Genome-enabled methods for predicting litter size in pigs: a comparison

L. Tusell1†, P. Pérez-Rodríguez1,2, S. Forni3, X.-L. Wu1,4 and D. Gianola1,4,5

1Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA; 2Colegio de Postgraduados, Km. 36.5 Carretera México, Texcoco, Montecillo, Estado de México, 56230, México; 3Genus Plc, Hendersonville, TN, USA; 4Department of Dairy Science, University of Wisconsin, Madison, WI 53706, USA; 5Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53706, USA

(Received 19 December 2012; Accepted 31 May 2013; First published online 24 July 2013)

Predictive ability of models for litter size in swine on the basis of different sources of genetic information was investigated. Data represented average litter size on 2598, 1604 and 1897 60K genotyped sows from two purebred and one crossbred line, respectively. The average correlation (r) between observed and predicted phenotypes in a 10-fold cross-validation was used to assess predictive ability. Models were: pedigree-based mixed-effects model (PED), Bayesian ridge regression (BRR), Bayesian LASSO (BL), genomic BLUP (GBLUP), reproducing kernel Hilbert spaces regression (RKHS), Bayesian regularized neural networks (BRNN) and radial basis function neural networks (RBFNN). BRR and BL used the marker matrix or its principal component scores matrix (UD) as covariates; RKHS employed a Gaussian kernel with additive codes for markers whereas neural networks employed the additive genomic relationship matrix (G) or UD as inputs. The non-parametric models (RKHS, BRNN, RBFNN) gave similar predictions to the parametric counterparts (average r ranged from 0.15 to 0.23); most of the genome-based models outperformed PED (r = 0.16). Predictive abilities of linear models and RKHS were similar over lines, but BRNN varied markedly, giving the best prediction (r = 0.31) when G was used in crossbreds, but the worst (r = 0.02) when the G matrix was used in one of the purebred lines. The r values for RBFNN ranged from 0.16 to 0.23. Predictive ability was better in crossbreds (0.26) than in purebreds (0.15 to 0.22). This may be related to family structure in the purebred lines.

Keywords: genomic prediction, Bayesian regression, RKHS, neural networks, litter size

Implications

Genomics is now used extensively in livestock and plant breeding, because of the availability of a large number of molecular markers. An important issue to be addressed before implementation of genome-enabled breeding programs is how the marker information can be incorporated into statistical models, so as to enhance prediction of yet to be observed phenotypes. Performance of genome-enabled linear and non-linear models using different sources of genomic information for predicting yet to be observed litter size records is evaluated in three pig lines. Possible factors influencing predictive ability of the models are discussed.

Introduction

The potential value of genomic information is based on the expectation that linkage disequilibrium between dense markers and quantitative trait loci (QTL) can be exploited to improve selection reliability, decrease breeding costs and generation intervals, and enhance the rate of genetic gain (Dekkers, 2004). An important issue to be addressed before implementation of genome-enabled breeding programs is how the marker information can be incorporated into statistical models, so as to enhance prediction of yet to be observed phenotypes.

Many genome-enabled prediction models differing in their specifications have been proposed (e.g. de los Campos et al., 2012a). Differences result from: (1) type of relationship among target traits and explanatory covariates (i.e. linear or non-linear relationships); (2) type of genomic information employed (use of all or a reduced subset of marker genotypes, or of measures of genomic similarities among individuals as conveyed by some genomic relationship matrix) and (3) strategies used to confront the ‘small n, large p’ estimation problem, both within classical or Bayesian frameworks.

Because selective breeding relies heavily on additive inheritance, linear models including only additive genetic effects have been used in the main. However, the additivity assumptions methods may not be adequate when dominance
or epistasis exists as one would expect, for example, in inbred or crossbred lines. Accounting for dominance in a parametric model may result in a better predictive accuracy, but non-parametric methods are more flexible. For example, models that are non-linear in either features (single-nucleotide polymorphisms, SNPs) or on parameters have been proposed for enhancing genome-enabled prediction of complex traits (Gianola et al., 2006; González-Camacho et al., 2012). Properly specified non-linear methods may be able to capture complex signals from the data and result in a better predictive accuracy, even if the trait is additive (Pérez et al., 2013; Pérez-Rodríguez et al., 2012). The present study aims at comparing the performance of several genome-enabled models for predicting litter size in pigs, and at evaluating the impact of different sources of genomic information using data from two purebred and one crossbred swine lines.

Material and methods
Animal care and use committee approval was not needed for this study since data came from an existing database.

Phenotype and genotype data
Phenotypes and genotypes for sows of two purebred Landrace-origin lines (A and B) and their commercial cross (AB) were provided by Genus-Plc (Hendersonville, TN, USA).

Individual litter size (total number of piglets born per litter) records of sows genotyped were pre-corrected by fitting a linear mixed model with the lme4 (Bates et al., 2012) R-package (R Core Team, 2012), to take into account the environmental random effects of number of services, mating type (natural or artificial insemination), parity order within farm, farm and month of farrow and age at first farrowing. Adjusted litter size records were obtained after subtracting the average of adjusted litter size records per sow (PB) was the target phenotype to be predicted. Histograms of the raw litter size records, adjusted litter size records and average of adjusted litter size records are given in Supplementary Figure S1, suggesting that approximated normality holds. The number of sows with records was 2598, 1604 and 1879 in the A, B and AB lines, respectively. Sows were genotyped using the Illumina PorcineSNP60 BeadChip. SNPs with minor allele frequency <0.05 and missing genotype frequency >0.95 were removed. For the remaining loci, missing genotypes were imputed using a naïve method based on allele frequencies at each locus. Table 1 shows summary statistics of phenotypic and genotypic data for the animals in the A, B and AB lines.

Models
The predictive ability of six linear and five non-linear (either in the features or on the parameters) models was compared using cross-validation, with phenotypes in testing sets viewed as yet to be observed in the training sets. Comparisons were on a within line basis. A summary of these models is listed in Table 2.

Linear genome-enabled models. Let $y$ be a vector of average pre-corrected number of piglets per sow, following the linear regression model:

$$
y = 1\mu + X\beta + \varepsilon \tag{1}
$$

where $y = (y_1, \ldots, y_n)$ is a vector of residuals. In this model, the number $r_i$ is the number of records used for computing the average and $b$ indexes an individual record; $\mu$ is an effect common to all records; $\beta = (\beta_j)$ is the vector of allele substitution effects of $p$ markers; $X = (x_{ij})$ is a matrix of marker covariates of dimensions $n \times p$ ($j = 1, \ldots, p$ and $i = 1, \ldots, n$, with $n$ being the number of phenotypes), and $\varepsilon = (\varepsilon_i)$ is the vector of residuals. In this model, the number of location parameters to be inferred largely exceeds the number of phenotypes. Hence, marker effects cannot be estimated uniquely from the likelihood function.

One approach for dealing with this issue is to use Bayesian shrinkage regression methods in which marker effects are treated as random variables. Assume that residuals follow the normal distribution, where $\sigma^2$ is the residual variance of individual records. Residual correlation in the data was not accounted for in the analysis, whereas heterogeneity of residual variance was partially alleviated by weighting the

<table>
<thead>
<tr>
<th>Parameters</th>
<th>A</th>
<th>B</th>
<th>AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average (standard deviation) litter size of females genotyped</td>
<td>12.16 (3.25)</td>
<td>13.03 (3.30)</td>
<td>13.75 (3.30)</td>
</tr>
<tr>
<td>Number of records of females genotyped</td>
<td>11 192</td>
<td>9438</td>
<td>9089</td>
</tr>
<tr>
<td>Number of sows genotyped with phenotypic records</td>
<td>2598</td>
<td>1604</td>
<td>1879</td>
</tr>
<tr>
<td>Average (standard deviation) litter size per sow genotyped</td>
<td>12.23 (2.41)</td>
<td>13.02 (2.26)</td>
<td>13.47 (2.72)</td>
</tr>
<tr>
<td>Minimum, mean and maximum number of records per sow</td>
<td>1, 4, 3, 8</td>
<td>1, 5.9, 8</td>
<td>1, 4, 3, 8</td>
</tr>
<tr>
<td>Number of SNPs, initial</td>
<td>64 232</td>
<td>64 232</td>
<td>64 232</td>
</tr>
<tr>
<td>Number of SNPs, after editing</td>
<td>46 855</td>
<td>45 597</td>
<td>50 151</td>
</tr>
<tr>
<td>Average frequency of imputation, per SNP</td>
<td>0.007</td>
<td>0.003</td>
<td>0.006</td>
</tr>
<tr>
<td>Average minor allele frequency</td>
<td>0.278</td>
<td>0.277</td>
<td>0.288</td>
</tr>
<tr>
<td>Average frequency of heterozygote SNPs</td>
<td>0.355</td>
<td>0.356</td>
<td>0.421</td>
</tr>
</tbody>
</table>

SNP = single-nucleotide polymorphism.
residual variance by \( r_i \) (de los Campos et al., 2012a). Then from equation (1), the likelihood function stems from the sampling model:

\[
p(y|\beta, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}(y_i|\mu + \sum_{j=1}^{p} x_{ij}\beta_j, \sigma^2/r_i) \tag{2}
\]

In the Bayesian framework prior assumptions are made about the unknown parameters, and differences in assumptions lead to different Bayesian regression models (de los Campos et al., 2009b). In what follows, the two Bayesian linear regression models employed here are described.

Let all \( p \) marker effects be assigned identical and independent distributions with joint prior density \( p(\beta|\omega) = \prod_{j=1}^{p} p(\beta_j|\omega) \), where \( \omega \) represents a collection of known parameters. Then, the posterior density of \( \beta \) is:

\[
p(\beta|y, \sigma^2, \omega) \propto p(y|\beta, \sigma^2) p(\beta|\omega)
\]

\[
= \prod_{i=1}^{n} \mathcal{N}(y_i|\mu + \sum_{j=1}^{p} x_{ij}\beta_j, \sigma^2/r_i) \prod_{j=1}^{p} p(\beta_j|\omega) \tag{3}
\]

Bayesian ridge regression (BRR) assumes that the prior density of a marker effect, \( p(\beta_j|\omega) \), is that of a Normal distribution with mean 0 and variance \( \sigma^2 \), common to all markers, that is, \( p(\beta_j|\sigma^2) = \mathcal{N}(\beta_j|0, \sigma^2) \), with \( \omega = \sigma^2 \). Bayesian LASSO (BL; Park and Casella, 2008), assigns a double-exponential (DE) distribution to marker effects, with mean zero and inverse scale parameter \( \sqrt{\sigma^2} \), that is \( p(\beta_j|\sigma^2) = \text{DE}(\beta_j|0, \frac{\sigma^2}{\sqrt{\sigma^2}}) \). Typically, \( \lambda^2 \) (the square of the regularization parameter) is assigned a Gamma or a Beta distribution. The DE distribution assigns a larger mass near 0 than the Normal, but it also allows for larger marker effects because it has thicker tails. Statistically, the DE density can be represented as a mixture of scaled-normal densities, which allows structuring the original problem in terms of a hierarchical model that can be implemented via Gibbs sampling (e.g. Park and Casella, 2008; de los Campos et al., 2009b).

Assuming additive allele substitution effects, SNP genotypes can be coded as \(-1, 0 \) and \( 1 \) for the homozygote for the minor allele, heterozygote and homozygote for the other allele. This leads to \( X = (x_{ij}) \), a matrix of dimension \( n \times p \) for all SNP genotypes of all individuals. For convenience, we denote BRR_X and BL_X as the Bayesian ridge regression and the Bayesian LASSO models, respectively, each using marker codes in \( X \) as the predictor variables. Aiming at reducing the number of parameters to be estimated, but without losing much of the information contained in \( X \), we also used as covariables the principal component scores (UD) of \( X \) in our BRR and BL models. Here, \( UD \) is an \( n \times n \) matrix, which is obtained from the singular value decomposition of \( X \) (Mandel, 1982):

\[
X = UDV' \tag{4}
\]

where the columns of \( U \) and \( V \) are the left and right singular vectors of \( X \), respectively, and \( D \) is a diagonal matrix containing the non-zero-singular values of \( X \) (square roots of the non-zero eigenvalues of either \( XX' \) or \( X'X \)). The cumulative percentage of variance explained by the components of the singular value decomposition for the three lines is shown in Figure 1. This figure illustrates the reason why we choose to keep all components of \( UD \) and not less. Since the square of the singular value is proportional to the variance explained by each singular vector, it can be seen from the histograms that each left singular vector kept in \( UD \) contributed a small amount to the variability of \( X \).

If \( UDV' \) is used instead of \( X \), model (1) can be re-parameterized as:

\[
y = \mu 1 + UD\beta + \varepsilon = \mu 1 + UD\beta^* + \varepsilon \tag{5}
\]

Here, \( \beta^* = V\beta \), so that \( \beta^* \) is assigned the same prior distribution as that of \( \beta \) in the BRR and BL models, for example, \( \beta^* \sim \mathcal{N}(0, I_{p^2}) \) in ridge regression because \( VV' = I \). We use BRR_UD and BL_UD to represent BRR and BL models that use the \( UD \) matrix as inputs, respectively.
Intuitively, a genome-enabled best linear unbiased prediction model genomic BLUP (GBLUP) is obtained by replacing $X\beta$ by $u$ in model (1):

$$y = 1\mu + u + \varepsilon$$  \hspace{1cm} (6)

where the residuals follow the normal distribution described previously. The vector of additive genomic effects, or genomic breeding values (GBV), $u = \{u_i\}$ is assumed to be normally distributed, $p(u|\sigma_u^2) = N(0, \sigma_u^2)$, where $\sigma_u^2$ is the additive genomic variance, and $G$ is the genomic relationship matrix (VanRaden, 2008).

$$G = \frac{(X - E)(X - E)'}{2 \sum_{j=1}^{p} q_j (1 - q_j)}$$  \hspace{1cm} (7)

Here, the $G$ matrix can be viewed as a ‘standardized’ (co)variance matrix of genomic similarities among individuals, because genotypes in the $X$ matrix are centered by subtracting their expected frequencies at each locus ($E$ columns). VanRaden (2008) stated that allele frequencies at each locus $q_j$ for $j = 1, \ldots, p$ should be those of a base population under Hardy–Weinberg equilibrium, in order to avoid biased estimates of degrees of relationships and of inbreeding. However, such information was not available to us. Hence, allele frequencies were estimated from the available data. Prior distributions assigned to $\sigma_g^2$, $\sigma_d^2$ and $\sigma_\varepsilon^2$ were proper scaled inverse chi-squared distributions with degrees of freedom parameter, $df_l$ and scale parameter $S_l$ assigned arbitrarily so that $p(\sigma_l^2|df_l, S_l) = \chi^{-2}(\sigma_l^2|df_l, S_l)$, for $l \in \{\beta, u, \varepsilon\}$.

**Non-linear genome-enabled models.** The standard assumption made in linear methods, that is, fitting additive effects of markers only, prevents from capturing complex interactions that may exist, leading to incorrectly specified models. The use of non-parametric methods may deliver a higher predictive accuracy. In this study, three types of non-linear models were used: reproducing kernel Hilbert spaces (RKHS), Bayesian regularized neural networks (BRNN) and radial basis functions neural networks (RBFNN). RKHS is non-linear in the inputs (SNPs) but linear in the parameters; on the other hand, BRNN and RBFNN are doubly non-linear.

The RKHS model was introduced in quantitative genomics by Gianola et al. (2006) and it can be represented as follows:

$$y = 1\mu + g + \varepsilon$$  \hspace{1cm} (8)

where $g = [g(x_1), \ldots, g(x_n)]$ the vector of conditional expectation functions, maps each input vector of marker information $x_i$ of order $p \times 1$, onto the real line. Here, the distribution of the residuals is as in model (1). The prior distribution for $g$ is normal: $p(g|\sigma_g^2) = N(0, \sigma_g^2)$, where $\sigma_g^2$ is an unknown variance and $K$ is an $n \times n$ positive definite matrix, called the reproducing kernel matrix. $K$ can be viewed as an expected correlation structure between pairs of points evaluated in the conditional expectation function: $\text{cov}(g_j, g_k) \propto K(x_j, x_k)$ (de los Campos et al., 2009a). Conveniently, $K$ can be built using a Gaussian basis function: $K = K(x_j, x_k) = \exp(-h ||x_j - x_k||^2)$, where the bandwidth parameter $h$ determines how fast the correlation between two points drops and $||x_j - x_k||$ is the Euclidean distance between pairs of marker genotypes. Note that the inputs (markers) enter non-linearly in $K$ but that the RKHS model is linear in $g$. Prediction using RKHS methods is very sensitive with respect to values of the bandwidth parameter, making the choice of an appropriate $h$ value important. This issue was addressed by de los Campos et al. (2009a), who proposed a Bayesian approach where several kernels are employed together in a model as if these represented different random factors in a linear model. Even though markers were coded for additive effects, this multi kernel approach should be capable of capturing some non-additivity (Gianola et al., 2006). Following this idea, we fitted three different kernels ($d = 1, 2, 3$) each with a different $h$. The $\sigma_g^2$ and the $\sigma_{d,\varepsilon}^2$ variances were assigned scaled inverse $\chi^2$ prior distributions.

Another type of non-linear model, both in marker covariates and parameters, that was considered for prediction was neural networks (NN). Unlike RKHS, where the basis functions used in the regression are imposed a priori, NN methods are more flexible because they are potentially capable of learning those functions from the data (González-Camacho et al., 2012). In this study, two types of single hidden layer NN were investigated: BRNN and RBFNN.

A single hidden layer feed forward NN can be seen as a two-step regression. First, information coming from the
predictors or inputs \((x_d)\) is combined with a vector of connection strengths \(b_l = \{b_l\}\) at each of \(t\) hidden neurons \((t = 1, 2, \ldots, s)\), plus an intercept or ‘bias’ \((b_0)\) giving the following linear score: \(v_i^t = b_0 + \sum_j^p b_j^t x_{ij}\) at neuron \(t\). Each of the neuron scores is then transformed non-linearly using an activation function, \(z_i^t = g_i(v_i^t)\). In the output layer, the \(s\) data-derived functions are combined to yield the predicted phenotype (i.e. the output, \(y\)) and the model becomes:

\[
y_i = \mu + \sum_{t=1}^s w_t z_i^t + \epsilon_i
\]

where \(w_t\) is the regression of the phenotype on the transformed score \(z_i^t\). The regression parameters are then \(\beta = (\beta_1, \beta_2, \ldots, \beta^y)\) and \(w = (w_1, w_2, \ldots, w_s)^\top\). All neural network models used here employed the identity function in the output layer. The two variants of NN used here basically differed in the basis function used in the neurons of the hidden layer, and in the way the network was constructed and its parameters estimated. Although NN methods have a high ability of reproducing complex patterns when learning from data, they are prone to over fitting the training dataset, impeding good predictions once new data is presented to the network. This can be avoided, in part, by using regularized NN methods in which a compromise between model goodness of fit and complexity of the network is achieved (e.g. Gianola et al., 2011; Okut et al., 2011).

Gianola et al. (2011) adapted the methodology for BRNN (e.g. MacKay, 1992) to the prediction of complex phenotypes through the use of pedigree and genomic data. The infinitesimal model and the linear regressions on markers are special cases. Let \(\Psi\) be the parameters in the neural network (connection strengths, hidden to output weights and hidden and output biases). Under a BRNN, estimates of parameters are obtained in two steps. In the first step, the posterior modes of \(\Psi\) given variance parameters \(\sigma^2_x, \sigma^2_y\) are obtained. This is equivalent to minimizing the ‘augmented’ sum of squares, for example, Gianola et al. (2011),

\[
F(\Psi) = \eta \sum_{i=1}^n e_i^2 + \alpha \sum_{j=1}^m \Psi_j^2
\]

where \(\eta = (2\sigma^2_x)^{-1}, \alpha = (2\sigma^2_y)^{-1}\), and \(e_i\) is the model residual. Estimation of the variance components in the second step is via maximization of an approximation to the marginal likelihood of the data, \(p(y|\sigma^2_x, \sigma^2_y)\), a procedure known as empirical Bayes. Steps one and two are iterated until convergence.

RBFNN (Broomhead and Lowe, 1988) is another NN method that uses a Gaussian radial basis function in the hidden neurons:

\[
z_i^t = \exp[-h_t \|x_i - c_t\|^2]
\]

The function is defined by the Euclidean norm between the input vector \(x_i\) of the individual and a centroid vector \(c_t\). Here, \(h_t\) is a bandwidth parameter that needs to be determined empirically. An algorithm was used to select orthogonal centroids sequentially in order to decrease the mean square error between the observed response and the predicted output (Chen et al., 1991). Number of centroids, that is, the number of neurons, grows up to a certain desired precision or up to the number of inputs. To make computations feasible, we used as inputs to the networks either the additive genomic relationship matrix (BRNN_G and RBFNN_G models, for Bayesian regularized neural networks and radial basis functions neural networks, respectively) or principal component scores of the SNP matrix (BRNN_UD and RBFNN_UD models, for Bayesian regularized neural networks and radial basis functions neural networks, respectively).

**Pedigree-based linear model.** As a benchmark, we also assessed the predictive ability of a pedigree-based linear mixed-effects model (PED), as follows:

\[
y = \mu + a + \varepsilon
\]

where \(a\) is the vector of additive genetic effects which, according to the standard assumptions of the infinitesimal model, is assumed to be normally distributed as: \(N(a|0, A\sigma^2_a)\), where \(\sigma^2_a\) is the additive genetic variance and \(A\) is the numerator additive genetic relationship matrix among individuals, obtained from the pedigree. Residuals were normally distributed, as described previously. The prior distributions for \(\sigma^2_a\) and \(\sigma^2\) were scaled inverse \(\chi^2\) distributions, each with 4 degrees of freedom and a scale parameter equal to 1.

**Markov chain Monte Carlo (MCMC) implementation and software used**

Hyper-parameters \(d_f\) and \(S_i\) of the scaled inverse \(\chi^2\) distribution of \(\sigma^2_f\), for \(l \in \{b, u, c\}\), were set to 4 and 1 in each of the BRR, BL and GBLUP models; this gives a prior density with a prior expectation of one half of the sample variance of the standardized phenotypes (González-Camacho et al., 2012). The rate and shape parameters of the prior gamma distribution in the BL models were set to \(1 \times 10^{-4}\) and 0.6, respectively. More details regarding specifications, guidelines for choosing hyper-parameters and computational implementation of these models can be found in de los Campos et al. (2009b) and Pérez et al. (2010). The degrees of freedom and scale parameters of the scaled inverse \(\chi^2\) distributions of \(\sigma^2_f\) and the \(\sigma^2_v\) of the RKHS models, were set to 5 and to half the variance of the phenotypes, respectively. The R-package (R Core Team, 2012) BLR (de los Campos and Pérez, 2012) was used to implement BRR and BL. All RKHS models and the PED model were implemented using a script provided by de los Campos et al. (2010). Briefly, single chains of 100 000 iterations were run, with the first 50 000 iterations discarded as burn-in. Samples drawn from the posterior distributions of the parameters of interest were saved every 100 rounds and then used for statistical inference, with means used as point estimates. The network architecture used in the BRNN models was a single hidden layer with four...
neurons, with the activation functions being hyperbolic-tangent. Models were run using a C++ program (written by the authors and available upon request) but an R-package (R Core Team, 2012) version is also available (BRNN; Pérez et al. 2013). Radial basis functions NN models were implemented using the newrb neural network function from the MATLAB Neural Network Toolbox™ [MATLAB 7.11.0.584 (R2010b), The MathWorks Inc., Natick, MA, USA, 2000]. For each of the RBFNN, the median of the Euclidean distance between each pair of vectors of the input matrix (i.e. elements of G or from the UD) was the value used as bandwidth. A sum-squared error of 0.02 was the precision goal assigned for stopping the algorithm.

Assessment of predictive ability of yet to be observed phenotypes

A 10-fold cross-validation was used to test the predictive performance of these models within each of the lines A, B and AB. For each line, the dataset was randomly split into approximately 10 equal parts. Marker effects and the overall mean were estimated based on a collection of nine portions (90%) of the data, referred to as training set, and predictive ability was assessed in the remaining (10%), used as testing set. Such a training–testing cycle was repeated ten times in total by rotating the 10 portions for training and testing, with each portion used for testing only once and for training nine times. Predictive ability of each model for each dataset was evaluated via Pearson's correlation (r) between observed and predicted phenotypes in testing sets.

Results and discussion

Predictive ability across models

The three panels in Figure 2 show the average correlations between observed and predicted litter size phenotypes in the testing sets from the 10-fold CV in lines A, B and AB, respectively. In general, the ability of predicting litter size was low (r ranged from 0.15 to 0.23 with a standard deviation across the 10-folds that ranged from 0.05 to 0.11 across models). This could be because of the low heritability and repeatability of litter size. Heritability estimates for average litter size obtained with the PED model were 0.21, 0.14 and 0.19 for A, B and AB lines, respectively. Reliabilities of genome-based predictions of fertility traits in dairy cattle have been lower than those obtained for more heritable traits, such as milk production and composition (Hayes et al., 2009). Nevertheless, several authors have argued that genome-based prediction could be of special interest for increasing rate of genetic gain for lowly heritable traits if a sufficiently large number of records per individual is available for training the models (Ibáñez-Escriche and González-Recio, 2011). If one assumes a repeatability of 0.15 for litter size, the correlation between one record or the average of three records of an individual and a future record (Falconer and MacKay, 1996) is expected to range between 0.15 and 0.35, respectively. Hence, our correlations were similar to those expected from a repeatability model with only one or two records available for prediction of future litter size. However, genomic information allows obtaining predictions much earlier in life than when using phenotypic data only. Selection decisions for litter size are made before a progeny record is available, so a model with predictive ability equivalent to that from a record of one progeny represents significant cost-savings to the industry.

Although genomic selection (GS) has been already applied in dairy cattle breeding programs, research on benefits for other species are scarce (Ibáñez-Escriche and González-Recio, 2011). Lillehammer et al. (2011) reported that use of genomic information in pig breeding could improve genetic gain by reducing generation interval, increasing selection accuracy, and reducing inbreeding rate, when compared with conventional selection without progeny testing. Further-
more, there is early evidence from commercial data that pig breeding can benefit from GS as well. For example, Cleveland et al. (2010), using de-regressed EBVs as phenotypes, reported accuracies of GEBV of 0.65 and 0.50 for total number born and stillborn, respectively, which are similar to those reported in dairy cattle (VanRaden et al., 2009).

**Predictive ability of genome-based linear models**

The five linear genome-enabled prediction models showed small differences in predicting future litter size in the three datasets. On average, the correlation ranged from 0.20 to 0.23 (Figure 2). Within the linear models, BRR_UD had the lowest predictive performance in each of the three lines. In general, predictive ability was not affected much by type of shrinkage (i.e. prior distribution on substitution effects) or by type of genomic information (individual markers v. principal component scores).

**Predictive ability of genome-based non-linear models**

RKHS performed similarly to the linear models and, on average, the correlation was −0.23 in all three datasets. RKHS can be implemented easily using standard mixed model packages (de los Campos et al., 2009a; Piepho, 2009), and it is computationally efficient because it has $n + 1$ unknowns only, which is typically much smaller than the $p$ unknowns in marker-based BRR or BL. Use of RKHS maybe of special interest in crossbred populations, because dominance can be represented in an additional kernel, in which case there would be an additional $n$ unknowns, as opposed to $2p$ effects in a model with additive and dominance effects of each of the SNPs. It has been shown that RKHS can outperform linear models in predictive ability in chickens (González-Recio et al., 2008) and plants (Crossa et al., 2010; González-Camacho et al., 2012). Heslot et al. (2012) compared the predictive performance of different linear and non-linear GS models for several traits in plants and found that, overall, linear and non-linear models performed similarly. RKHS seemingly gave the best predictive ability: for example, in 16 out of their 18 comparisons RKHS outperformed Bayes Cπ (Habier et al., 2011).

The two neural network models gave mixed results. On average, the correlation ranged from 0.16 to 0.23 for RBFNN and BRNN. The predictive ability of BRNN varied drastically in the three lines. They delivered the worst predictions in A but were best in AB. The BRNN_UD and BRNN_G yielded the best predictive ability in the crossbred dataset, where non-additive genetic effects such as dominance are to be expected. Using different network architectures, Gianola et al. (2011) found that BRNN outperformed a linear model when predicting yield traits in dairy cows (Gianola et al., 2011). By reducing the number of markers from all the SNPs available, Okut et al. (2011), also found that BRNN could produce a predictive performance similar to linear regression models using mouse data. Pérez-Rodríguez et al. (2012), evaluating a wheat population also found that the predictive ability of BRNN was similar to that of the BL and other linear models. These results suggest that NN models, which are flexible non-linear systems, may predict complex phenotypes better than linear models. Unfortunately our study with litter size could not corroborate this unambiguously, because of the large cross-validation uncertainty. Presumably use of dominance code inputs could enhance the prediction ability of BRNN and RKHS models even further. We found that the predictive abilities of the RBFNN models were higher when using UD than when using G as inputs to the network. González-Camacho et al. (2012) assessed the predictive performance of BL, RKHS and RBFNN in simulated data and in several trait-environment combinations of maize datasets. They found similar predictive ability among different models, but RKHS and RBFNN performed more consistently across traits than BL. Our results indicate that NN have the potential of outperforming other linear and non-linear GS models when predicting litter size under certain conditions, but care must be taken to avoid overfitting. Semi-parametric genomic prediction methods can deliver a better predictive performance than the linear methods provided that kernels, network architecture and inputs are carefully selected (Gianola et al., 2006).

**Predictive ability by type of genomic information used**

A model can be viewed as providing a direct GS approach if the genomic estimated breeding values (GEBV) are estimated directly, as the PED model does in a single step, or as an indirect GS approach if individual marker effects are estimated first, and then the GEBV are computed as the sum of marker effects (times genotype codes) for each individual. For example, GBLUP and RKHS are direct GS approaches, and BRR and BL are indirect. Several simulation studies have shown that the predictive ability may differ between these two types of approaches, depending on marker coverage of the genome and, perhaps, the genetic architecture of the target trait (de los Campos et al. 2012a). However, some models that can be formulated in both manners, such as GBLUP and Ridge regression BLUP should not show differences in prediction ability. In our data, different GS models had similar predictive ability. Likewise, Maltecca et al. (2012) did not find marked advantages of several direct GS models (BL, Bayes A and semi-parametric methods) over GBLUP in Holstein data. An advantage of direct GS methods is that they deal conveniently with the $p >> n$ issue because the large number of SNP genotypes is used only for computing a (co)variance matrix of genomic kinships among individuals, of dimension $n \times n$, before model fitting. In GBLUP, the $G$ matrix was obtained here following VanRaden (2008), with kinships among individuals based on additive similarities at each locus. In RKHS, Euclidean norms between pairs of vectors of marker genotypes are used to compute kernels. Both types of relationships, that is, $G$ and the kernel matrices, were constructed under the assumption that all SNPs had the same contribution toward explaining the degree of relatedness among individuals. This could create a possible ‘dilution effect’ of those that truly contribute to signal. For instance, alleles with a low allelic frequency that are present in a pair of individuals may be
more informative of kinship than if the two individuals share common alleles.

A better predictive ability from using principal component scores instead of all original marker scores may be because of dimensionality reduction and elimination of redundancy of explanatory variables. However, we did not observe improvement in predictive ability when using principal component-based genomic models. Actually, the predictive ability was lower when principal component scores were used as explanatory variables in BRR (−3.4% on average), but this difference was smaller in BL. Use of the UD matrix as an input led to improved predictive ability (~4% on average) in the RBFNN models, relative to what was observed when the G matrix was the input. Perhaps principal component scores are more informative as an input than the G matrix. However, this was not corroborated with the other type of NN used, since BRNN_G obtained the best predictive results in the B and AB lines and the worst in A. We did not construct a cross-validation distribution, so it may be that these differences are because of chance.

In most cases, the predictive ability of a genome-based model was higher than that from PED. This was probably because of the fact that realized relationships among individuals are captured by marker information. This is in agreement with previous findings (see e.g. de los Campos et al., 2012a). For example, predictive accuracy using various types of genomic relationship matrices was higher than when using a pedigree-based kinship matrix in pigs (Forni et al., 2010). Likewise, reliabilities for predicting daughter yield deviations of several traits for Holstein bulls were higher when genomic information was used than when only parent averages were employed (VanRaden et al., 2009). Our results are consistent with these observations, that is, predictive accuracy can be increased if genomic information is used for predicting the phenotypes of candidates.

**Predictive ability in the different lines**

The average correlation between observed and predicted phenotypes in testing sets from all models was higher in the crossbred (0.26) than in the purebred lines (0.15 and 0.22, for A and B lines, respectively). Posterior mean estimates of the additive genetic variance obtained with PED were 0.94, 0.50 and 1.02 for A, B and AB lines, respectively suggesting that the amount of genetic variance in crossbreds was larger. Within purebred lines, predictive ability for the B population was higher than in A, although the B line had fewer animals (1604 v. 2598 sows) and heritability was slightly lower than in A (0.21 and 0.14 for A and B lines, respectively). These differences in predictive ability may be because of differences between lines in effective sample size, or perhaps are related to hidden population structure. Summary statistics for the three datasets are given in Table 1. Raw average litter size per sow was slightly larger in the AB line. This is probably a consequence of inbreeding depression recovery, capture of heterosis and complementarity effects (Falconer and Mackay, 1996) on reproductive performance. The average number of records per sow varied from 2.26 in the B line to 2.72 in the AB line. Average minor allele frequency was similar for the three lines. Sows from the AB line had a higher level of heterozygosity (7% more) than in purebred lines. This is also reflected by the size of diagonal elements of the G matrices of the three lines (Figure 3a). The magnitudes of the diagonal entries of G increase with the number of loci that are homozygote in an individual, and these elements were larger in the A and B lines than in AB. Possibly, selection for improvement of maternal traits may have may have caused fixation of favorable alleles at loci in chromosome regions in strong linkage disequilibrium with causal variants (Falconer and Mackay, 1996). This would also contribute toward an increased degree of inbreeding in the purebred lines, which dissipates in their cross as it can be noticed in the magnitudes of the diagonal entries of A (Figure 3b).

Histograms of the off-diagonal elements of the G matrices (Figure 4a–c) indicated that these were larger and more variable in purebred lines than in AB, and the distribution was bimodal in B and skewed in A. Predictive ability should be better if individuals in training and testing have more molecular similarity or genetic relatedness (de los Campos et al., 2012b). However, if genomic similarity among individuals is

![Figure 3 Boxplots of the diagonal elements of (a) the genomic relationship matrices (G) and (b) the additive genetic kinship matrix obtained from the pedigree (A) for the two purebred (A and B) and the crossbred lines (AB).](https://www.cambridge.org/core/core/terms, https://doi.org/10.1017/S1751731113001389)
high in a population, this reduces effective population size, accordingly (Falconer and MacKay, 1996). If this is the case, a large sample size may be required for obtaining reliable predictions, impairing predictive ability within purebreds relative to what would be observed in crossbred lines. The bimodal or skewed distribution of off-diagonal elements of $G$ in the purebred lines may be suggesting hidden population structure in these lines. Histograms of the off-diagonal elements of the $A$ matrices (Figure 4d–f) were markedly bimodal. Although being more similar in their genome, the degree of relatedness among purebred individuals was lower than in the crossbreds. This could have also contributed to a weaker degree of genetic connectedness between training and testing sets in purebred lines, thus impairing predictive ability.

We examined this issue further by plotting the first two principal components of $G$ and $A$ (Figure 5). Plots of the two first principal components of $G$ showed the presence of clusters only in the A and B lines, but not in line AB. Within purebreds, clustering was more noticeable in A (where the worst predictions were attained) than in B. These clusters indicate that molecular similarity in sows of purebred lines was not homogeneous over individuals, probably resulting from family aggregation, given that clusters were also observed for the purebred lines in the plots of the 2 first principal components of the additive genetic kinship matrix obtained from the pedigree. This also suggests that a cross-validation distribution from random partitions into training and testing datasets may be misleading, as one would expect that predictive ability is larger within than between clusters (Habier et al., 2010). The level of linkage disequilibrium (LD) between markers and QTLs impacts prediction accuracy, and this is impaired if individuals in training and testing are not genetically related (de los Campos et al.,
Acknowledgments

for litter size in pigs. NN showed some promising results enabled prediction outperforms pedigree-based predictions to some extent with how the genome-based information probably effective sample size had an influence on predictive especially for lowly heritable traits, such as litter size. Char-
a higher predictive ability than a pedigree-based model. The SNPs was compared when predicting litter size phenotypes
The performance of several models including information on accounting for population structure in prediction were
between markers and QTL playing a smaller role. Methods for accounting for population structure in prediction were investigated by Janns et al. (2012).

Conclusions

The performance of several models including information on SNPs was compared when predicting litter size phenotypes in three pig lines. Overall, predictive abilities were low and similar over models, but all genome-enabled models had a higher predictive ability than a pedigree-based model. The superior performance of SNP-based models is encouraging, especially for lowly heritable traits, such as litter size. Characteristics of the datasets such as population structure and probably effective sample size had an influence on predictive performance of the models. Predictive ability also varied to some extent with how the genome-based information was treated, for example, models on markers, on principal components or on genomic relationships.

In conclusion, our results support the view that genome-enabled prediction outperforms pedigree-based predictions for litter size in pigs. NN showed some promising results under certain scenarios, but were quite vanish.

Acknowledgments

This research was supported by the Wisconsin Agriculture Experiment Station and by Genus PIC (Hendersonville, TN, USA). The authors thank Juan Manuel González Camacho (Colegio de Postgraduados, Texcoco, Montecillo, Estado de México, México) for useful comments and suggestions.

Supplementary materials

For supplementary material referred to in this article, please visit http://dx.doi.org/10.1017/S1751731113001389

References


package=BLR.


Downloaded from https://www.cambridge.org/core. IP address: 54.191.40.80, on 19 Apr 2017 at 06:21:32, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. doi:10.1017/S1751731113001389
Genome-enabled prediction of litter size in pigs