DOI: 10.1534/genetics.108.088575

Performance of Genomic Selection in Mice

Andrés Legarra,[1] Christèle Robert-Granié, Eduardo Manfredi and Jean-Michel Elsen

INRA, UR 631, F-31326 Auzeville, France

Manuscript received February 27, 2008
Accepted for publication July 10, 2008

ABSTRACT

Selection plans in plant and animal breeding are driven by genetic evaluation. Recent developments suggest using massive genetic marker information, known as "genomic selection." There is little evidence of its performance, though. We empirically compared three strategies for selection: (1) use of pedigree and phenotypic information, (2) use of genomewide markers and phenotypic information, and (3) the combination of both. We analyzed four traits from a heterogeneous mouse population (http://gscan.well.ox.ac.uk/), including 1884 individuals and 10,946 SNP markers. We used linear mixed models, using extensions of association analysis. Cross-validation techniques were used, providing assumption-free estimates of predictive ability. Sampling of validation and training data sets was carried out across and within families, which allows comparing across- and within-family information. Use of genomewide genetic markers increased predictive ability up to 0.22 across families and up to 0.03 within families. The latter is not statistically significant. These values are roughly comparable to increases of up to 0.57 (across family) and 0.14 (within family) in accuracy of prediction of genetic value. In this data set, within-family information was more accurate than across-family information, and populational linkage disequilibrium was not a completely accurate source of information for genetic evaluation. This fact questions some applications of genomic selection.

THIS work evaluates the empirical performance of the so-called genomewide selection strategy for marker-assisted selection (MAS) in a mouse population, using massive molecular marker information. MAS techniques are advocated as a tool for more efficient selection schemes in plant and animal populations (Dekkers and Hospital 2002). In general, MAS techniques are based on tracing the inheritance of the quantitative trait loci (QTL) of interest throughout the pedigree with the help of molecular markers. However, the use of MAS techniques in animal populations is still not much extended, because of its complexity in practice (Boichard *et al.* 2006) and relatively small additional gains. For example, Chamberlain and Goddard (2006) estimated by cross-validation an increase in prediction accuracy over pedigree index (no use of MAS) of, at best, 0.02. Recent developments in massive single-nucleotide polymorphism (SNP) marker genotyping increased the interest for MAS techniques. Dense marker maps capture much richer information, including not only recombination events in the genotyped pedigree (*i.e.*, linkage analysis) but also the populational linkage disequilibrium pattern in the genome, *i.e.*, the possibility of predicting alleles at some loci on the basis of alleles in other (possibly close) loci. This allows for a much finer description of the genome.

[1]*Corresponding author:* INRA-SAGA, BP52627, 31326 Castanet Tolosan Cedex, France. E-mail: andres.legarra@toulouse.inra.fr

Genetic evaluation methods and application issues in MAS in livestock have been extensively described (Fernando and Grossman 1989; Fernando and Totir 2003; Boichard *et al.* 2006). Roughly, this is a two-step process: first, putative QTL locations have to be found in a resource population; later, inheritance of these QTL loci is traced through linkage analysis, and this information is used to estimate breeding values. There are two sources of inefficiency in this approach: first, the fact of having to "declare" (usually by a statistical test in a QTL detection experiment) QTL locations implies that only a few QTL are used, due to lack of power, and their sizes are usually biased upward by the "Beavis" effect (Lynch and Walsh 1998). Second, at least at the first stage, linkage equilibrium has to be assumed between markers and QTL at the founders; this results in loss of across-family information and lower accuracies. However, linkage disequilibrium (LD) analysis (*i.e.*, association between markers and QTL) is more effective because it applies to within- and across-families selection and because the phase of the QTL can be predicted across families (Boichard *et al.* 2006).

Marker-assisted selection techniques considering several QTL loci exist (Lande and Thompson 1990; Villanueva *et al.* 2005). Genomewide selection or genomic selection is a term used by Meuwissen *et al.* (2001). These authors overcome the problems of linkage analysis by fitting a mixed linear model with the effects of thousands of two-marker haplotypes or individual marker loci. In their simulations, breeding

values were estimated with high accuracies, up to 0.85. There were two major insights in their work: to reject the previous stage of QTL position testing and to assume that, for dense marker maps, LD information alone is enough to inform about QTL effects. Piyasatian *et al.* (2007) showed that genomic selection is also valuable in the case of crossings between two inbred lines, requiring a much smaller number of genetic markers.

The main assumption of both methods is that most QTL explaining genetic variation are in linkage disequilibrium with available genetic markers, an assumption that is met in the simulations (Meuwissen *et al.* 2001), in the inbred populations (Piyasatian *et al.* 2007), and, to some extent, in crossings of outbred populations (Pérez-Enciso and Varona 2000). It is unknown to what extent this holds in outbred populations. Nevertheless, if promises from genomewide selection are fulfilled, very economically efficient selection schemes can be set up (Schaeffer 2006; Dekkers 2007).

There is little empirical evidence of practical performance of genomewide selection. Sölkner *et al.* (2007) compiled several approaches with evidence of high accuracies in dairy bulls, using progeny-test estimated breeding values as a proxy for true genetic values.

The objective of this article is to test the performance of genomewide selection and genetic evaluation, using data from a heterogeneous stock mouse population (Valdar *et al.* 2006a), including 1884 individuals (168 full-sib families) and 10,946 SNP markers and four different traits (weight, growth slope, body length, and body mass index). In addition, to gain insight on models and traits, variance components were estimated for different linear mixed models including genomic information.

Predictive ability (the correlation between predicted and observed phenotypes) was estimated using cross-validation. Its connection with genetic gain and accuracy (correlation between true and predicted genetic value) is shown. Cross-validation techniques considered sampling across families (*i.e.*, choosing entire full-sib families) or within families (splitting families into two), thus disentangling across- and within-family information.

MATERIALS AND METHODS

Mouse population: Recently (Valdar *et al.* 2006a), a population of heterogeneous stock mice was used to finely describe the sources of quantitative genetic variation. This population has been extensively described and analyzed (Mott *et al.* 2000; Mott 2006; Solberg *et al.* 2006; Valdar *et al.* 2006a,b). We refer here to the relevant aspects for this work. The data are freely available at http://gscan.well.ox.ac.uk/. The origin of this population is a crossing of eight inbred strains, followed by 50 generations of pseudorandom mating. This population is valuable for testing genomewide selection because, due to the high number of markers, it is expected that many (about three of every five) QTL loci will be in complete LD with marker loci (Mott *et al.* 2000). Indeed, the extent of LD in this population is small (Valdar *et al.* 2006a), which indicates high resolution: average R^2 among two loci falls from 0.5 within 2 Mb to 0.2 within 8 Mb, and average R^2 among adjacent loci is 0.62. The family structure and history of the population are known and therefore interpretation of the results is easy.

Only animals with available phenotype and genotype were retained for data analysis. Details on the genotyping techniques and choice of SNP can be found in Valdar *et al.* (2006a). We discarded animals with <10,000 genotyped SNPs. Our data set was composed of 1884 individuals with 10,946 polymorphic loci (SNPs). Of these, some genotypes were missing, in a very low frequency of 0.001. These missing values should have a negligible effect on the analysis. To simplify the analysis, we imputed them at random from their allelic frequencies; no attempt of reconstruction based on family information was made. Pedigree extended over 2272 individuals. Genealogical information is available on parents of phenotyped mice but not on their grandparents. No parent of a phenotyped animal has been phenotyped itself. This genealogy is roughly organized into 168 full-sib families with 11.21 offspring on average.

We chose four morphological traits: weight at 6 weeks (hereinafter weight), growth slope, body mass index, and body length. The heritabilities of these traits are 0.74, 0.30, 0.21, and 0.13 (Valdar *et al.* 2006b). Environmental covariates affecting those traits include sex for weight and growth slope, and body weight, season, month, and day for body mass index and body length; moreover, there is a "cage" effect considered as random (Valdar *et al.* 2006b). To simplify the analysis, we used the precorrected (by fixed effects, but not cage) data, which are available at http://gscan.well.ox.ac.uk/. Analysis with true phenotypes for weight gave very similar results. An overall mean was added to these residuals and included in the model.

A note of caution has to be made about the cage effect. The allocation of animals to cages is not at random—most animals in the cage are full sibs. From 359 cages in the data, there are just 8 cages with offspring from more than one sire; conversely, each full-sib group is allocated to an average of 2.84 cages. Therefore, it can be considered that cage is a random effect almost nested within the sire effect. This means that, in the absence of a polygenic additive effect, the cage effect might take into account part of the (genetic) family effect.

Two types of methods were needed in this study. The first type is the use of different statistical/genetic models to estimate genetic value, conditionally on marker genotypes, phenotypes, and pedigree. These models are described in the following. The second type is the empirical evaluation of these estimates by cross-validation.

Models for genetic evaluation: The following describes the linear mixed models (in the spirit of BLUP; Lynch and Walsh 1998) that were used for prediction of genetic and environmental values. In short, the models were as follows: model 1, including polygenic (or infinitesimal) effects, without using genomic information (this is the most typical model in genetic improvement nowadays); model 2, including genomic information (SNP genotypes) but not polygenic effects; and model 3, which considers both. In addition to the genetic effects, all models included a random cage effect. The details of the models are as follows.

Model 1—classical polygenic model: This is the model of choice nowadays in applied animal breeding and does not rely on molecular information. This model can be expressed, in matrix algebra notation, as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Tu} + \mathbf{Sc} + \mathbf{e},$$

where **b** is a vector of environmental effects (an overall mean), **c** is a vector of cage effects, and **u** is a vector of additive genetic polygenic effects; and **X**, **S**, and **T** are the corresponding

design matrices. As usual, residuals **e** are assumed independent and to follow a normal distribution, $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$. We assumed **c** and **u** to be random normal effects with *a priori* normal distributions

$$\mathbf{c} \sim N(\mathbf{0}, \mathbf{I}\sigma_c^2), \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2),$$

where **I** is the identity matrix and **G** is the additive genetic relationship matrix (Lynch and Walsh 1998). It is worth remarking that the purpose of genomewide selection and in general of any MAS strategy is to be of better predictive ability than this model.

Model 2—marker-locus effects model: The basic model including SNP effects can be described as follows. Consider n SNP loci. In the jth locus, there are two possible alleles for each SNP (say 1 and 2), and there are three possible genotypes: "11," "12," and "22." We arbitrarily assign the value $+\frac{1}{2}a_j$ to the allele 1 and the value $-\frac{1}{2}a_j$ to the allele 2. This follows a classical parameterization in which a_j is half the difference between the two homozygotes (Lynch and Walsh 1998). These are the additive effects of the SNPs and they can be thought of as classical substitution effects in the polygenic model. It is possible to further postulate a "dominant" effect, assigning the value d_j to the heterozygous genotype, 12. After preliminary analysis we discarded this option as it did not increase predictive ability of the different methods (not shown). Therefore the effects of the different genotypes are $+a_j$ for 11, 0 for 12, and $-a_j$ for 22. The effects of the different genotypes at the n loci sum up to form the genetic effect. The model for the phenotype (ignoring other effects for the sake of clarity) is

$$y_i = \sum_{j=1}^{n} (z_{ij} a_j) + e_i,$$

where y_i is the phenotype of the ith animal, z_{ij} is an indicator covariate for the ith animal and the jth SNP locus, and e_i is a residual term. Hereinafter and for the sake of clarity we refer to $+a_j$ as "marker-locus effects." A marker-locus effect represents the effects on phenotype of unobserved genes (QTL) that are in partial or complete linkage disequilibrium with the marker locus.

If environmental effects are included, and in matrix notation, the model becomes

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Sc} + \mathbf{e},$$

where **a** is the marker-locus effects and **Z** is the corresponding design matrix. It is possible to fit **a** as a "fixed" effect, but for the case of large number of effects and small number of records, the predictive ability will be very poor (Lande and Thompson 1990; Miller 1990; Meuwissen *et al.* 2001). Therefore, we assume that **a** follows a normal distribution, $\mathbf{a} \sim N(\mathbf{0}, \mathbf{I}\sigma_a^2)$. Meuwissen *et al.* (2001) used *a priori* information for σ_a^2 (they divided the polygenic variance by the number of SNP loci), which matched their simulated population. Our attempt to do so resulted in worse predictive abilities (not shown). As for the BayesA and BayesB approaches, substantial *a priori* information is needed (number of segregating loci and variances) that we did not try to guess at.

Model 3—marker-locus effects model and polygenic component: A simple extension of the previous model is to consider, in addition to marker-locus effects, polygenic components:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Tu} + \mathbf{Sc} + \mathbf{e}.$$

The polygenic component **u** here can be thought of as fitting the genes not accounted for by the marker-locus effects in **a**.

Cross-validation: There is extensive literature in model selection techniques, some of whose criteria have been applied in animal breeding. In this work we used cross-validation. This is a robust, nonparametric technique for model selection. The method consists of splitting the data **y** into a training data set ($\mathbf{y}_1$) and a validation data set ($\mathbf{y}_2$). Model parameters are estimated in the training data set. Parameter estimates from $\mathbf{y}_1$ are then used to predict observations in the validation data set (*i.e.*, $\hat{\mathbf{y}}_2 \,|\, \mathbf{y}_1$). A function of interest among the predicted and true observations summarizes the performance of the model and is assumption free and comparable across models. We used Pearson's correlation among predicted and realized observations in the data set. Cross-validation has also an interpretation in terms of efficiency of genetic improvement; this is further developed in the discussion.

In the following, we talk of the correlation $r(\hat{\mathbf{y}}_2, \mathbf{y}_2)$ as the "predictive ability" (of unobserved records), whereas we keep the term "accuracy" for the correlation $r(g, \hat{g})$ between total genetic value of an individual (g) and its estimate ($\hat{g}$). Accuracy can be approximately estimated from predictive ability, as shown in the appendix. In our work, $\hat{y} = \hat{g} + \hat{c}$, where c is the cage effect. Differences in accuracies among models can be estimated by $\Delta r(g, \hat{g}) = \Delta r(y, \hat{y})/(H\Omega)$, where $H^2 = \sigma_g^2/(\sigma_g^2 + \sigma_c^2 + \sigma_e^2)$ and $\Omega^2 = \sigma_{\hat{g}}^2/(\sigma_{\hat{g}}^2 + \sigma_{\hat{c}}^2)$. Estimates for these variance components were obtained from model 1 in Table 1.

Training and validation sets: A key feature in cross-validation is the choice of the training and validation sets. The first choice is the size of each set, as there is a trade-off between precision of the model in the training set and overfitting in the validation set. Usual recommendations are the validation set to be one-fifth or one-tenth of the full data set. However, we have chosen to split the data set into half for training and half for validation because we consider that, in this context, 1000 animals should be enough to get good estimates of the model. For example, Meuwissen *et al.* (2001) fitted 50,000 effects to a data set comprising 2000 records.

The second and more critical choice is how to split the data into training and validation sets. As explained above, the mouse population is composed of several full-sib families with little or no known relationship among them. We devised two options. The first option is to sample whole families; *i.e.*, we use across-family information. The second option is to randomly split every family into two; *i.e.*, we use within-family information. Note that prediction is based on different kinds of information in each setting. For example, when within-family information is used, performance is predicted basically from full sibs using relationships (in models 1 and 3) and from full-sibs and other families via genomic information (in models 2 and 3). On the other hand, when across-family information is used, performance is predicted from other families via genomic information (in models 2 and 3), but it is not possible to use relationships across families (in models 1 and 3) as these are unknown. Loosely speaking, across families, the genetic ties between training and validation data sets are distant relationships (*i.e.*, at the level of grandparents) and populational LD. However, within families, the genetic ties are much stronger, being close relationships as well as populational LD. In practical selection schemes, this is a more likely setting; for instance, prospective bulls are chosen among sons of bulls with good estimated breeding values. The fact of splitting data in two different ways implicitly evaluates the relative weight of each source of information. Splitting was repeated at random 10 times to ensure that the results were not due to random sampling of the data, providing empirical estimates of the standard errors.

ANALYSIS

Variance component estimates with the full data set: The first set of analyses consisted of parameters estimation for the different models, using the whole data set. Although these estimates are not of direct interest for the main subject of this work (efficiency of genomic selection), they help to clarify the different models. Parameters estimates were obtained in a Bayesian framework from the posterior distributions, using MCMC (in particular, Gibbs sampling).

Cross-validation of the genomewide predictive ability: Values for the different unknowns (marker locus effects, polygenic effects, cage effects) were estimated, from the training data set, using Henderson's mixed-model equations (Lynch and Walsh 1998). Means of the marginal posterior distributions for the unknowns in the model were estimated in a Bayesian framework, using Gibbs sampling as well. This marginalization maximizes accuracy and expected genetic progress (Gianola and Fernando 1986), accounting for uncertainty in parameters (in particular, variance components), which might be very high for such small data sets. As discussed before, it was difficult to come up with adequate priors. Flat priors were thus used for variance components and fixed effects. In the cross-validation step, the correlation between the observed and predicted—from the estimates in the previous steps—performances in the validation data set was computed.

Computational issues: In variance component estimation and cross-validation, we used Henderson's mixed-model equations (Lynch and Walsh 1998) to compute solutions for the different models, using homemade software (available on request from the authors). Flat priors were used for variance components. Computing requirements (time and memory) of the mixed-model equations under these models are formidable, because for these models the matrix of crossproducts $\mathbf{Z}'\mathbf{Z}$ to be included is of big size (10,946 × 10,946) and almost 100% dense. To alleviate this problem, the Gauss–Seidel with residual updating strategy was used (Legarra and Misztal 2008).

RESULTS

Variance component estimates with the full data set: Table 1 summarizes estimates. The results differ for weight with respect to the other three traits. For growth slope, body length, and body mass index, estimates of cage and residual variance are fairly constant across models. This is not the case, though, for weight, where cage variance is inflated when the polygenic effect was not included in the model (model 2). Estimates of the residual variance indicate roughly the same fit for all models for growth slope, body length, and body mass index, but a loss of fit for weight when the polygenic term was not fit.

TABLE 1

Variance components estimates for different models of genomic selection

Model	σ^2_a	σ^2_u	σ^2_c	σ^2_e
		Weight		
1		4.59	2.12	0.16
2	3.52*E*-04		3.34	1.94
3	2.52*E*-04	3.56	2.15	0.19
		Growth slope		
1		8.37*E*-04	9.72*E*-04	8.22*E*-04
2	1.04*E*-07		10.30*E*-04	10.79*E*-04
3	1.00*E*-07	2.36*E*-04	9.65*E*-04	9.57*E*-04
		Body length		
1		0.040	0.048	0.146
2	9.09*E*-06		0.051	0.150
3	8.58*E*-06	0.010	0.048	0.144
		Body mass index		
1		2.49*E*-04	3.91*E*-04	18.72*E*-04
2	0.80*E*-07		3.94*E*-04	18.46*E*-04
3	0.77*E*-07	0.67*E*-04	3.75*E*-04	18.08*E*-04

Estimated variance components are shown for marker-locus effects **a**, random cage effects **c**, polygenic additive genetic effects **u**, and residual **e**.

As for the effects of individual SNPs (**a**), they roughly follow normal distributions *a posteriori*. This is as expected because by the nature of the mixed model, they are severely shrunken toward a mean of 0. As an example, for body length the estimated (posterior means) **a** effects in model 2 range between -1.61×10^{-3} and 1.51×10^{-3}, for a trait with a polygenic additive variance of 0.02.

Cross-validation of the genomewide predictive ability: Results are shown in Table 2. Model 1 is the reference, as it is based on phenotype and pedigree information only. Across families, model 1 has a low predictive ability, because there is no family information to rely on, just the common cage environmental effect. Therefore, models 2 and 3 are expected to perform better, due to the molecular information. This is the case and the difference is significant, going up to an increase of 0.22 in predictive ability.

Within families, models including genomewide information slightly outperform model 1 in predictive ability. This gain in predictive ability is not significant, but nevertheless suggestive and fairly consistent across traits. Model 2 always has the better predictive ability in spite of being the simpler one.

Changes in accuracy of the genetic value are shown in Table 3. For the across-families case, accuracies increased up to ~0.5. This is actually not surprising as these accuracies were close to 0 in model 1. For the within-families case, the increase in accuracy in prediction of genetic value ranged from 0 up to 0.14 by using genomic information. It has to be kept in mind, though, that in this case these values are not significant.

TABLE 2

Predictive ability of different models for genomic selection

Trait	Model 1	Model 2	Model 3
	Across families[a]		
Weight	0.07	0.25	0.20
Growth slope	0.04	0.26	0.19
Body length	0.05	0.16	0.12
Body mass index	0.06	0.17	0.12
	Within families[b]		
Weight	0.67	0.67	0.63
Growth slope	0.54	0.55	0.51
Body length	0.24	0.27	0.25
Body mass index	0.32	0.35	0.33

[a] Standard errors ~0.03.
[b] Standard errors ~0.02.

TABLE 3

Approximate increase in accuracy of estimation of the genetic value from model 1 to model 2

Trait	Across families	Within families
Weight	0.26	0.00
Growth slope	0.57	0.03
Body length	0.40	0.11
Body mass index	0.51	0.14

DISCUSSION

Global performance: Genomewide selection tools show in general similar or better predictive ability than classical polygenic methods. This is as expected if data are adequate and underlying assumptions (additivity of QTL effects, strong linkage disequilibrium among at least some markers and QTL) are true (Meuwissen *et al.* 2001). The increase in accuracy of prediction of the genetic value in our study varied, but it was at best at ~0.14. This value is comparable to values found by simulations or other data analysis (Meuwissen *et al.* 2001; Sölkner *et al.* 2007). However, it has to be kept in mind that we have only suggestive results for the within-families case. The estimators of accuracy are approximated and dependent on variance components, which are estimated (see materials and methods). This might be a source of error. However, this is not the case for the predictive ability shown in Table 2. This difference in predictive ability is assumption free and exact (up to numerical error) and clearly shows better predictive abilities of the models including genomic information.

Model 3 shows lower predictive ability than model 2 and even sometimes than model 1 (Table 2). One could expect the opposite, because the polygenic term is expected to catch all the genetic variability not traced by genetic markers. The most likely explanation is double: first, markers capture polygenic resemblance between relatives (see discussion below); second, for this reason, polygenic genetic values and "marker-explained" global genetic values are expected to be extremely collinear, which deteriorates performance of the estimation.

Intriguingly, results (Tables 2 and 3) show more benefit from using genomic selection for decreasing heritabilities. Whereas predictive ability is, as expected, lower for low-heritable traits, the difference in predictive ability and accuracy of the models including SNP information increases with decreasing heritability, with respect to the polygenic model (model 1). If this is confirmed, genomic selection would be a good tool for selecting low-heritable traits.

Role of within- and across-family information: We have split the data in two ways for the cross-validation approach: across and within families. The comparison between these two ways shows the relative performance of each source of information, either close or distant relatives, respectively. In this work, the information from distant relatives is equivalent to the population-level information (indeed even the most distant individuals in a population are distant relatives). Clearly, in this data set, information from distant relatives has a poorer predictive ability (and thus accuracy in prediction of the genetic value) than information from close relatives, as shown in Table 2. Therefore family information should not be discarded for practical use. This fact partly invalidates the assumption of Meuwissen *et al.* (2001) that most genetic variation can be traced by the use of populational linkage disequilibrium between markers and QTL. This also invalidates some proposals of genomic selection (Schaeffer 2006) that assume no need of genotyping and phenotyping close-relative animals. The point has been explored in further detail by Habier *et al.* (2007), who show that some methods (BayesB and a fixed regression) capture better the population LD, which is more useful in the long range, *i.e.*, after several generations.

Model 2, using genomewide SNP information, shows good accuracies and predictive abilities, in spite of not using explicitly pedigree information. This implies that genomewide selection might be of interest for species with difficult or no pedigree tracing, like fish (Martinez 2006), self-pollinating crops, or trees (Bauer *et al.* 2006). The molecular information would not be used to reconstruct the pedigree (probably with errors) but would be used "as is." The accuracy will depend on whether individuals analyzed together are close relatives or not, but these relationships do not need to be known.

Accuracies: In our work we estimated by cross-validation the *increase* in predictive ability (Table 2). These estimates can be compared to accuracies found in simulations (Meuwissen *et al.* 2001) or other real data analysis (Sölkner *et al.* 2007). These authors showed accuracies in the prediction of genetic value of 0.81–0.85, for traits of heritability of 0.5 (Meuwissen *et al.*

2001) and ~1 (progeny-tested estimated breeding values in bulls; Sölkner *et al.* 2007).

Unfortunately, they did not compare their genomewide genetic evaluations with a polygenic model strategy without genomic information, such as model 1. They suggest comparison with parents' information in a polygenic model, which is at best 0.71. Further increase could be achieved, in a polygenic model framework, only by means of the progeny information. While this is true for the simulations in Meuwissen *et al.* (2001), we consider that this assertion is false in Sölkner *et al.* (2007), who found accuracies of ~0.8. They used a complex, real pedigree of dairy bulls. In a cross-validation approach, they sampled four-fifths of these bulls for training and one-fifth for validation. However, most of these bulls are related. It is thus likely that bulls in the validation data set have some descendants in the training data set. For example, assume a bull breeding value is estimated from the information from his father, maternal grandfather, and four sons. This would result in a theoretical accuracy of 0.81 (Van Vleck *et al.* 1987), without using genomic information.

At any rate, their increase in accuracy might be considered to be ~0.10–0.14. Our results (Table 3) are comparable. This is uncertain, though, for differences are not always significant and are higher for low-heritable traits. Conversely, for a trait with heritability of 0.50, the increase of 14% in accuracy found by Meuwissen *et al.* (2001) would be reflected in an increase in predictive ability of 10%, which is comparable as well to our results in Table 2.

Cross-validation in a genetic improvement context: We addressed validation of the genomewide genetic evaluation by cross-validation. Cross-validation has a clear interpretation in a genetic improvement context because it mimics a genetic improvement process. The objective of any breeding program is to improve future performance of the individuals in the population. In practice, the breeding process goes on through the analysis of a series of phenotypes ($\mathbf{y}_1$), to estimate breeding values, these estimators being used to select the next generation that in turn will express its phenotypes ("future" performances, $\hat{\mathbf{y}}_2$). If these predicted future performances are used to make breeding decisions (*e.g.*, producing selected animals in the population expressing $\mathbf{y}_2$), the observed phenotypic gain ΔP depends on the correlation r between $\hat{\mathbf{y}}_2$ and $\mathbf{y}_2$, $\Delta P = i \cdot r \cdot \sigma_{y_2}$, where i is the selection intensity. This equation reduces to the usual breeders' equation (Lynch and Walsh 1998) under the usual assumptions of the polygenic model $P = G + E$, if there is no other random effect.

This holds as well for the across-family cross-validation, where we might want to select individuals on the basis of information from other families. Therefore the correlation among predicted and observed performances in a cross-validation setting is a direct measure of the efficiency of a breeding scheme applying the proposed model to this set of data. Although this approach is robust and assumption free, its interpretation in genetic terms remains problematic as far as there are environmental effects. Another possibility for validation is the use of quasi-true estimated breeding values (Sölkner *et al.* 2007), although this neglects the role of other phenomena such as genotype–environment interactions, epistasis, or dominance.

Other models for genetic evaluation: There is an equivalence between the genomewide marker-locus effects models and models using markers as indicators of relatedness (identity-by-state, IBS) (Caballero and Toro 2002; Habier *et al.* 2007). We can summarize the overall genetic value due to marker-locus effects (v) of individual i as $v_i = \sum_{j=1}^{n} z_{ijk} a_{jk}$, where a_{jk} are individual SNP effects (a_{j1} for allele "1" and $a_{j2}$for allele "2") and z_{ijk} are indicator variables. Assuming small a_{jk} effects, it comes out that the joint distribution of $\mathbf{v}$ is approximately multivariate normal, with mean = 0 and variance = $\mathbf{ZZ}'\sigma_a^2$. Note that v is a "genomic" counterpart of u, the polygenic breeding value; v might be called genomewide or genomic breeding value. It can be shown that $\mathbf{ZZ}'$ is related to the matrix of IBS probabilities, as IBS = $\mathbf{ZZ}'/4n$. The IBS matrix would give the same information as the identical-by-descent (IBD) (the coefficient of coancestry) probabilities if SNPs were fully informative about their origin [*i.e.*, infinite loci and different alleles at each loci for every individual in the base population (Caballero and Toro 2002)]. However, this is not the case here and two copies of the same locus might be IBS without being IBD.

Using genomewide breeding values $\mathbf{v}$, a linear model can be constructed as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Tv} + \mathbf{Sc} + \mathbf{e}$$

$$\mathbf{v} \sim N(\mathbf{0}, \mathbf{ZZ}'\sigma_a^2).$$

The model is equivalent to model 2 in the sense of Henderson (1985), *i.e.*, after appropriate transformation solutions for $\mathbf{v}$ are identical. The joint distribution of $\mathbf{v}$ is well defined for positive-definite $\mathbf{ZZ}'$; *i.e.*, its rank is equal to the size of $\mathbf{v}$ (otherwise, likelihood is 0). We tried this approach, but it led to serious numerical problems ($\mathbf{ZZ}'$ positive definite but of extremely small determinant) and we did not further pursue this option.

IBS can be thought of as a molecular counterpart of the additive genetic relationship matrix $\mathbf{G}$, with more and less information at the same time: more, because we can trace meiosis sampling among full sibs; less, because this matrix assumes that two animals sharing the same molecular information are identical in spite of the possibility of this being just by chance, without one being related to the other. A more refined approach is to use as well the IBD information conditional on molecular and genealogical information, to reflect relationships among relatives. Similar ideas have been used in different con-

texts: genomic control (Yu *et al.* 2006), refining of polygenic models (Visscher *et al.* 2006), or genetic evaluation (Villanueva *et al.* 2005). The latter three condition the IBD state to the available genealogy. The fact that predictive ability is better using within-family information suggests that within-family IBS (*i.e.*, IBD) is important and might be worth modeling.

Two such models are (1) the segment-mapping approach (Pérez-Enciso and Varona 2000), which models total breeding value as a sum of small segments of the genome, allowing for linkage, and (2) the use of marker-assisted relatedness (Villanueva *et al.* 2005; Visscher *et al.* 2006), which models the covariance between relatives using molecular information conditional on genealogy. Another possibility is to avoid explicit modeling; nonparametric methods might have better predictive abilities and are robust to departures from the assumed theory (Fox 2000; Gianola *et al.* 2006).

Conclusion: Our results suggest, but do not prove, that genomewide genetic evaluation and selection have better accuracies and predictive abilities than the classical polygenic model. More traits and studies across different species need to be carried out to further confirm this hypothesis. Results also prove that within-family information, for this data set, is a more accurate source of information than across-family information. This information is relevant for the setup of genetic improvement programs. Cross-validation has been shown to be a valuable tool for this study. Results also show good properties of genomic selection for the case of unrecorded pedigrees, where available tools are scarce. As for its practical implementation, the use of genomic selection will depend on a cost-benefit analysis of recording of DNA samples against expected additional economic gains.

We thank Johann Sölkner, Zulma Vitezica, and Miguel Ángel Toro for discussions. We gratefully acknowledge The Wellcome Trust Centre for Human Genetics, Oxford, for making the heterogeneous stock data available at http://gscan.well.ox.ac.uk.

LITERATURE CITED

Bauer, A. M., T. C. Reetz and J. Léon, 2006 Estimation of breeding values of inbred lines using best linear unbiased prediction (blup) and genetic similarities. Crop Sci. **46:** 2685–2691.

Boichard, D., S. Fritz, M. Rossignol, F. Guillaume, J. J. Colleau *et al.*, 2006 Implementation of marker assisted selection: practical lessons from dairy cattle. Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, MG, Brazil, CD-ROM communication 22–11.

Caballero, A., and M. A. Toro, 2002 Analysis of genetic diversity for the management of conserved subdivided populations. Conserv. Genet. **3:** 289–299.

Chamberlain, A. J., and M. E. Goddard, 2006 Testing marker assisted selection in a real breeding program. Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, MG, Brazil, CD-ROM communication 22–12.

Dekkers, J. C. M., 2007 Marker-assisted selection for commercial crossbred performance. J. Anim. Sci. **85:** 2104–2114.

Dekkers, J. C. M., and F. Hospital, 2002 The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. **3:** 22–32.

Fernando, R. L., and M. Grossman, 1989 Marker assisted prediction using best linear unbiased prediction. Genet. Sel. Evol. **21:** 467–477.

Fernando, R. L., and L. R. Totir, 2003 Incorporating molecular information in breeding programmes: methodology, pp 537–548 in *Poultry Genetics, Breeding and Biotechnology*, edited by W. M. Muir and S. E. Aggrey. CAB International, Wallingford, UK.

Fox, J., 2000 *Nonparametric Simple Regression: Smoothing Scatterplots.* Sage Publications, Thousand Oaks, CA.

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

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

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

Henderson, C. R., 1985 Equivalent linear models to reduce computations. J. Dairy Sci. **68:** 2267–2277.

Lande, R., and R. Thompson, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics **124:** 743–756.

Legarra, A., and I. Misztal, 2008 Genome-wide selection computing strategies. J. Dairy Sci. **91:** 360–366.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.

Martinez, V., 2006 Importance and implementation of molecular markers in selective breeding programs for aquaculture species. Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, MG, Brazil, CD-ROM communication 09–01.

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

Miller, A. J., 1990 *Subset Selection in Regression.* Chapman & Hall, London/New York.

Mott, R., 2006 Finding the molecular basis of complex genetic variation in humans and mice. Philos. Trans. R. Soc. Lond. B Biol. Sci. **361:** 393–401.

Mott, R., C. J. Talbot, M. G. Turri, A. C. Collins and J. Flint, 2000 A method for fine mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. USA **97:** 12649–12654.

Pérez-Enciso, M., and L. Varona, 2000 Quantitative trait loci mapping in F2 crosses between outbred lines. Genetics **155:** 391–405.

Piyasatian, N., R. L. Fernando and J. C. M. Dekkers, 2007 Genomic selection for marker-assisted improvement in line crosses. Theor. Appl. Genet. **115:** 665–674.

Schaeffer, L. R., 2006 Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. **123:** 218–223.

Solberg, L. C., W. Valdar, D. Gauguier, G. Nunez, A. Taylor *et al.*, 2006 A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice. Mamm. Genome **17:** 129–146.

Sölkner, J., B. Tier, R. Crump, G. Moser, P. Thomson *et al.*, 2007 A comparison of different regression methods for genomic-assisted prediction of genetic values in dairy cattle. Proceedings of the 58th Annual Meeting of the European Association for Animal Production, Dublin, p. 161.

Valdar, W., L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman *et al.*, 2006a Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat. Genet. **38:** 879–887.

Valdar, W., L. C. Solberg, D. Gauguier, W. O. Cookson, J. N. P. Rawlins *et al.*, 2006b Genetic and environmental effects on complex traits in mice. Genetics **174:** 959–984.

Van Vleck, L. D., E. J. Pollak and E. A. B. Oltenacu, 1987 *Genetics for the Animal Sciences.* W. H. Freeman, New York.

Villanueva, B., R. Pong-Wong, J. Fernández and M. A. Toro, 2005 Benefits from marker-assisted selection under an additive polygenic genetic model. J. Anim. Sci. **83:** 1747–1752.

Visscher, P. M., S. E. Medland, M. A. R. Ferreira, K. I. Morley, G. Zhu *et al.*, 2006 Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. **2**(3): e41.

Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38:** 203–208.

Communicating editor: C. Haley

APPENDIX: RELATION BETWEEN ACCURACY IN THE ESTIMATION OF GENETIC VALUE AND PREDICTIVE ABILITY:

Let us assume y and $\hat{y}$ are random variables denoting the realization and the prediction of a phenotype. For a realization of y, we can write the simple model $y = g + e$, where g is the overall genetic value and e is a residual term. Variables g and e are assumed uncorrelated. On the other hand, $\hat{y} = \hat{g}$. Therefore, the correlation between observed and predicted phenotype

$$r(y, \hat{y}) = \frac{\mathrm{Cov}(g + e, \hat{g})}{\sqrt{\mathrm{Var}(g + e)\mathrm{Var}(\hat{g})}}$$

can be reduced to

$$r(y, \hat{y}) = \frac{r(g, \hat{g})\sigma_g\sigma_{\hat{g}}}{\sqrt{\sigma_g^2 + \sigma_e^2}\sigma_{\hat{g}}} = r(g, \hat{g})H,$$

where $H^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$, *i.e.*, the broad-sense heritability. Therefore, the "true" accuracy might be obtained as $r(g, \hat{g}) = r(y, \hat{y})/H$.

In the case of our work we assumed three random variables in y: a genetic effect g (with different modelizations across models), the cage effect c, and the residual e. Therefore $\hat{y} = \hat{g} + \hat{c}$. Following a development as above, we have

$$r(y, \hat{y}) = \frac{\mathrm{Cov}(y, \hat{y})}{\sqrt{\mathrm{Var}(y)\mathrm{Var}(\hat{y})}},$$

which might be simplified and split as

$$r(y, \hat{y}) = \frac{\mathrm{Cov}(g, \hat{y})}{\sigma_y\sigma_{\hat{y}}} + \frac{\mathrm{Cov}(c, \hat{y})}{\sigma_y\sigma_{\hat{y}}}.$$

As an approximation, the second term might be assumed to be constant across models, because all of them fitted the cage effect. As for the first term, it can be further developed by expanding $\hat{y}$ into $\hat{g} + \hat{c}$:

$$\frac{\mathrm{Cov}(g, \hat{y})}{\sigma_y\sigma_{\hat{y}}} = r(g, \hat{g})H\frac{\sigma_{\hat{g}}}{\sqrt{\sigma_{\hat{g}}^2 + \sigma_{\hat{c}}^2}} + r(g, \hat{c})H\frac{\sigma_{\hat{c}}}{\sqrt{\sigma_{\hat{g}}^2 + \sigma_{\hat{c}}^2}}.$$

Assuming $r(g, \hat{c}) = 0$ (*i.e.*, the cage effect does not capture genetic information), the second term can be again neglected (this is likely an approximation). In this expression there are pseudoheritability terms that indicate the amount of variation in the prediction explained by the genetic part. This can be assumed to be fairly constant across models.

It is thus possible to refer, at least approximately, differences in the predictive ability $r(y, \hat{y})$ among models to differences in the predictive ability for the genetic component, $r(g, \hat{g})$, as

$$\Delta r(g, \hat{g}) \approx \frac{\Delta r(y, \hat{y})}{H\Omega},$$

where $H^2 = \sigma_g^2/(\sigma_g^2 + \sigma_c^2 + \sigma_e^2)$ and $\Omega^2 = \sigma_{\hat{g}}^2/(\sigma_{\hat{g}}^2 + \sigma_{\hat{c}}^2)$.