set to the default values given by Neal (1996), s_e=0.05, df_e=0.5. These values were also used by Lampinen and Vehtari (2001).

b) Connection strengths, weights, and biases: Neal (1996) suggested dividing the network parameters in $\boldsymbol{\theta}$ into groups and then using hierarchical models for each group of parameters; for example, connection strengths ($\beta_1^{[1]}$, . . ., $\beta_p^{[1]}$;. . .; $\beta_1^{[S]}$, . . ., $\beta_p^{[S]}$), biases (b_1,. . .,b_S) of the hidden layer, and output weights (w_1,. . .,w_S), and general mean or bias (μ) of the linear output layer. Suppose that u_1,. . .,u_k are parameters of a given group; then assume

$$p(u_1,\ldots,u_k|\sigma_u^2) = (2\pi)^{-k/2}\sigma_u^k \exp\left\{-\frac{1}{2\sigma_u^2}\sum_{k=1}^{S} u_k^2\right\}$$

And, at the last stage of the model, assign the prior $\sigma_u^2 \sim$ Inv-Gamma(s_u, df_u). The scale parameter of the distribution associated with the group of parameters containing the connection strengths ($\beta_1{}^{[1]}, \ldots, \beta_p{}^{[1]}$;. . .; $\beta_1{}^{[S]}, \ldots, \beta_p{}^{[S]}$) changes according to the number of inputs, in this case, $s_u = (0.05/p^{1/df_u})^2$ with $df_u = 0.5$ and p is the number of markers in the data set.

By using Markov chain Monte Carlo (MCMC) techniques through an algorithm called hybrid Monte Carlo, Neal (1996) developed a software termed flexible Bayesian modeling (FBM) capable of obtaining samples from the posterior distributions of all unknowns in a neural network (as in Figure 1).

Reproducing kernel Hilbert spaces regression: RKHS models have been suggested as an alternative to multiple linear regression for capturing complex interaction patterns that may be difficult to account for in a linear model framework (Gianola *et al.* 2006). In RKHS model, the regression function takes the form

$$f(\mathbf{x}_i) = \mu + \sum_{i'=1}^{n} \alpha_{i'} K(\mathbf{x}_i, \mathbf{x}_{i'}) \tag{6}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ and $\mathbf{x}_{i'} = (x_{i'1}, \ldots, x_{i'p})'$ are input vectors of marker genotypes in individuals i and i'; $\alpha_{i'}$ are regression coefficients; and $K(\mathbf{x}_i, \mathbf{x}_{i'}) = \exp(-h\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2)$ is the reproducing kernel defined (here) with a Gaussian RBF, where h is a bandwidth parameter and $\|\mathbf{x}_i - \mathbf{x}_{i'}\|$ is the Euclidean norm between each pair of input vectors. The strategy termed "kernel averaging" for selecting optimal values of h within a set of candidate values was implemented using the Bayesian approach described in de los Campos *et al.* (2010). Similarities and connections between the RKHS and the RBFNN are given in González-Camacho *et al.* (2012).

Assessment of the models' predictive ability

The predictive ability of the models given above was compared using Pearson's correlation and predictive mean-squared error (PMSE) using predicted and realized values. A total of 50 random partitions were generated for each of the data sets, and each partition randomly assigned 90% of the lines to the training set and the remaining 10% to the validation set. The partition scheme used was similar to that in Gianola *et al.* (2011) and González-Camacho *et al.* (2012).

All scripts were run in a Linux work station; for Bayesian ridge regression and Bayesian LASSO, we used the R package BLR (de los Campos and Perez 2010), whereas for RKHS, we used the R implementation described in de los Campos *et al.* (2010), which was kindly provided by the authors. In the case of Bayes A and Bayes B, we used a program described by Hickey and Tier (2009), which is freely available at http://sites.google.com/site/hickeyjohn/alphabayes. For the BRNN, we used the FMB software available at http://www.cs.toronto.edu/~radford/fbm.software.html. Because the computational time required to evaluate the predictive ability of the BRNN network was great, we used the Condor high throughput computing system at the University of Wisconsin-Madison (http://research.cs.wisc.edu/condor). The RBFNN model was run using Matlab 2010b for Linux. The differences in computing times between the models were great. The computing times for evaluating the prediction ability of the 50 partitions for each trait were as follows, 10 min for RBFNN, 1.5 hr for RKHS, 3 hr for BRR, 3.5 hr for BL, 4.5 hr for Bayes B, 5.5 hr for Bayes A, and 30 days for BRNN. In the case of RKHS, BRR, BL, Bayes A, and Bayes B, inferences were based on 35,000 MCMC samples, and on 10,000 samples for BRNN. The estimated computing times were obtained using, as reference, a single Intel Xeon CPU 5330 2.4 GHz and 8 Gb of RAM memory. Significant reduction in computing time was achieved by parallelizing the tasks.

RESULTS

Data from replicated experiments in 2010 were used to calculate the broad-sense heritability for each trait in each environment (Table 1). Broad-sense heritability across locations for 2010 data were 0.67 for GY and 0.92 for DTH. These high estimates can be explained, at least in part, by the strict environmental control of trials conducted at CIMMYT's experiment station at Ciudad Obregon. The heritability of the two traits for 2009 was not estimated because the only available phenotypic data were adjusted means for each environment.

Predictive assessment of the models

The predictive ability of the different models for GY and DTH varied among the 12 environments. The model deemed best using correlations (Table 2) tended to be the one with the smallest average PMSE (Table 3). The three non-parametric models had higher predictive correlations and smaller PMSE than the linear models for both GY and DTH. Within the linear models, the results are mixed, and all models gave similar predictions. Within the non-parametric models, RBFNN and RKHS always gave higher correlations between predicted values and realized phenotypes, and a smaller average PMSE than the BRNN. The mean of the correlations and the associated standard errors can be used to test for statistically significant improvements in the predictability of the non-linear models *vs.* the linear models. The t-test (with $\alpha = 0.05$) showed that RKHS gave significant improvements in prediction in 13/19 cases (Table 3) compared with the BL, whereas RBFNN was significantly better than the BL in 10/19 cases. Similar results were obtained when comparing RKHS and RBFNN with Bayes A and Bayes B.

Correlations between observed and predicted values for DTH were lowest overall in environments 4 and 8, in Cd. Obregon, 2009, and in Toluca, 2009. Average PMSE was in agreement with the findings based on correlations. Although accuracies in environment 4 were much lower than in other environments, the higher accuracy of the non-parametric models (RKHS, RBFNN, and BRNN) over that of the linear models (BL, BRR, Bayes A, and Bayes B) was consistent with what was observed in the other environments. Figures 3 and 4 give scatter plots of the correlations obtained with the three non-parametric models *vs.* the BL for DTH and GY, respectively; each circle represents the estimated correlations for each of the two models included in the plot. In Figure 3, A–C, DTH had a total of 500 points (10 environments and 50 random training-testing partitions). In Figure 4, A–C, GY had a total of 350 points (7 environments and 50 random partitions in each environment). A point above the 45-degree line represents an analysis where the method whose predictive correlation is given on the vertical axis (RKHS, RBFNN, BRNN) outperformed the one whose correlation is given on the horizontal axis (BL). Both figures show that although there is a great deal of variability due to partition, for both DTH and GY, the overall superiority of RKHS and RBFNN over the linear model BL is clear. For both traits, BL had slightly better prediction accuracy than the BRNN in terms of the number of individual correlation points. It is interesting to note that some cross-validation partitions picked subsets of training data that had negative, zero, or very low correlations with the observed values in

■ **Table 2 Average correlation (SE in parentheses) between observed and predicted values for grain yield (GY) and days to heading (DTH) in 12 environments for seven models**

Trait	Environment	BL	BRR	Bayes A	Bayes B	RKHS	RBFNN	BRNN
DTH	1	0.59 (0.11)	0.59 (0.11)	0.59 (0.11)	0.56 (0.11)	0.66 (0.09)	0.66 (0.10)	0.64 (0.11)
	2	0.58 (0.14)	0.57 (0.14)	0.61 (0.12)	0.57 (0.13)	0.63 (0.13)	0.61 (0.13)	0.62 (0.13)
	3	0.60 (0.13)	0.60 (0.12)	0.62 (0.11)	0.60 (0.12)	0.68 (0.10)	0.69 (0.10)	0.67 (0.11)
	4	0.02 (0.18)	0.07 (0.17)	0.06 (0.17)	0.06 (0.17)	0.12 (0.18)	0.16 (0.18)	0.02 (0.19)
	5	0.65 (0.09)	0.64 (0.10)	0.66 (0.09)	0.66 (0.09)	0.69 (0.08)	0.68 (0.08)	0.68 (0.08)
	8	0.36 (0.15)	0.37 (0.15)	0.36 (0.15)	0.35 (0.14)	0.46 (0.13)	0.46 (0.14)	0.39 (0.15)
	9	0.59 (0.12)	0.59 (0.11)	0.53 (0.12)	0.52 (0.11)	0.62 (0.11)	0.63 (0.11)	0.61 (0.12)
	10	0.54 (0.14)	0.52 (0.14)	0.56 (0.13)	0.54 (0.14)	0.61 (0.13)	0.62 (0.12)	0.57 (0.13)
	11	0.52 (0.15)	0.52 (0.16)	0.53 (0.13)	0.51 (0.13)	0.58 (0.14)	0.59 (0.13)	0.55 (0.14)
	12	0.45 (0.19)	0.42 (0.18)	0.45 (0.18)	0.45 (0.18)	0.47 (0.18)	0.39 (0.19)	0.35 (0.19)
	Average	0.59 (0.12)	0.58 (0.12)	0.60 (0.12)	0.57 (0.12)	0.65 (0.10)	0.48 (0.14)	0.48 (0.14)
GY	1	0.48 (0.13)	0.43 (0.14)	0.48 (0.13)	0.46 (0.13)	0.51 (0.12)	0.51 (0.12)	0.50 (0.13)
	2	0.48 (0.14)	0.41 (0.17)	0.48 (0.14)	0.48 (0.14)	0.50 (0.14)	0.43 (0.16)	0.43 (0.16)
	3	0.20 (0.21)	0.29 (0.22)	0.20 (0.22)	0.18 (0.22)	0.37 (0.20)	0.42 (0.21)	0.32 (0.24)
	4	0.45 (0.15)	0.46 (0.13)	0.43 (0.15)	0.42 (0.15)	0.53 (0.12)	0.55 (0.11)	0.49 (0.14)
	5	0.59 (0.14)	0.56 (0.16)	0.75 (0.11)	0.74 (0.12)	0.64 (0.13)	0.66 (0.13)	0.63 (0.13)
	6	0.70 (0.10)	0.67 (0.11)	0.73 (0.08)	0.71 (0.08)	0.73 (0.08)	0.71 (0.08)	0.69 (0.10)
	7	0.46 (0.14)	0.50 (0.14)	0.42 (0.14)	0.40 (0.15)	0.53 (0.13)	0.54 (0.14)	0.50 (0.14)
	Average	0.62 (0.10)	0.57 (0.14)	0.69 (0.10)	0.70 (0.09)	0.67 (0.09)	0.56 (0.12)	0.65 (0.10)

Fitted models were Bayesian LASSO (BL), RR-BLUP (BRR), Bayes A, Bayes B, reproducing kernel Hilbert spaces regression (RKHS), radial basis function neural networks (RBFNN) and Bayesian regularized neural networks (BRNN) across 50 random partitions of the data with 90% in the training set and 10% in the validation set. The models with highest correlations are underlined.

the validation set. These results indicate that lines in the training set are not necessarily related to those in the validation set.

DISCUSSION AND CONCLUSIONS

Understanding the impact of epistasis on quantitative traits remains a major challenge. In wheat, several studies have reported significant epistasis for grain yield and heading or flowering time (Goldringer *et al.* 1997). Detailed analyses have shown that vernalization, day-length sensitivity, and earliness *per se* genes are mainly responsible for regulating heading time. The vernalization requirement relates to the sensitivity of the plant to cold temperatures, which causes it to accelerate spike primordial formation. Transgenic and mutant analyses, for example, have suggested a pathway involving epistatic interactions that combines environment-induced suppression and upregulation of several genes, leading to final floral transition (Shimada *et al.* 2009).

There is evidence that the aggregation of multiple gene × gene interactions (epistasis) with small effects into small epistatic networks

■ **Table 3 Predictive mean- squared error (PMSE) between observed and predicted values for grain yield (GY) and days to heading (DTH) in 12 environments for seven models**

Trait	Environment	BL	BRR	Bayes A	Bayes B	RKHS	RBFNN	BRNN
DTH	1	13.02	13.18	12.72	13.23	11.02	10.85	11.52
	2	11.89	12.37	10.65	11.28	10.19	10.72	10.44
	3	8.18	8.44	7.31	7.59	6.29	6.25	6.63
	4	21.59	22.27	21.79	21.67	21.14	22.64	21.49
	5	8.86	9.23	8.48	8.37	7.95	8.02	8.21
	8	14.72	15.22	14.54	14.58	13.12	13.19	14.81
	9	21.38	21.44	23.71	23.93	20.50	19.84	20.62
	10	7.72	8.51	7.27	7.57	6.66	6.51	7.36
	11	6.83	7.12	6.59	6.74	6.03	5.96	6.51
	12	13.60	14.42	13.56	13.46	13.25	14.86	15.75
	Average	6.09	6.47	5.99	6.28	5.31	9.12	9.25
GY	1	0.07	0.09	0.07	0.07	0.07	0.07	0.07
	2	0.06	0.08	0.06	0.06	0.06	0.07	0.07
	3	0.06	0.07	0.06	0.06	0.05	0.05	0.05
	4	0.22	0.24	0.23	0.23	0.20	0.19	0.21
	5	0.39	0.44	0.26	0.27	0.35	0.33	0.36
	6	0.13	0.15	0.12	0.13	0.12	0.13	0.13
	7	0.40	0.41	0.43	0.44	0.38	0.37	0.39
	Average	0.06	0.07	0.05	0.05	0.05	0.07	0.06

Fitted models were Bayesian LASSO (BL), RR-BLUP (BRR), Bayes A, Bayes B, reproducing kernel Hilbert space regression (RKHS), radial basis function neural networks (RBFNN) and Bayesian regularized neural networks (BRNN) across 50 random partitions of the data with 90% in the training set and 10% in the validation set. The models with lowest PMSE are underlined.

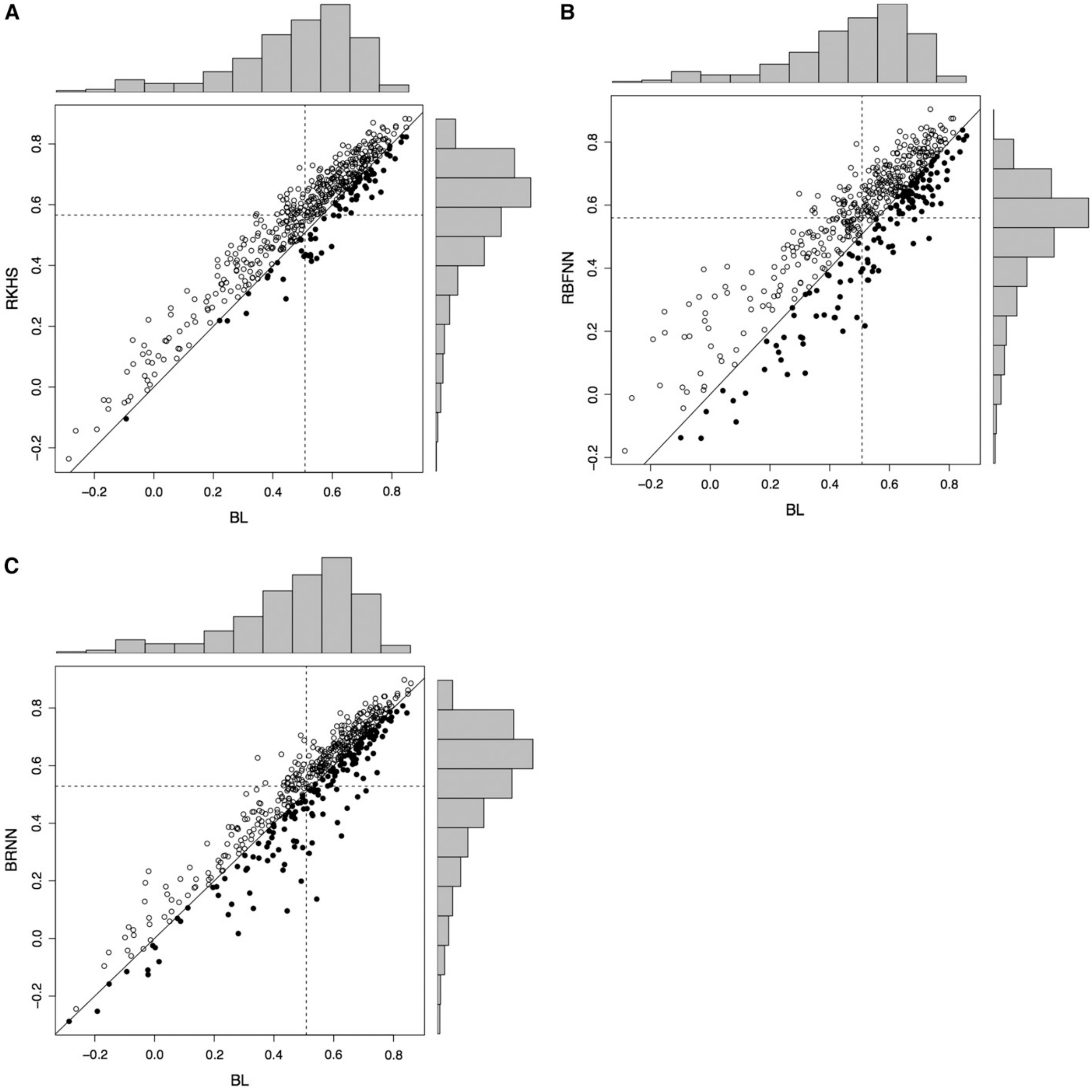

Figure 3 Plots of the predictive correlation for each of 50 cross-validation partitions and 10 environments for days to heading (DTH) in different combinations of models. (A) When the best non-parametric model is RKHS, this is represented by an open circle; when the best linear model is BL, this is represented by a filled circle. (B) When the best non-parametric model is RBFNN, this is represented by an open circle; when the best linear model is BL, this is represented by a filled circle. (C) When the best non-parametric model is BRNN, this is represented by an open circle; when the best linear model is BL, this is represented by a filled circle. The histograms depict the distribution of the correlations in the testing set obtained from the 50 partitions for different models. The horizontal (vertical) dashed line represents the average of the correlations for the testing set in the 50 partitions for the model shown on the Y (X) axis. The solid line represents Y = X; *i.e.* both models have the same prediction ability.

is important for explaining the heritability of complex traits in genome-wide association studies (McKinney and Pajewski 2012). Epistatic networks and gene × gene interactions can also be exploited for GS via suitable statistical-genetic models that incorporate network complexities. Evidence from this study, as well as from other research involving other plant and animal species, suggests that models that are non-linear in input variables (*e.g.* SNPs) predict outcomes in testing sets better than standard linear regression models for genome-enabled prediction. However, it should be pointed out that better predictive ability can have several causes, one of them the ability of some non-linear models to capture epistatic effects. Furthermore, the random cross-validation scheme used in this study was not designed to

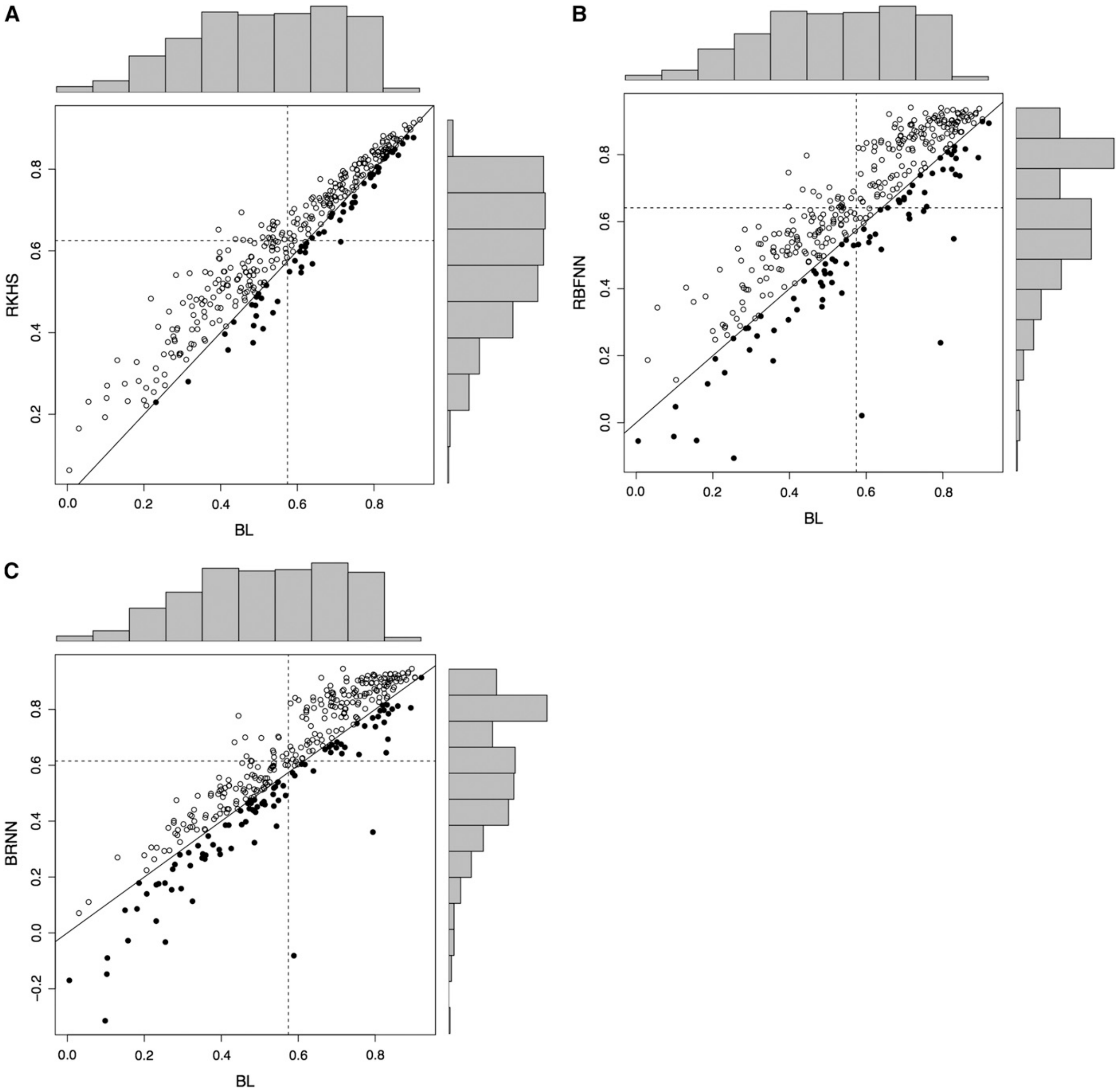

Figure 4 Plot of the correlation for each of 50 cross-validation partitions and seven environments for grain yield (GY) in different combinations of models. (A) When the best model is RKHS, this is represented by an open circle; when the best model is BL, this is represented by a filled circle. (B) When best model is RBFNN, this is represented by an open circle; when the best model is BL, this is represented by a filled circle. (C) When the best model is BRNN, this is represented by an open circle; when the best model is BL, this is represented by a filled circle. The histograms depict the distribution of the correlations in the testing set obtained from the 50 partitions for different models. The horizontal (vertical) dashed line represents the average of the correlations for the testing set in the 50 partitions for the model shown on the Y (X) axis. The solid line represents Y = X; *i.e.* both models have the same prediction ability.

specifically assess epistasis but rather to compare the models' predictive ability.

It is interesting to compare results from different predictive machineries when applied to either maize or wheat. Differences in the prediction accuracy of non-parametric and linear models (at least for the data sets included in this and other studies) seem to be more pronounced in wheat than in maize. Although differences depend, among other factors, on the trait-environment combination and the number of markers, it is clear from González-Camacho *et al.* (2012) that for flowering traits (highly additive) and traits such as grain yield (additive and epistatic) in maize, the BL model performed very similarly to the RKHS and RBFNN. On the other hand, in the present study, which involves wheat, the RKHS, RBFNN, and BRNN models clearly had a markedly better predictive accuracy than BL, BRR, Bayes A, or Bayes B. This may be due to the fact that, in wheat, additive × additive epistasis plays an important role in grain yield, as found by

Crossa *et al.* (2006) and Burgueño *et al.* (2007, 2011) when assessing additive, additive × additive, additive × environment, and additive × additive × environment interactions using a pedigree-based model with the relationship matrix **A**.

As pointed out first by Gianola *et al.* (2006) and subsequently by Long *et al.* (2010), non-parametric models do not impose strong assumptions on the phenotype-genotype relationship, and they have the potential of capturing interactions among loci. Our results with real wheat data sets agreed with previous findings in animal and plant breeding and with simulated experiments, in that a non-parametric treatment of markers may account for epistatic effects that are not captured by linear additive regression models. Using extensive maize data sets, González-Camacho *et al.* (2012) found that RBFNN and RKHS had some similarities and seemed to be useful for predicting quantitative traits with different complex underlying gene action under varying types of interaction in different environmental conditions. These authors suggested that it is possible to make further improvements in the accuracy of the RKHS and RBFNN models by introducing differential weights in SNPs, as shown by Long *et al.* (2010) for RBFs.

The training population used here was not developed specifically for this study; it was made up of a set of elite lines from the CIMMYT rain-fed spring wheat breeding program. Our results show that it is possible to achieve good predictions of line performance by combining phenotypic and genotypic data generated on elite lines. As genotyping costs decrease, breeding programs could make use of genome-enabled prediction models to predict the values of new breeding lines generated from crosses between elite lines in the training set before they reach the yield testing stage. Lines with the highest estimated breeding values could be intercrossed before being phenotyped. Such a "rapid cycling" scheme would accelerate the fixation rate of favorable alleles in elite materials and should increase the genetic gain per unit of time, as described by Heffner *et al.* (2009).

It is important to point out that proof-of-concept experiments are required before genome-enabled selection can be implemented successfully in plant breeding programs. It is necessary to test genomic predictions on breeding materials derived from crosses between lines of the training population. If predictions are reliable enough, an experiment using the same set of parental materials could be carried out to compare the field performance of lines coming from a genomic-assisted recurrent selection program scheme *vs.* lines coming from a conventional breeding scheme. The accuracies reported in this study represent prediction of wheat lines using a training set comprising lines with some degree of relatedness to lines in the validation set. When the validation and the training sets are not genetically related (unrelated families) or represent populations with different genetic structures and different linkage disequilibrium patterns, then negligible accuracies are to be expected. It seems that successful application of genomic selection in plant breeding requires some genetic relatedness between individuals in the training and validation sets, and that linkage disequilibrium information *per se* does not suffice (*e.g.* Makowsky *et al.* 2011).

ACKNOWLEDGMENTS

Financial support by the Wisconsin Agriculture Experiment Station and the AVIAGEN, Ltd. (Newbridge, Scotland) to Paulino Pérez and Daniel Gianola is acknowledged. We thank the Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT) researchers who carried out the wheat trials and provided the phenotypic data analyzed in this article.

LITERATURE CITED

Bernardo, R., and J. M. Yu, 2007 Prospects for genome-wide selection for quantitative traits in maize. Crop Sci. 47(3): 1082–1090.

Broomhead, D. S., and D. Lowe, 1988 Multivariable functional interpolation and adaptive networks. Complex Systems 2: 321–355.

Burgueño, J., J. Crossa, P. L. Cornelius, R. Trethowan, G. McLaren *et al.*, 2007 Modeling additive × environment and additive × additive × environment using genetic covariances of relatives of wheat genotypes. Crop Sci. 47(1): 311–320.

Burgueño, J., J. Crossa, J. M. Cotes, F. San Vicente, and B. Das, 2011 Prediction assessment of linear mixed models for multienvironment trials. Crop Sci. 51(3): 944–954.

Chen, S., C. F. N. Cowan, and P. M. Grant, 1991 Orthogonal least squares learning algorithm for radial basis function networks. Neural Networks, IEEE Transactions on 2(2): 302–309.

Cockram, J., H. Jones, F. J. Leigh, D. O'Sullivan, W. Powell *et al.*, 2007 Control of flowering time in temperate cereals: genes, domestication, and sustainable productivity. J. Exp. Bot. 58(6): 1231–1244.

Conti, V., P. F. Roncallo, V. Beaufort, G. L. Cervigni, R. Miranda *et al.*, 2011 Mapping of main and epistatic effect QTLs associated to grain protein and gluten strength using a RIL population of durum wheat. J. Appl. Genet. 52(3): 287–298.

Crossa, J., J. Burgueño, P. L. Cornelius, G. McLaren, R. Trethowan *et al.*, 2006 Modeling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Sci. 46(4): 1722–1733.

Crossa, J., G. de los Campos, P. Perez, D. Gianola, J. Burgueño *et al.*, 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2): 713–724.

Crossa, J., P. Perez, G. de los Campos, G. Mahuku, S. Dreisigacker *et al.*, 2011 Genomic selection and prediction in plant breeding. J. Crop Improv. 25(3): 239–261.

de los Campos, G., and P. Perez, 2010. BLR: Bayesian Linear Regression R package, version 1.2.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1): 375–385.

de los Campos, G., D. Gianola, G. J. M. Rosa, K. A. Weigel, and J. Crossa, 2010 Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92(4): 295–308.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2012 Whole genome regression and prediction methods applied to plant and animal breeding. Genetics DOI: 10.1534/genetics.112.14331.

Foresee, D., and M. T. Hagan, 1997. Gauss-Newton approximation to Bayesian learning. International Conference on Neural Networks, June 9–12, Houston, TX.

Gianola, D., and J. B. C. H. M. van Kaam, 2008 Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178(4): 2289–2303.

Gianola, D., R. L. Fernando, and A. Stella, 2006 Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173(3): 1761–1776.

Gianola, D., H. Okut, K. A. Weigel, and G. J. M. Rosa, 2011 Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. 12: 87.

Goldringer, I., P. Brabant, and A. Gallais, 1997 Estimation of additive and epistatic genetic variances for agronomic traits in a population of doubled-haploid lines of wheat. Heredity 79: 60–71.

González-Camacho, J. M., G. de los Campos, P. Perez, D. Gianola, J. Cairns *et al.*, 2012 Genome-enabled prediction of genetic values using radial basis function. Theor. Appl. Genet. 125: 759–771.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrik, 2011 Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

Hastie, T., R. Tibshirani, and J. Friedman, 2009 *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Ed. 2. Springer, New York.

Heffner, E. L., M. E. Sorrells, and J. L. Jannink, 2009 Genomic selection for crop improvement. Crop Sci. 49(1): 1–12.

Heslot, N., H. P. Yang, M. E. Sorrells, and J. L. Jannink, 2012 Genomic selection in plant breeding: a comparison of models. Crop Sci. 52(1): 146–160.

Hickey, J. M., and B. Tier, 2009 *AlphaBayes (Beta): Software for Polygenic and Whole Genome Analysis. User Manual.* University of New England, Armidale, Australia.

Hoerl, A. E., and R. W. Kennard, 1970 Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1): 55–67.

Holland, J. B., 2001 Epistasis and plant breeding. Plant Breeding Reviews 21: 27–92.

Holland, J. B., 2008 Theoretical and biological foundations of plant breeding, pp. 127–140 in *Plant Breeding: The Arnel R. Hallauer International Symposium*, edited by K. R. Lamkey and M. Lee. Blackwell Publishing, Ames, IA.

Lampinen, J., and A. Vehtari, 2001 Bayesian approach for neural networks - review and case studies. Neural Netw. 14(3): 257–274.

Laurie, D. A., N. Pratchett, J. W. Snape, and J. H. Bezant, 1995 RFLP mapping of five major genes and eight quantitative trait loci controlling flowering time in a winter × spring barley (*Hordeum vulgare* L.) cross. Genome 38(3): 575–585.

Long, N. Y., D. Gianola, G. J. M. Rosa, K. A. Weigel, A. Kranis *et al.*, 2010 Radial basis function regression methods for predicting quantitative traits using SNP markers. Genet. Res. 92(3): 209–225.

MacKay, D. J. C., 1992 A practical Bayesian framework for backpropagation networks. Neural Comput. 4(3): 448–472.

MacKay, D. J. C., 1994 Bayesian non-linear modelling for the prediction competition. ASHRAE Transactions 100(Pt. 2): 1053–1062.

Makowsky, R., N. M. Pajewski, Y. C. Klimentidis, A. I. Vazquez, C. W. Duarte *et al.*, 2011 Beyond missing heritability: prediction of complex traits. PLoS Genet. 7(4): e1002051.

McKinney, B. A., and N. M. Pajewski, 2012. Six degrees of epistasis: statistical network models for GWAS. Front. Genet. 2: 109.

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 (4): 1819–1829.

Neal, R. M., 1996. *Bayesian Learning for Neural Networks (Lecture Notes in Statistics)*, Vol. 118. Springer-Verlag, NY.

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu *et al.*, 2012 Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. PLoS Genet. 8(5): e1002685.

Okut, H., D. Gianola, G. J. Rosa, and K. A. Weigel, 2011 Prediction of body mass index in mice using dense molecular markers and a regularized neural network. Genet. Res. Camb. 93: 189–201.

Park, T., and G. Casella, 2008 The Bayesian LASSO. J. Am. Stat. Assoc. 103: 681–686.

Perez, 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 Genome 3(2): 106–116.

Poggio, T., and F. Girosi, 1990 Networks for approximation and learning. Proc. IEEE 78(9): 1481–1497.

Resende, M. F. R., P. Muñoz, M. D. V. Resende, D. J. Garrick, R. L. Fernando *et al.*, 2012 Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 4: 1503–1510.

Shimada, S., T. Ogawa, and S. Kitagawa, 2009 A genetic network of flowering-time genes in wheat leaves, in which an *APETALA1/FRUITFULL*-like gene, VRN-1, is upstream of *FLOWERING LOCUS T*. Plant J. 58: 668–681.

Wang, C. S., J. J. Rutledge, and D. Gianola, 1994 Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genet. Sel. Evol. 26: 91–115.

Zhang, K., J. Tian, L. Zhao, and S. Wang, 2008 Mapping QTLs with epistatic effects and QTL × environment interactions for plant height using a doubled haploid population in cultivated wheat. J. Genet. Genomics 35 (2): 119–127.

Communicating editor: J. B. Holland