Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models

Fernando Macedo, A. Reverter, Andres Legarra

J. Dairy Sci. 103:529–544
https://doi.org/10.3168/jds.2019-16603

Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models

F. L. Macedo,[1,2]* A. Reverter,[3] and A. Legarra[1]
[1]INRA, GenPhySE, Castanet-Tolosan 31320, France
[2]Facultad de Veterinaria, Universidad de la República, 11600 Montevideo, Uruguay
[3]CSIRO Agriculture and Food, St. Lucia 4067, Australia

ABSTRACT

Bias in genetic evaluations has been a constant concern in animal genetics. The interest in this topic has increased in the last years, since many studies have detected overestimation (bias) in estimated breeding values (EBV). Detecting the existence of bias, and the realized accuracy of predictions, is therefore of importance, yet this is difficult when studying small data sets or breeds. In this study, we tested by simulation the recently presented method Linear Regression (LR) for estimation of bias, slope, and accuracy of pedigree EBV. The LR method computes statistics by comparing EBV from a data set containing old, partial information with EBV from a data set containing all information (old and new, a whole data set) for the same individuals. The method proposes an estimator for bias $\left(\widehat{\Delta_p}\right)$, an estimator of slope $\left(\widehat{b_p}\right)$, and 3 estimators related to accuracies: the ratio between accuracies $\left(\hat{\rho}_{w,p}\right)$, the reliability of the partial data set $\left(\widehat{acc_p^2}\right)$, and the ratio of reliabilities $\left(\widehat{\rho_{p,w}^2}\right)$. We simulated a dairy scheme for low (0.10) and moderate (0.30) heritabilities. In both cases, we checked the behavior of the estimators for 3 scenarios: (1) when the evaluation model is the same as the model used to simulate the data; (2) when the evaluation model uses an incorrect heritability; and (3) when the data includes an environmental trend. For scenarios in which the evaluation model was correct, the LR method was capable of correctly estimating bias, slope, and accuracies, with better performance for higher heritability [i.e., $corr\left(b_p, \widehat{b_p}\right)$ was 0.45 for $h^2 = 0.10$ and 0.59 for $h^2 = 0.30$]. In cases of the use of incorrect heritabilities in the evaluation model, the bias was correctly estimated in direction but not in magnitude. In the same way, the magnitudes of bias and of slope were underestimated in scenarios with environmental trends in data, except for cases in which contemporary groups were random and greatly shrunken. In general, accuracies were well estimated in all scenarios. The LR method is capable of checking bias and accuracy in all cases, if the evaluation model is reasonably correct or robust, and its estimations are more precise with more information (e.g., high heritability). If the model uses an incorrect heritability or a hidden trend exists in the data, it is still possible to estimate the direction and existence of bias and slope but not always their magnitudes.

Key words: genetic evaluation, BLUP, bias, accuracy

INTRODUCTION

The study of bias has become more relevant in the last years, as several works have shown differences between the estimated genetic value of top young bulls at genomic prediction and after progeny results (Spelman et al., 2010; Sargolzaei et al., 2012). The most frequently used statistics to analyze bias in selection schemes are as follows: $b_0 = \bar{\hat{u}} - \bar{u}$ [the difference between the averages of estimated breeding values $\hat{u}$ (**EBV**) and true breeding values u (**TBV**)], associated with the genetic gain, and $b_1 = \frac{cov(u,\hat{u})}{var(\hat{u})}$ (slope of the regression of TBV on EBV), related to the dispersion of the EBV. Values of $b_0 < 0$ underestimate and $b_0 > 0$ overestimate TBV. Similarly, values of $b_1 < 1$ represent an overestimation of selected animals. Both biases produce variation in the expected genetic gain, with implications at the moment of selection (Boichard et al., 1995; Mäntysaari et al., 2010).

Studies in Lacaune sheep have shown overestimation of genetic gain ($b_0 > 0$) as well as overdispersion ($b_1 < 1$) of the genomic estimated breeding values (**GEBV**), with more effect in those traits under important selection pressure (Astruc et al., 2014; Baloche et al., 2014). The origin of these biases is unknown, and they should

Received March 12, 2019.
Accepted September 13, 2019.
*Corresponding author: fernando.macedo@inra.fr

not occur under standard assumptions of animal breeding (Henderson, 1984). In pedigree-based predictions, several situations can produce bias, such as the use of incorrect heritability (**h^2**) in genetic evaluations, selective reporting, incorrect modeling of the age effect, an ill-defined contemporary group (**CG**) effect, or the use of genetic groups in pedigrees. In genomic predictions, incorrect models can also generate bias.

Currently, the most widely used tool in animal breeding to benchmark genetic models and detect bias is time truncation of data and prediction of future records or averages of records (e.g., daughter yield deviations, **DYD**). However, this is difficult to do in certain contexts—for instance, in selection programs with small numbers of sires and small numbers of daughters each, or for traits with low heritability (Legarra and Reverter, 2017). In the case of Pyrenean dairy sheep breeds, one of the problems for forward prediction is the existence of few sires, each with small progeny groups (Barillet et al., 2016).

In 2018, Legarra and Reverter presented the Linear Regression (**LR**) method, based on the comparison of EBV obtained from old records (partial data sets) with a data set containing both old and new records (a whole data set). The LR method does not require accurate EBV or precorrected phenotypes and can be used for any kind of traits (e.g., maternal effects on offspring). At the same time, VanRaden and O'Connell (2018) also proposed the use of changes in GEBV to validate published genomic reliabilities, although they did not address the existence of bias per se.

The LR method was formally presented and applied to an example data set (Legarra and Reverter, 2018), but it was not verified in depth. In particular, it assumes that the heritability and the evaluation model are the correct ones, but these assumptions are not always true. In fact, it is of most interest to know whether the LR method can detect an incorrect model. In this work, we used simulations to analyze the potential of method LR to estimate the bias, the slope, and the accuracies of different scenarios: first when the evaluation model is correct, second when the heritability used for genetic evaluations is not correct, and finally when there is an environmental trend in the data that is not explicitly accounted for by the model. These cases may not be the most urgent of topics at present—for instance, bias due to ignoring genomic preselection in BLUP evaluations may be more urgent (Patry and Ducrocq, 2011)—but the aim of this study was to gain a general view of the capabilities of the LR method, especially when the model is reasonable. Only pedigree-based evaluations were considered, given the complexity of genomic predictions for the simulated data.

MATERIALS AND METHODS

Simulations

We simulated a dairy cattle breeding scheme with partially overlapping generations, progeny testing, and selection. Only females were phenotyped, with only 1 record each, because of limitations of the simulation software. Two heritabilities ($h^2 = 0.10$ and 0.30) were simulated. We used the QMSim v. 1.10 software program (Sargolzaei and Schenkel, 2009), and the main parameters of the simulation are shown in Table 1 and the parameter file in Appendix 1. In each generation, 8% of born males and 45% of born females were selected to join the pool of reproducers, provided their EBV was high enough. Accordingly, animals with the lowest EBV in the pool were discarded. The pool of reproducers contains, potentially, animals of all previous generations, and therefore parents of a given generation may came from any of the preceding generations. For instance, in Figure 1, we show an example of the generation of origin of parents of individuals in generation 7. It can be observed that, of 45,000 animals born in generation 7, 1,800 sires were born in generation 6, 1,192 were born in generation 5, and so on. All born females have a single performance. The mating system seeks to minimize inbreeding (mating design = *minf* in QMSim parameter file; Sonesson and Meuwissen, 2000), achieving an average inbreeding, for all generations, close to zero. Instead of using QMSim internal BLUP evaluations, genetic evaluations were performed at the end of each generation, using as external software blupf90 (Misztal et al., 2002). Then QMSim selects individuals with higher external EBV to be parents for the next generation. This scheme allowed us the flexibility required to explore competing scenarios.

We considered 3 different strategies to evaluate the individuals in the population: (1) using the same model as the one used in simulation, (2) using a different h^2 for evaluation, or (3) adding an environmental trend.

Table 1. Main parameters used to simulate populations in QMSim software program (Sargolzaei and Schenkel, 2009)

Parameter	Value
Replicates	20
Generations	10
Sex ratio	0.5
Total animals in populations	~450,000
Phenotype	Only 1 measure in females
Mating system	Inbreeding control
Selection	Higher EBV (BLUP)
Number of chromosomes	30
Number of QTL per chromosome	333

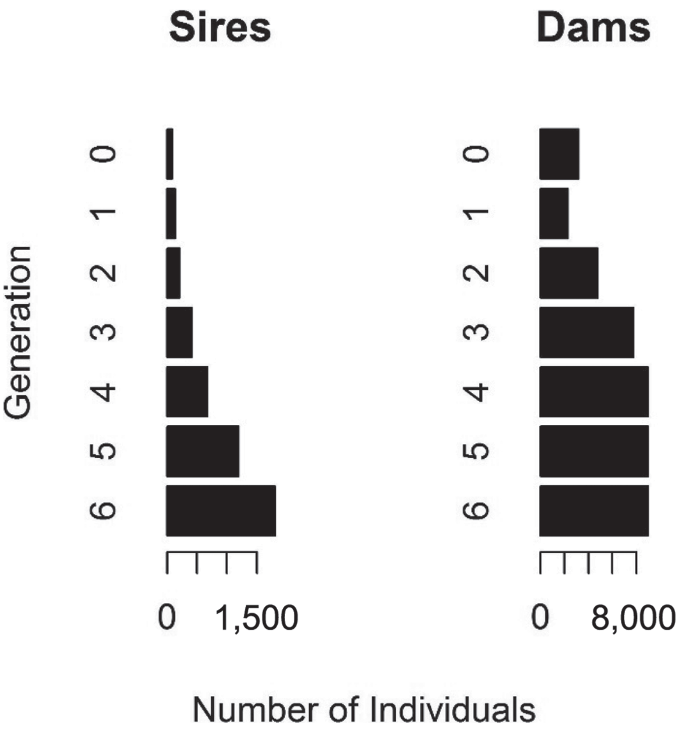

Figure 1. Generation of birth of the parents of 45,000 individuals of the seventh generation. Example of the first replicate of the simulation scenario T10 ($h^2 = 0.10$).

In total, 11 scenarios were obtained: 2 using the correct model to evaluate, 4 using an incorrect h^2, and 5 using an environmental trend effect. In all cases, TBV were simulated as the sum of QTL effects, sampled from a gamma distribution. All simulations used a genetic variance of 1, which implies that units (e.g., of bias) are in genetic standard deviations.

Correct Genetic Model. Phenotypes were simulated, adding an overall mean and a residual deviate to TBV with 2 heritabilities: h^2 of 0.10 (scenario **T10**) or 0.30 (scenario **T30**). These heritabilities mimic, respectively, health traits with low heritability, such as subclinical mastitis, and moderately heritable production traits. The population was evaluated assuming the infinitesimal model (whereas the simulation uses a finite genome) $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \boldsymbol{e}$, where $\mathbf{u} \sim N\left(0, \mathbf{A}\sigma_u^2\right)$, $\mathbf{y}$ is the vector of observations, μ is the overall mean, $\mathbf{Z}$ is the incidence matrix that relates the records to animals, $\boldsymbol{e}$ is the residual, $\mathbf{A}$ is the relationship matrix, and σ_u^2 is the genetic variance; and assuming the variance components used in simulations.

Incorrect Heritability. Phenotypes were simulated as above, with the same 2 heritabilities. However, the models used for genetic evaluation used wrong heritabilities. For simulations performed with an h^2 of 0.1, we used h^2 of 0.05 (scenario **W05**) and 0.15 (scenario **W15**) in the evaluation models, and for data simulated with an h^2 of 0.3, the models for evaluation used h^2 of 0.25 (scenario **W25**) and 0.35 (scenario **W35**).

Environmental Trends. Phenotypes were simulated as the sum of TBV, residual, and environmental trends, as follows. At each generation, an environmental trend was added of the form $t \times k$, where t is the generation number and k is equal to half the genetic progress per generation. An example of phenotypic, genetic, and environmental trend is shown in Figure 2. Then, at each generation, 9 CG with no effect were simulated, and the individuals were assigned randomly to each one. To guarantee genetic connections, CG included 5,000 individuals. The reason for this is that the number of daughters per male is low (approx. 11) and little overlap across generations occurs. Hence, to ensure connectedness, large groups are needed. Previous experimentation with 500 individuals provided very low connectivity, but results were qualitatively similar (data not shown). A sensible model (the "correct" one) for genetic evaluations for these conditions would be a regression on time to account for environmental trend, plus CG: $y_{ij} = t \times k + CG_i + u_j + e_{ij}$.

We tried 2 approaches to perform the genetic evaluation: CG as fixed effect or as random effect. In the

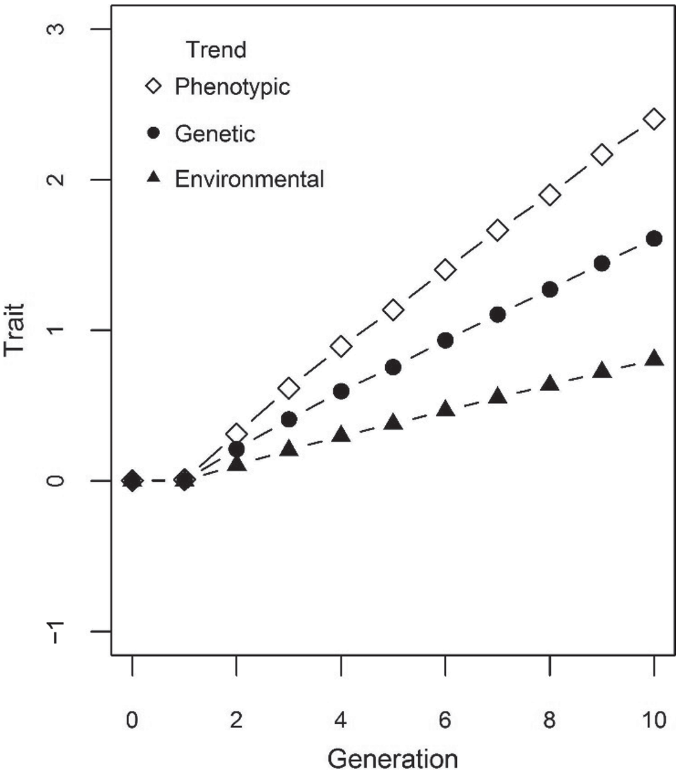

Figure 2. Phenotypic, genetic, and environmental trends corresponding to the first replicate for the simulation scenario FCG30 (environmental trend, $h^2 = 0.30$).

first approach, CG was included as a fixed effect, $y_{ij} = CG_i + u_j + e_{ij}$. We expected that CG would capture the environmental trend. We simulated 2 heritabilities, 0.10 (scenario **FCG10**) and 0.30 (scenario **FCG30**). In the second approach, CG was included as a random effect in the evaluation model, so that CG estimates would be reduced and may not fully capture the environmental trend. This second approach may therefore yield biased evaluations. We tried this approach using different variances of 0.0001 (scenario **RCG0001**), 0.001 (scenario **RCG001**), and 0.01 (scenario **RCG01**). For this second approach, we performed simulations only for a heritability of 0.30.

Data Analysis

For each scenario, 20 replicates were obtained with 10 generations each, and the LR method was applied starting in generation 5. After each generation we ran a BLUP genetic evaluation using blupf90 (Misztal et al., 2002). Thus, for each replicate there are 10 BLUP genetic evaluations. The LR method proceeds by comparing, only for individuals of interest (focal individuals), EBV with little information (partial) at genetic evaluation n and EBV with more information (whole) at genetic evaluation $n + 1$. Individuals of interest were males (approx. 1,800 in each generation), with parent average information during genetic evaluation n, and performance from daughters during genetic evaluation $n + 1$. Then the EBV of these individuals in the partial and whole evaluations are compared. Thus we proceed by comparing EBV across pairs of partial and whole evaluations. These individuals are selected by QMSim based on parent average, which has consequences for the estimated accuracy, as will be discussed later. In this manner, we have 5 comparisons per replicate (5 with 6, 6 with 7, and so on until 9 with 10). We estimated the bias, slope, and accuracies using the formulas shown below, and we compared these with true bias, slope, and accuracies. The true values of bias, slope and accuracy were obtained by comparing the EBV in genetic evaluation n with TBV.

Estimators

The LR method proposes estimators of bias $\left(\hat{\Delta}_p\right)$, slope $\left(\hat{b}_p\right)$, ratios of accuracies $\left(\hat{\rho}_{w,p}\right)$, reliability $\left(\widehat{acc_p^2}\right)$, and ratios of reliabilities $\left(\widehat{\rho^2}_{w,p}\right)$. Accuracies and reliabilities are "selected" ones, in the spirit of Dekkers (1992) and Bijma (2012); in other words, they are lower if the animals of interest are selected. For a deeper description of the statistics, see Legarra and Reverter (2018). All the estimators can be used in multiple trait evaluations as well.

To check the capability of the estimators of bias, slope, and accuracy, we report (a) means and standard deviation of true and estimated values and (b) correlations between true and estimated values. The purpose of reporting the means is to verify whether the LR method is a consistent estimator. For instance, if true slope is 0.9, we want find an average of approximately 0.9, not of 0.7 or 1.1. The purpose of reporting the correlations is to verify the precision of the LR method. For instance, if the true ratio of accuracies is 0.5, we want the estimator to cluster near this value.

Bias. The formula we used for bias was $\hat{\Delta}_p = \overline{\hat{u}_p} - \overline{\hat{u}_w}$, where $\hat{u}_p$ are EBV based on partial data sets and $\hat{u}_w$ are EBV based on whole data sets. This statistic estimates the true bias (Δ_p) between EBV and TBV—that is, $\overline{\hat{u}_p} - \overline{u}$, where u represents TBV. In the absence of true bias, the expected value of $\hat{\Delta}_p$ is zero. A metric of possible interest is the intercept of the regression of $\hat{u}_w$ on $\hat{u}_p$, which is different from $\hat{\Delta}_p$ if $\frac{cov\left(\hat{u}_w, \hat{u}_p\right)}{var\left(\hat{u}_p\right)} \neq 1$ (Mäntysaari et al., 2010). However, we prefer not to consider this metric for our work, first because it does not check the property of BLUP that $E(\hat{u}) = E(u)$, regardless of selection; second because when making selection decisions, as on preselected candidates for selection, it is $\overline{\hat{u}_p}$ and not the intercept that is implicitly used to compare younger versus older generations. In our study, we considered a specific group of animals for which selection proceeds identically, by parent average. In more complex settings (for instance, when the focal group consists of a mixture of animals selected in different ways), it is unclear how selection across several pathways affects differences among average EBV. The standard intercept of the regression may be helpful in such a case, as a perhaps more robust indicator of bias across several groups of individuals selected in heterogeneous manners.

Slope. This is the formula for the slope of the regression of EBV with whole data set (EBV_w) on estimated breeding values with partial data set (EBV_p): $\hat{b}_p = \frac{cov\left(\hat{u}_w, \hat{u}_p\right)}{var\left(\hat{u}_p\right)}$. This is an estimator of the true slope: $b_p = \frac{cov\left(u, \hat{u}_p\right)}{var\left(\hat{u}_p\right)}$. This estimator is related to the dispersion of EBV, and the expected value of $\hat{b}_p$ in the absence of bias is 1. Values less than 1 indicate overdispersion of the EBV.

Ratio of Accuracies. This is the formula for the estimator of the ratio of accuracies:

$\hat{\rho}_{w,p} = \frac{cov\left(\hat{u}_p, \hat{u}_w\right)}{\sqrt{var\left(\hat{u}_p\right) var\left(\hat{u}_w\right)}}$. The expected value of this estimator is $\frac{acc_p}{acc_w}$, where acc_p is the true ("selected") accuracy in the partial data set and acc_w is the true accuracy in the whole data set. Thus, $\frac{1}{\hat{\rho}_{p,w}}$ is the relative increase of accuracy from partial to whole information. For instance, if $\hat{\rho}_{p,w}$ is equal to 0.5, the addition of information doubled the accuracy.

Accuracy of EBV from the Partial Data Set. The formula for accuracy in the partial data set is $\widehat{acc_p^2} = \frac{cov\left(\hat{u}_w, \hat{u}_p\right)}{\sigma^2_{g,i}}$, where $\sigma^2_{g,i}$ is the genetic variance of the individuals of interest. The original Legarra and Reverter (2018) paper suggests the formula $\widehat{acc_p^2} = \frac{cov\left(\hat{u}_w, \hat{u}_p\right)}{\left(1 + \bar{F} - 2\bar{f}\right)\sigma^2_{g,\infty}}$, where $\bar{F}$ is the average inbreeding coefficient, $2\bar{f}$ is the average relationship between individuals, and $\sigma^2_{g,\infty}$ is the genetic variance at equilibrium in a population under selection. However, this formula applies if animals of interest are representative samples of their generation—in other words, they are not yet selected. The formula that we present here is more general. This statistic estimates the "selected" reliability (square of the accuracy) on a partial data set, although it does not estimate model-based accuracy (Dekkers, 1992; Bijma, 2012). We verified that true acc_p^2 agreed with its expected value. The expected value was obtained considering the selection intensities used in the simulation; the model-based accuracies were obtained from the inverse of the Mixed-Model Equations in the BLUP evaluations and using the expressions described in Bijma (2012), as shown in Appendix 2.

To estimate $\sigma^2_{g,i}$ in our case (with true values known from simulation), we simply used

$$\sigma^2_{g,i} = \frac{1}{n}\sum u_j^2 - \left(\frac{1}{n}\sum u_j\right)^2,$$

which already considers the fact that animals may be related (although in our case, they were very little related). In real data sets, $\sigma^2_{g,i}$ can be estimated for any subset of individuals by Gibbs sampling (Sorensen et al., 2001; Lehermeier et al., 2017). If there is no selection, the following formula may be used: $\sigma^2_{g,i} = \left(1 + \bar{F} - 2\bar{f}\right)\sigma^2_{g,\infty} = \left(1 + \bar{F} - 2\bar{f}\right)\sigma^2_g$, as no Bulmer effect occurs, only drift. Thus, this estimator is of easy use for unselected individuals or traits.

Ratio of Reliabilities. We used the following formula to calculate the ratio of reliabilities: $\widehat{\rho^2_{p,w}} = \frac{cov\left(\hat{u}_p, \hat{u}_w\right)}{var\left(\hat{u}_w\right)}$. This is a measure of the inverse increase in ("selected") reliabilities from partial to whole, with an expected value $E\left(\widehat{\rho^2_{p,w}}\right) = \frac{acc_p^2}{acc_w^2}$.

RESULTS

Scenario 1: Correct Genetic Model

Figure 3 shows, across all replicates, true and estimated biases. Because the model used in the genetic evaluation was the same as that used to simulate the data, no bias is expected. Nevertheless, a small true bias was generated due to chance. For the 2 heritabilities, the estimator was able to indicate the true value of bias: $corr\left(\Delta_p, \widehat{\Delta_\mathrm{p}}\right) = 0.59$ for T10 (Table 2 and Figure 3, left-hand panel). The best estimation was in the higher-heritability scenario: $corr\left(\Delta_p, \widehat{\Delta_\mathrm{p}}\right) = 0.61$ for T30 (Table 2 and Figure 3, right-hand panel). In Figure 3, points of the same color belong to the same replicate, and it is clear that they do not cluster together. In other words, comparisons within replicates can be seen as independent.

Similar results were observed (Figure 4 and Table 2) for the estimator of the slope of EBV: $corr\left(b_p, \widehat{b_p}\right) = 0.45$ for T10 and $corr\left(b_p, \widehat{b_p}\right) = 0.59$ for T30. Thus, the true slope was more precisely estimated when heritability was high ($h^2 = 0.30$).

Figure 5 shows $\hat{\rho}_{w,p}$ and $\widehat{acc_p^2}$, the estimator of the accuracy gain from partial to whole data sets and the estimator of reliability for partial data, versus true values from the simulations: $\frac{acc_p}{acc_w}$ and acc_p^2, respectively. We found good agreement between estimators and true values; for instance, for scenario T10, $corr\left(\hat{\rho}_{w,p}, \frac{acc_p}{acc_w}\right) = 0.54$ and $corr\left(\widehat{acc_p^2}, acc_p^2\right) = 0.45$. For scenario T30, we found $corr\left(\hat{\rho}_{w,p}, \frac{acc_p}{acc_w}\right) = 0.62$ and $corr\left(\widehat{acc_p^2}, acc_p^2\right) = 0.53$. We also verified that values of true accuracy, acc_p^2, and estimated accuracy, $\widehat{acc_p}^2$, agree with expectations based on model-based accuracies and selection decisions and intensities (see Appendix 2). In particular, the low mean values of acc_p^2, 0.022 and 0.033, are due to preselection on males based on parent aver-

Table 2. Mean, SD, and correlation between estimated $\left(\hat{\Delta}_p\right)$ and true bias (Δ_p) and estimated $\left(\hat{b}_p\right)$ and true slope (b_p) when the h^2 used in the evaluation model was the correct one

Estimator	Scenario[1]	Estimated value (SD)	True value (SD)	Correlation estimated—true
$\hat{\Delta}_p$	T10	−0.001 (0.005)	−0.001 (0.010)	0.59
	T30	$-6.55e^{-05}$ (0.008)	$-5.76e^{-04}$ (0.014)	0.61
$\hat{b}_p$	T10	0.996 (0.067)	1.009 (0.167)	0.45
	T30	1.006 (0.069)	0.992 (0.141)	0.59

[1]Scenario T10: $h^2 = 0.10$; scenario T30: $h^2 = 0.30$.

age, whereas model-based (or unselected) reliabilities are 0.16 and 0.25.

The estimator $\widehat{\rho^2_{p,w}}$ behaved similarly to $\hat{\rho}_{w,p}$. For example, $corr\left(\hat{\rho}_{w,p}, \sqrt{\widehat{\rho^2_{p,w}}}\right) = 0.91$ for both heritabilities, 0.10 and 0.30.

Scenario 2: Incorrect Heritability in Evaluation Model

When we used the wrong h^2 in the model for evaluation, the largest differences could be seen in the estimation of bias (Figure 6 and Table 3). The use of an incorrect heritability generates a strong true bias. Similarly to the detection of bias, the estimator was able to indicate the bias in the correct direction, but the magnitude was underestimated. For instance, the real bias of scenario W05 is approximately 0.10, but the estimated bias is approximately 0.05. These differences are more pronounced for lower h^2.

In the case of the estimation of slope, Table 3 and Figure 7 show that the use of incorrectly high heritability results in true values of slope b_p less than 1, as indicated by Reverter et al. (1994a), with the effect more important for the scenario with a simulated heritability of 0.10 (mean b_p of 0.83 in scenario W15 and 0.97 in scenario W35). In addition, it is possible to observe that there is no important difference among means of the estimators of slopes across heritabilities, but differences do exist with respect to the variation of the estimators, with the estimators of W05 and W15 being more variable than those of W25 and W35. Nevertheless, in all scenarios the slope could be estimated, albeit with low precision (Figure 7 and Table 3): $corr\left(b_p, \hat{b}_p\right)$ for scenario W05 = 0.53, W15 = 0.44, W25 = 0.46, and W35 = 0.46. We observe that for scenario W05, true b_p was close to 1, whereas it should be higher; we have no explanation for this. Table 4 shows the results of the estimations of accuracies. In general, it is possible to estimate both the ratio of accuracies $\left(\frac{acc_p}{acc_w}\right)$ and the squared accuracy $\left(acc_p^2\right)$. Note that the values of the squared accuracies in $\widehat{acc_p^2}$ are very small,

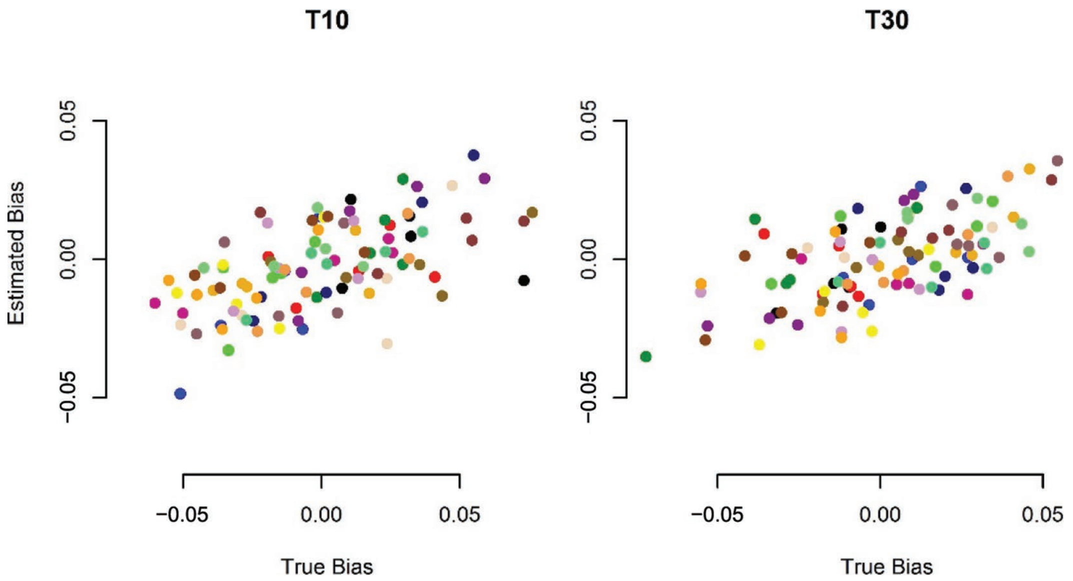

Figure 3. Estimated versus true bias, simulation scenarios T10 ($h^2 = 0.10$) and T30 ($h^2 = 0.30$). Different colors are used for each replicate.

because these animals have very little information when selected as candidates for selection: a phenotyped dam and possibly a few phenotyped half-sibs.

It is possible to observe a particular behavior in scenario W05. For instance, this scenario estimates incorrect values of $\hat{\rho}_{w,p}$ and of $\widehat{\rho^2_{p,w}}$. A possible explanation could be the use of excessively low heritability, where sires' EBV have a very small contribution from daughters' phenotypes, and the EBV in successive genetic evaluations tend to strongly resemble parent average EBV.

Scenario 3: Not Fitting Environmental Trend

When we used CG as a fixed effect, because the CG are large enough, they correctly capture the effect of the environmental trend, and there is almost no bias in the evaluations, only relatively small biases due to chance (approx. 0.05 genetic standard deviation). Figure 8 shows that this bias cannot be very well estimated: $corr\left(\Delta_p, \hat{\Delta}_p\right)$is 0.46 for scenario FCG30 and 0.41 for scenario FCG10. Additionally, its estimated magnitude is too small. The estimator of the slope (Figure 9 and Table 5), whose direction is well estimated—$corr\left(b_p, \hat{b}_p\right)$ equal to 0.52 for FCG10 and 0.60 for FCG30—but whose magnitude is underestimated. Accuracies are in general well estimated (Table 6).

When CG are used as random effect, at each generation the true bias increases, because the genetic trend captures the environmental trend (Figure 10). It is possible to observe that the confusion decreases as the variance used for the CG increases and the CG estimates are less reduced, but in no case is it possible to estimate the true bias. Regarding the remaining , $\hat{b}_p$ performed more poorly when CG were fit as random effects than when CG were used as fixed effects: $corr\left(b_p, \hat{b}_p\right)$ were 0.43, 0.45, and 0.49 for RCG0001, RCG001, and RCG01, respectively. Meanwhile, the estimators of accuracies presented similar values to those of the fixed CG scenarios but with less correlation between estimator and estimated (Table 6).

DISCUSSION

Several reports have showed some concern about the bias of the genomic predictions of young bulls with genomic predictions (Spelman et al., 2010; Sargolzaei et al., 2012; Mikshowsky, 2018). Using different methodologies, several studies have detected bias (Liu et al., 2016; Mikshowsky et al., 2017). In addition, bias is a problem that continues to motivate studies of dairy sheep. In Pyrenees dairy sheep breed selection schemes, some bias was found, ranging from 4.92 (Basco-Béarnaise) to 16.98 L of milk (Manech Tête Rousse) with pedigree evaluations and slopes of 0.44 (Manech Tête Noire) to 0.95 (Basco-Béarnaise; Legarra et al., 2014). This demonstrates that bias may be present in the genetic evaluations of some dairy sheep breeds. However, these studies relied on the use of precorrected data, and we were interested in the possibility of using official genetic evaluations to quantify biases and accuracies.

Studies searching for methods to analyze bias in genetic evaluations are not new. In 1994 Reverter et al. (1994b) presented 3 statistics related to dispersion,

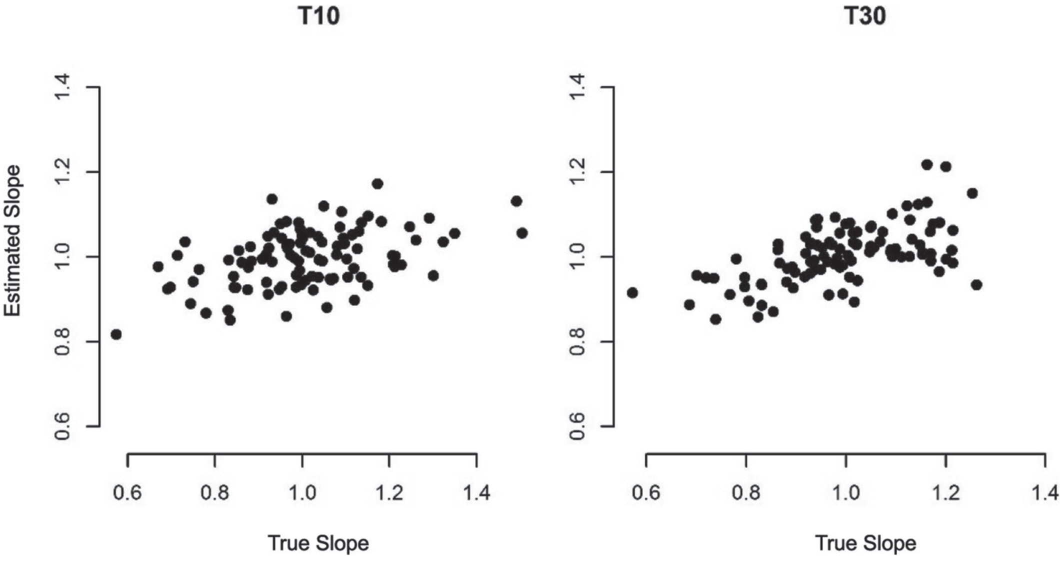

Figure 4. Estimated versus true slope, simulation scenarios T10 ($h^2 = 0.10$) and T30 ($h^2 = 0.30$).

accuracy, and genetic gain, obtained from subsets of EBV of successive evaluations. The following year, Boichard et al. (1995) presented 3 methods to check bias in genetic evaluations; for the first 2 methods, work with raw data is needed, but the last method is based on statistics obtained from EBV obtained from different data sets. Following the same principles, Mäntysaari et al. (2010) developed the Interbull validation test for genomic evaluations, using GEBV from a reduced data set and DYD from a full data set. However, this requires access to the raw data sets, and DYD are not always computable or reliable, as we have seen among sheep and swine. In addition, for traits that have been genomically preselected, the estimated genetic trends and DYD using pedigree information only are possibly biased (Sullivan, 2018). Yet these pedigree evaluations pass the Interbull test, although they may not pass the Mendelian sampling variance test (Sullivan, 2018; Tyrisevä et al., 2018). Because the LR method does not use DYD, it should not be affected by biased DYD.

Comparing successive EBV is advantageous because there is no need to access the full data, and also because the procedure is very simple to execute. This is why comparing EBV was proposed by Reverter et al. (1994b) and Boichard et al. (1995). The genetic interpretation of this comparison, according to Thompson (2001), is, “Informally this statistic is asking the question does the recent data change the prediction of early animals. In a sense this is looking backwards.” The LR method is an extension of the ideas of Reverter et al. (1994b). Using standard BLUP theory, Legarra and Reverter (2018) showed that, by comparing old and new EBV, it is possible to infer biases and also accuracies at the population level. However, the behavior of this method in practice is unknown. In particular, the LR method assumes that the model for genetic evaluation is perfect. In this work, we used simulation to verify that the LR method is robust to departures from the true model (generally speaking), which is very advantageous because analytical models are always compromises that do not perfectly reflect the state of nature.

One of our results is the correlation between true and estimated value, as of the estimated accuracy. This number reflects the ability to estimate, in a data

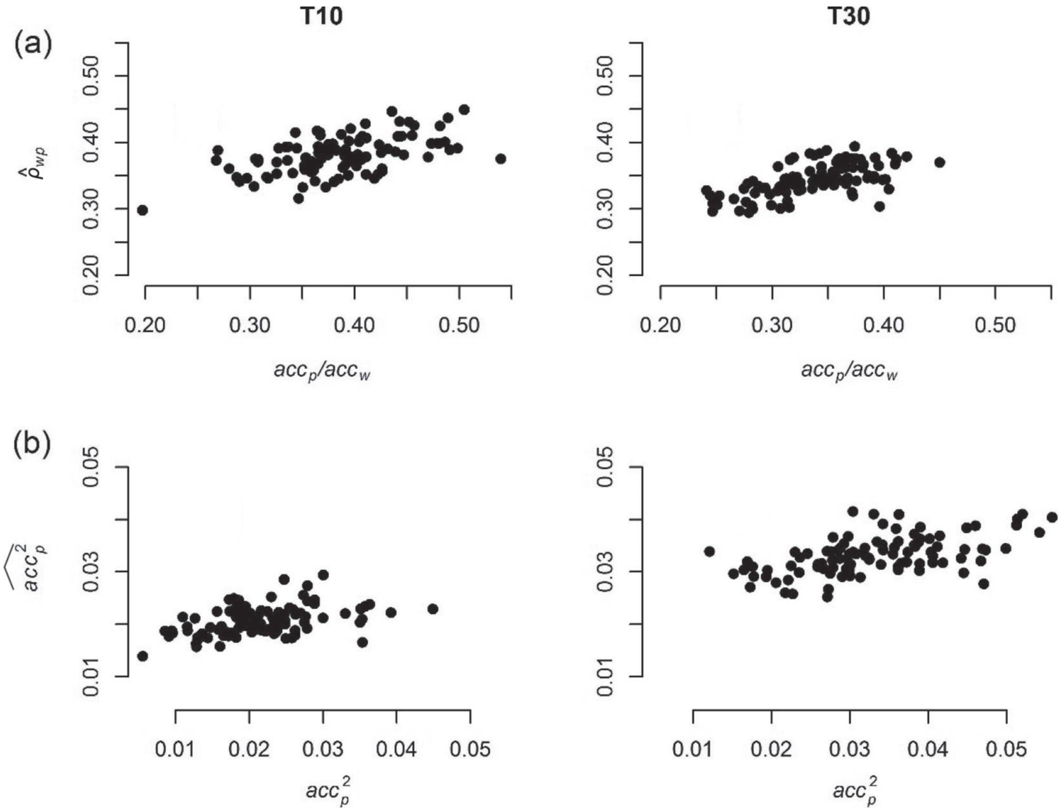

Figure 5. Estimations of accuracies, simulation scenarios T10 ($h^2 = 0.10$) and T30 ($h^2 = 0.30$). (a) Estimations of the inverse of relative gain in accuracy from partial to whole data sets $\left(\hat{\rho}_{w,p}\right)$, versus the ratio of the accuracy with partial data set to the accuracy with whole data set $\left(\frac{acc_p}{acc_w}\right)$. (b) Estimations of reliability on partial data set $\left(\widehat{acc_p^2}\right)$ versus true reliability on partial data set $\left(acc_p^2\right)$.

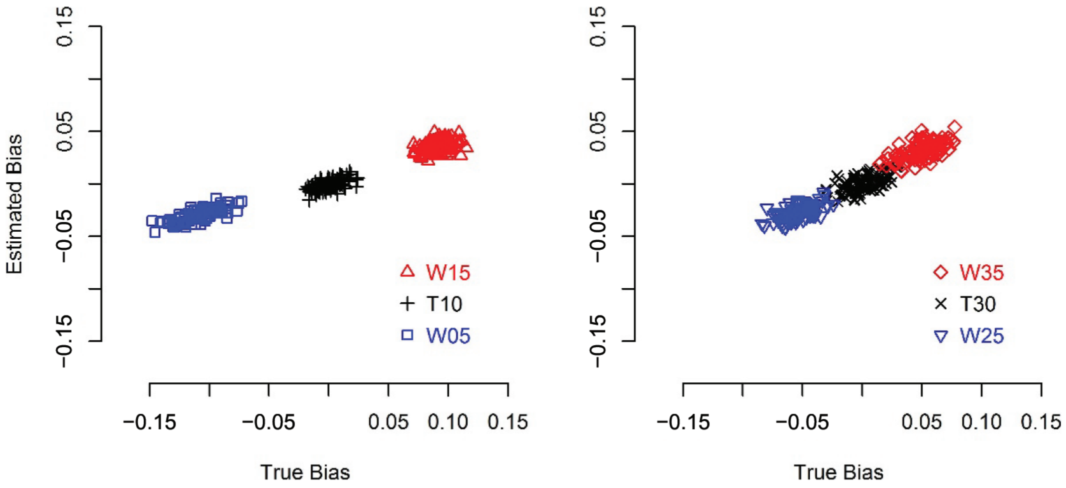

Figure 6. Estimated versus true bias when the evaluation model used incorrect heritability: simulation performed with $h^2 = 0.10$, evaluation model used $h^2 = 0.05$ (W05) or $h^2 = 0.15$ (W15); simulation performed with $h^2 = 0.30$, evaluation model used $h^2 = 0.25$ (W25) or $h^2 = 0.35$ (W35). Simulation scenarios T10 and T30 (when heritability used in the evaluation model was correct, $h^2 = 0.10$ or 0.30, respectively) were included for comparison.

set, the parameter of interest using the LR method, but the variation of the true parameter of interest is generally small, and therefore the correlation is not a good guide. In addition, the correlation between the estimator and the true value is not available for a single study with real data. A confidence interval around the estimated value would be more useful. For this, Legarra and Reverter (2018) suggested bootstrap. This deserves investigation.

When the model is wrong, clear indications might or might not be present. For instance, Table 3 points out that heritabilities fit in the model appear to be incorrect, and the model may be changed accordingly. However, the LR method cannot "see" (e.g., Figure 10) that the model for random CG is biased.

In several cases, we observed that the bias was correctly estimated in direction but not correctly estimated in magnitude: for example, when the wrong heritability was used in the evaluation model. This is because if estimated EBV are too greatly or too little regressed (as due to an incorrect model), the statistics used are, therefore, scaled, but the sign does not change. In our case, the difference between true and used heritability was not very large, which results in signals of bias that are not very strong (see Table 3). Still, method LR in this scenario generally pointed out that problems existed in the evaluation.

However, when an environmental trend was simulated and CG was used as a random effect (a very incorrect model of evaluation) the EBV captured an important part of the environmental trend, and consequently estimation of bias through the LR method became impossible. When the model for genetic evaluation was robust, no bias occurred, and the LR method reported correct results. Globally, these 2 scenarios (incorrect heritability and environmental trend) show that the LR

Table 3. Mean, SD, and correlation between estimated $\left(\hat{\Delta}_p\right)$ and true bias (Δ_p) and between estimated $\left(\hat{b}_p\right)$ and true (b_p) slope when the h^2 used in the evaluation model was incorrect

Estimator	Scenario[1]	Estimated value (SD)	True value (SD)	Correlation estimated—true
$\hat{\Delta}_p$	W05	−0.030 (0.006)	−0.111 (0.015)	0.77
	W15	0.035 (0.005)	0.091 (0.010)	0.36
	W25	−0.027 (0.007)	−0.054 (0.012)	0.55
	W35	0.032 (0.009)	0.050 (0.014)	0.63
$\hat{b}_p$	W05	1.091 (0.077)	0.976 (0.235)	0.54
	W15	0.931 (0.083)	0.826 (0.135)	0.44
	W25	1.026 (0.065)	1.059 (0.138)	0.46
	W35	0.980 (0.071)	0.969 (0.109)	0.46

[1] Scenario W05: true $h^2 = 0.10$, used $h^2 = 0.05$; W15: true $h^2 = 0.10$, used $h^2 = 0.15$; W25: true $h^2 = 0.30$, used $h^2 = 0.25$; W35: true $h^2 = 0.30$, used $h^2 = 0.35$.

Table 4. Mean, SD, and correlation between estimated $\left(\hat{\rho}_{w,p}, \widehat{acc_p^2}, \text{ and } \widehat{\rho_{p,w}^2}\right)$ and true values of accuracies $\left(\frac{acc_p}{acc_w}, acc_p^2, \text{ and } \frac{acc_p^2}{acc_w^2}, \text{ respectively}\right)$, when h^2 used in the evaluation model was incorrect; values for scenarios T10 and T30 (when h^2 used in the evaluation model was correct) are included for comparison[1]

Estimator	Scenario[2]	Estimated value (SD)	True value (SD)	Correlation estimated—true
$\hat{\rho}_{w,p}$	T10	0.381 (0.028)	0.385 (0.059)	0.54
	W05	0.587 (0.043)	0.366 (0.074)	0.41
	W15	0.305 (0.028)	0.360 (0.057)	0.43
	T30	0.344 (0.024)	0.336 (0.045)	0.62
	W25	0.371 (0.027)	0.340 (0.043)	0.50
	W35	0.319 (0.022)	0.349 (0.036)	0.45
$\widehat{acc_p^2}$	T10	0.021 (0.003)	0.022 (0.007)	0.45
	W05	0.020 (0.004)	0.018 (0.008)	0.32
	W15	0.025 (0.003)	0.018 (0.006)	0.48
	T30	0.033 (0.004)	0.033 (0.009)	0.53
	W25	0.030 (0.004)	0.033 (0.009)	0.45
	W35	0.036 (0.003)	0.035 (0.008)	0.44
$\widehat{\rho_{p,w}^2}$	T10	0.146 (0.016)	0.152 (0.046)	0.50
	W05	0.319 (0.051)	0.139 (0.055)	0.28
	W15	0.100 (0.011)	0.133 (0.042)	0.40
	T30	0.118 (0.011)	0.115 (0.030)	0.57
	W25	0.135 (0.014)	0.118 (0.030)	0.43
	W35	0.104 (0.008)	0.123 (0.025)	0.48

[1] $\hat{\rho}_{w,p}$ = estimator of the ratio of accuracies; $\widehat{acc_p^2}$ = estimator of the accuracy of EBV in partial data set; $\widehat{\rho_{p,w}^2}$ = estimator of the ratio of reliabilities; $\frac{acc_p}{acc_w}$ = ratio of accuracies; acc_p^2 = accuracy of EBV in partial data set; and $\frac{acc_p^2}{acc_w^2}$ = ratio of reliabilities.

[2] Scenario T10: h^2 = 0.10; scenario T30: h^2 = 0.30; scenario W05: true h^2 = 0.10, used h^2 = 0.05; W15: true h^2 = 0.10, used h^2 = 0.15; W25: true h^2 = 0.30, used h^2 = 0.25; W35: true h^2 = 0.30, used h^2 = 0.35.

method works reasonably well for detection of biases when the model is robust or close to the true one, and that it works well for estimation of accuracy even when the model is not good. This is because accuracies are correlations that are invariant to shift and scaling.

The most obvious use of statistics on bias is model selection. We suggest that a good model is one that is empirically (i.e., using the LR method or a similar one) unbiased (both in bias and slope) and that gives accurate predictions. For instance, it seems reasonable to

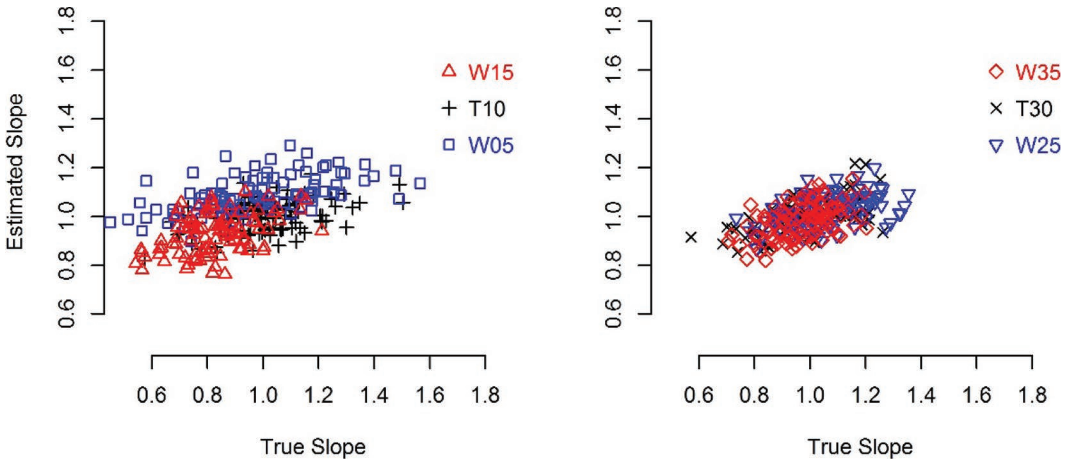

Figure 7. Estimated versus true slope when the evaluation model used an incorrect heritability: simulation performed with h^2 = 0.10, evaluation model used h^2 = 0.05 (W05) or h^2 = 0.15 (W15); simulation performed with h^2 = 0.30, evaluation model used h^2 = 0.25 (W25) or h^2 = 0.35 (W35). Simulation scenarios T10 and T30 (when heritability used in the evaluation model was correct, h^2 = 0.10 or 0.30, respectively) were included for comparison.

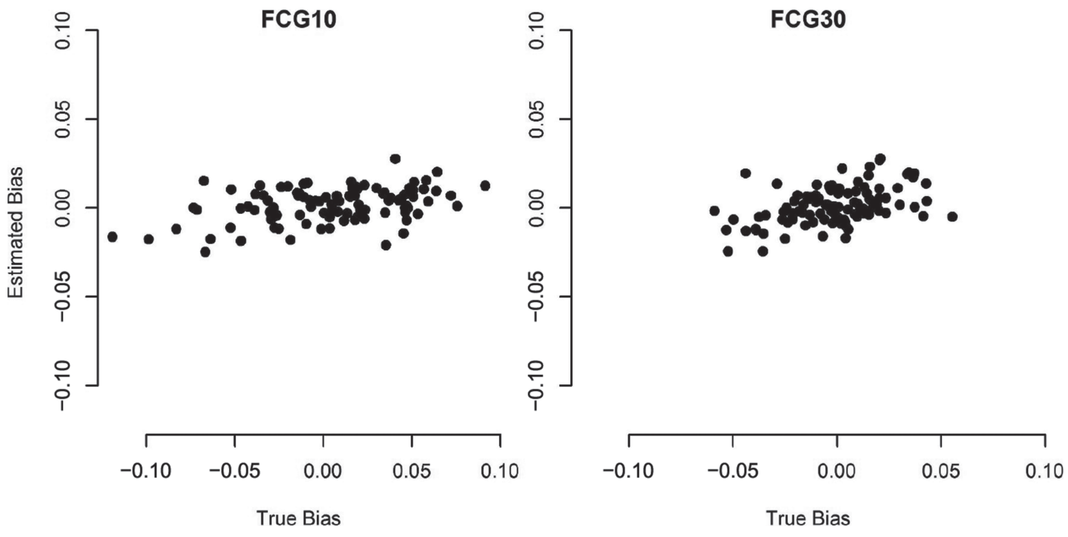

Figure 8. Estimated versus true bias when an environment trend effect was simulated: scenarios FCG10 ($h^2 = 0.10$) and FCG30 ($h^2 = 0.30$).

choose, between 2 competing heritabilities, the one that would give less bias, as on Δ_p. However, this seems to work only for minor changes in the model, given that Δ_p is not estimable if the model is too far from reality or not robust, as in the environmental trend and random CG scenario. Also, the theory only works within the model; that is, the results of checking $\hat{u}_p$ of model 1 against $\hat{u}_w$ of model 2 do not have theoretical support. Still, a model that is more coherent (empirically unbiased from run to run) always seems more attractive than one with erratic behavior, in which biases are observed.

We presented 3 estimators related to accuracies, 2 of them being ratios of accuracies $\hat{\rho}_{w,p}$ and $\widehat{\rho^2}_{w,p}$, which try to indicate the changes in accuracies due to the increment of information. Because they are ratios of the accuracy and the reliability, they should be equivalent (they are expectations of the same true values), but as the results show, they are not. One of the reasons is that expectations do not yield true values, so 2 expectations constructed differently may give different values. Another, more relevant, reason for the difference is that $\widehat{\rho^2}_{w,p}$ is influenced by the dispersion of EBV in the

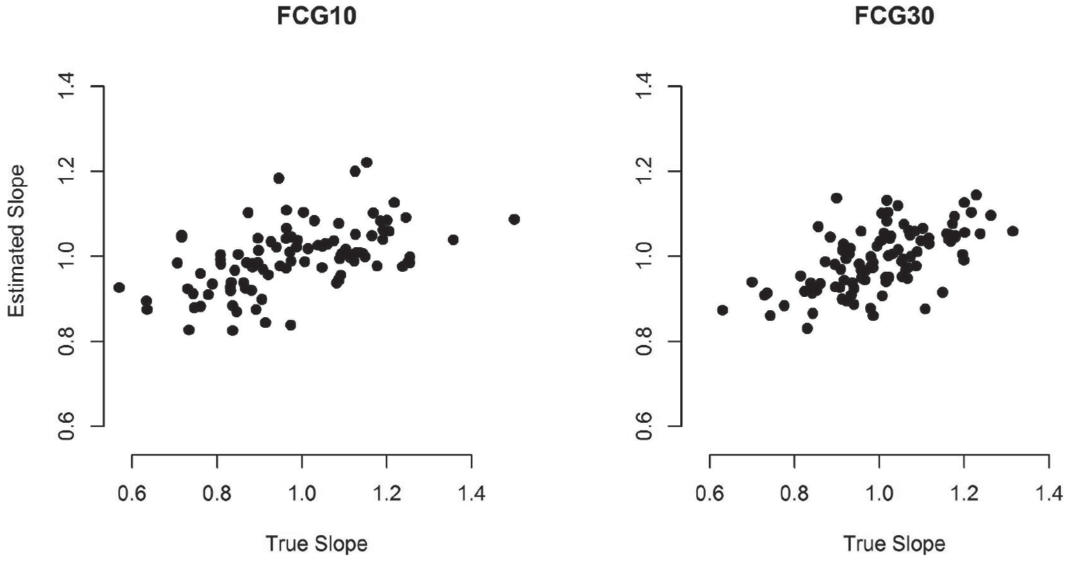

Figure 9. Estimated versus true slope when an environment trend effect was simulated and contemporary group (CG) is used as fixed effect in the model: scenarios FCG10 ($h^2 = 0.10$) and FCG30 ($h^2 = 0.30$).

Table 5. Mean, SD, and correlation between estimated $\left(\hat{\Delta}_p\right)$ and true bias (Δ_p) and between estimated $\left(\hat{b}_p\right)$ and true (b_p) slope when an environmental effect was simulated

Estimator	Scenario[1]	Estimated value (SD)	True value (SD)	Correlation estimated—true
$\hat{\Delta}_p$	FCG10	$4.34e^{-04}$ (0.003)	0.001 (0.013)	0.41
	FCG30	0.001 (0.006)	−0.001 (0.013)	0.46
	RCG0001	−0.121 (0.008)	0.404 (0.121)	0.13
	RCG001	−0.074 (0.012)	0.189 (0.075)	−0.78
	RCG01	−0.013 (0.006)	0.030 (0.022)	−0.08
$\hat{b}_p$	FCG10	0.995 (0.076)	0.984 (0.173)	0.52
	FCG30	0.993 (0.072)	1.003 (0.133)	0.60
	RCG0001	1.01 (0.056)	0.877 (0.112)	0.43
	RCG001	1.01 (0.064)	0.936 (0.122)	0.45
	RCG01	1.01 (0.064)	0.974 (0.137)	0.49

[1] Scenario FCG10: $h^2 = 0.10$; FCG30: $h^2 = 0.30$; RCG0001, RCG001, and RCG01: $h^2 = 0.30$, and variance of contemporary groups = 0.0001, 0.001, and 0.01, respectively.

partial and whole data sets, whereas $\hat{\rho}_{w,p}$ is not (Legarra and Reverter, 2018), so if the slope is not equal to 1, the estimators will differ. In that sense, $\hat{\rho}_{w,p}$ is robust to slopes not being 1.

All accuracies and reliabilities in this study are "selected" ones, meaning that they refer to a selected set of individuals. Therefore, they are affected by selection and much lower than model-based accuracies and reliabilities, as shown in Appendix 2. Biases and slopes may both be affected by selection. For instance, if $b_p < 1$ (inflation of EBV), prediction is unbiased, considering averages of all animals in the first generation. However, selected animals will be overdispersed, and their estimated mean will be lower than the true mean. If selected animals are used for the LR method, then $\hat{\Delta}_p$ will be different from zero, showing that BLUP is not biased for this group of animals, which is the property of interest for breeders.

The ultimate aim of the LR method, and that of this study, is to reliably detect systematic biases in genetic evaluations that, if ignored, would hamper genetic progress—as the overdispersion of EBV results in

Table 6. Mean, SD, and correlation between estimated $\left(\hat{\rho}_{w,p}, \widehat{acc_p^2}, \text{ and } \widehat{\rho_{p,w}^2}\right)$ and true values of accuracies $\left(\frac{acc_p}{acc_w}, acc_p^2, \text{ and } \frac{acc_p^2}{acc_w^2}, \text{ respectively}\right)$ when an environmental effect was simulated[1]

Estimator	Scenario[2]	Estimated value (SD)	True value (SD)	Correlation estimated—true
$\hat{\rho}_{w,p}$	FCG10	0.377 (0.031)	0.374 (0.061)	0.53
	FCG30	0.337 (0.024)	0.338 (0.042)	0.59
	RCG0001	0.382 (0.023)	0.340 (0.040)	0.54
	RCG001	0.364 (0.022)	0.340 (0.043)	0.48
	RCG01	0.344 (0.023)	0.333 (0.046)	0.54
$\widehat{acc_p^2}$	FCG10	0.020 (0.003)	0.020 (0.007)	0.39
	FCG30	0.032 (0.003)	0.033 (0.009)	0.58
	RCG0001	0.042 (0.004)	0.033 (0.008)	0.40
	RCG001	0.037 (0.004)	0.033 (0.009)	0.44
	RCG01	0.033 (0.003)	0.032 (0.009)	0.50
$\widehat{\rho_{p,w}^2}$	FCG10	0.143 (0.016)	0.144 (0.046)	0.42
	FCG30	0.114 (0.011)	0.116 (0.028)	0.57
	RCG0001	0.144 (0.012)	0.117 (0.027)	0.50
	RCG001	0.131 (0.010)	0.117 (0.029)	0.40
	RCG01	0.118 (0.011)	0.113 (0.030)	0.51

[1] $\hat{\rho}_{w,p}$ = estimator of the ratio of accuracies; $\widehat{acc_p^2}$ = estimator of the accuracy of EBV in partial data set; $\widehat{\rho_{p,w}^2}$ = estimator of the ratio of reliabilities; $\frac{acc_p}{acc_w}$ = ratio of accuracies; acc_p^2 = accuracy of EBV in partial data set; $\frac{acc_p^2}{acc_w^2}$ = ratio of reliabilities.

[2] Scenario FCG10: $h^2 = 0.10$; FCG30: $h^2 = 0.30$; RCG0001, RCG001, and RCG01: $h^2 = 0.30$, and variance of contemporary groups = 0.0001, 0.001, and 0.01, respectively.

choosing too many young animals and leads to slower genetic progress. Overestimating genetic progress for a trait may result in changes to selection objectives. This problem is not merely theoretical; for instance, Powell and Wiggans (1994) describe a bias in the US national evaluation that generated overprediction of breeding values of US bulls in France (Bonaiti and Barbat, 1993).

Efron (2004) showed that parametric and nonparametric (cross-validation) prediction error estimates are related, and, when the model used for genetic evaluations is believable, estimation of error using parametric methods is more precise than the results of a nonparametric method. Therefore, as an ancillary property, the LR method can assist finding a believable model from which statistics of interest (biases and accuracies) can be obtained parametrically.

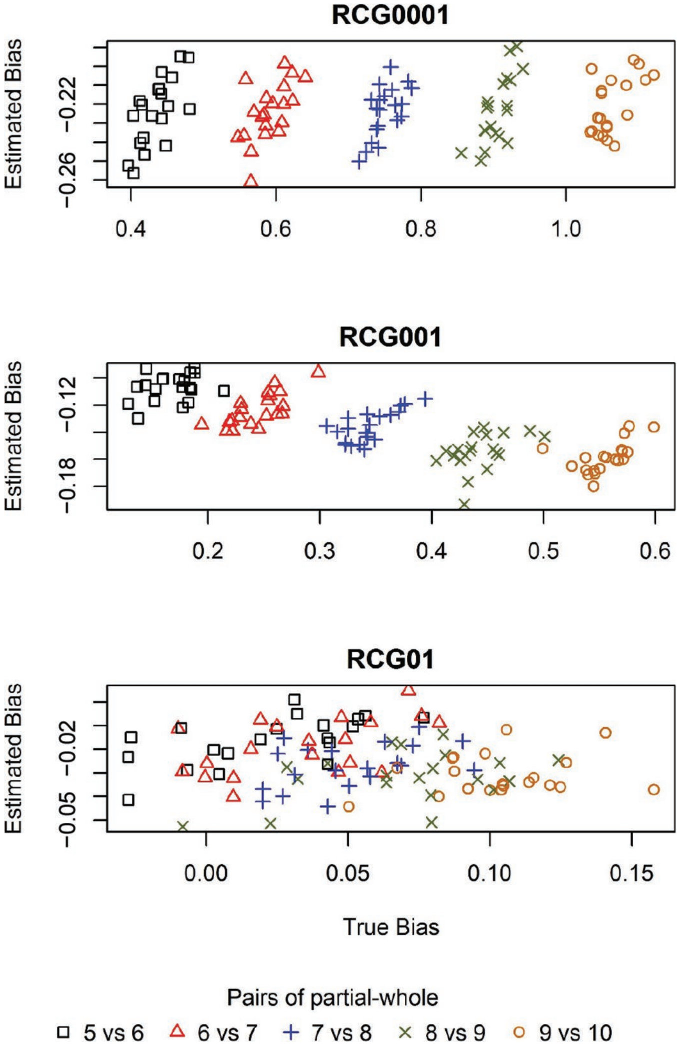

Figure 10. Estimated versus true bias when an environment trend effect was simulated and contemporary group (CG) is used as random effect in the model (RCG0001, RCG001, and RCG01 represent variances of 0.0001, 0.001, and 0.01, respectively). Different colors are used for different pairs of comparisons between partial and whole data sets.

CONCLUSIONS

The LR method is capable of estimating bias and accuracies if the model is reasonably correct or robust, and its estimates of bias and accuracies improve as information increases (that is, when the heritability of the trait is high). For incorrect genetic models—in our case, if the heritability used in genetic evaluations was wrong, or if there were hidden trends in the data such as an environmental trend—it is still possible to estimate bias if the model is robust. The direction of the bias will be correctly pointed out but not its magnitude. However, if the model is seriously mis-specified (in our work, such that environmental trend could not be accommodated), the LR method cannot estimate the bias. However, the estimators of slope and accuracies generally performed well for all scenarios. Further research is warranted, using the LR method with real data.

ACKNOWLEDGMENTS

The authors thank projects ARDI funded by INTERREG POCTEFA (Programa INTERREG V-A España-Francia-Andorra, Jaca, Spain), Metaprogram SELGEN of INRA, the Department of Animal Genetics of INRA and La Région Occitanie, which financed this research. In addition, we are grateful to the Genotoul Bioinformatics Platform Toulouse Midi-Pyrenees (Bioinfo Genotoul; Toulouse, France) for providing computing and storage resources. This work has also received funding from the European Union Horizon 2020 Research and Innovation program under grant agreement N°772787, SMARTER. We further thank Vincent Ducrocq (INRA, Jouy-en-Josas, France), Alain Charcosset (INRA, Gif-sur-Yvette, France), Catherine Larzul (INRA, Toulouse, France), Anne Ricard (Institut Français du Cheval et de l'Equitation, Toulouse), and Leopoldo Sanchez-Rodriguez (INRA, Orleans, France), and the two anonymous reviewers for their helpful advice.

REFERENCES

Astruc, J. M., G. Baloche, F. Barillet, and A. Legarra. 2014. Genomic evaluation validation test proposed by Interbull is necessary but not sufficient because it does not check the correct genetic trend. Page 50 in Proc. 39th ICAR Annual Conference. International Committee for Animal Recording, Berlin, Germany.

Baloche, G., A. Legarra, G. Sallé, H. Larroque, J.-M. Astruc, C. Robert-Granié, and F. Barillet. 2014. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. J. Dairy Sci. 97:1107–1116. https://doi.org/10.3168/jds.2013-7135.

Barillet, F., G. Lagriffoul, P. Marnet, H. Larroque, R. Ruup, D. Portes, F. Bocquier, and J. M. Astruc. 2016. Objectifs de sélection et stratégie raisonnée de mise en œuvre à l'échelle des populations de brebis laitières françaises. INRA Prod. Anim. 29:19–40.

Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. J. Anim. Breed. Genet. 129:345–358. https://doi.org/10.1111/j.1439-0388.2012.00991.x.

Boichard, D., B. Bonaiti, A. Barbat, and S. Mattalia. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. J. Dairy Sci. 78:431–437. https://doi.org/10.3168/jds.S0022-0302(95)76652-8.

Bonaiti, B., and D. B. A. Barbat. 1993. Problems arising with genetic trend estimation in dairy cattle. Interbull Bull. 8:1–8.

Dekkers, J. C. M. 1992. Asymptotic response to selection on best linear unbiased predictors of breeding values. Anim. Prod. 54:351–360. https://doi.org/10.1017/S0003356100020808.

Efron, B. 2004. The estimation of prediction error: Covariance penalties and cross-validation. J. Am. Stat. Assoc. 99:619–632. https://doi.org/10.2307/27590436.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447. https://doi.org/10.2307/2529430.

Henderson, C. R. 1984. Applications of Linear Models in Animal Breeding. University of Guelph, Guelph, Canada.

Legarra, A., G. Baloche, F. Barillet, J. M. Astruc, C. Soulas, X. Aguerre, F. Arrese, L. Mintegi, M. Lasarte, F. Maeztu, I. Beltrán de Heredia, and E. Ugarte. 2014. Within- and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J. Dairy Sci. 97:3200–3212. https://doi.org/10.3168/jds.2013-7745.

Legarra, A., and A. Reverter. 2017. Can we frame and understand cross-validation results in animal breeding? Proc. Assoc. Advmt. Anim. Breed. Genet. 22:73–80. https://doi.org/10.14800/ics.95.

Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genet. Sel. Evol. 50:53. https://doi.org/10.1186/s12711-018-0426-6.

Lehermeier, C., G. de los Campos, V. Wimmer, and C. C. Schön. 2017. Genomic variance estimates: With or without disequilibrium covariances? J. Anim. Breed. Genet. 134:232–241. https://doi.org/10.1111/jbg.12268.

Liu, Z., H. Alkhoder, F. Reinhardt, and R. Reents. 2016. Accuracy and bias of genomic prediction for second-generation candidates. Interbull Bull. 50:24–28.

Mäntysaari, E., Z. Liu, and P. Vanraden. 2010. Interbull validation test for Genomic evaluations. Page 17 in Proc. Interbull International Workshop—Genomic Information in Genetic Evaluations. Paris, France. Mar. 4 to 5, 2010. Interbull Bull. 41:4–5.

Mikshowsky, A. 2018. Can you really trust dairy genomics? The Bullvine. Accessed Mar. 11, 2019. http://www.thebullvine.com/news/can-you-really-trust-dairy-genomics/.

Mikshowsky, A. A., D. Gianola, and K. A. Weigel. 2017. Assessing genomic prediction accuracy for Holstein sires using bootstrap aggregation sampling and leave-one-out cross validation. J. Dairy Sci. 100:453–464. https://doi.org/10.3168/jds.2016-11496.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. BLUPF90 and related programs (BGF90). Page 21 in Proc. 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France. Aug. 19 to 23, 2002. INRA, Paris, France.

Patry, C., and V. Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J. Dairy Sci. 94:1011–1020. https://doi.org/10.3168/jds.2010-3804.

Powell, R. L., and R. Wiggans. 1994. Impact of changes in U.S. evaluations on conversions and comparisons. Proc. annual meeting of the International Bull Evaluation Service. Ottawa, Ontario, Canada. Aug. 6, 1994. Interbull Bull. 10:1–2.

Reverter, A., B. L. Golden, R. M. Bourdon, and J. S. Brinks. 1994a. Method R variance components procedure: Application on the simple breeding value model. J. Anim. Sci. 72:2247–2253. https://doi.org/10.2527/1994.7292247x.

Reverter, A., B. L. Golden, R. M. Bourdon, and J. S. Brinks. 1994b. Technical note: Detection of bias in genetic predictions. J. Anim. Sci. 72:34–37. https://doi.org/10.2527/1994.72134x.

Robertson, A. 1977. The effect of selection on the estimation of genetic parameters. Z. Tierzuecht. Zuechtungsbiol. 94:131–135. https://doi.org/10.1111/j.1439-0388.1977.tb01542.x.

Sargolzaei, M., J. Chesnais, and F. Schenkel. 2012. Assessing the bias in top GPA bulls. Canadian Dairy Network. Accessed Mar. 11, 2019. https://www.cdn.ca/Articles/GEBOCT2012/Jacques_Bias_in_Top_Bulls.pdf.

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: A large-scale genome simulator for livestock. Bioinformatics 25:680–681. https://doi.org/10.1093/bioinformatics/btp045.

Sonesson, A. K., and T. H. Meuwissen. 2000. Mating schemes for optimum contribution selection with constrained rates of inbreeding. Genet. Sel. Evol. 32:231–248.

Sorensen, D., R. Fernando, and D. Gianola. 2001. Inferring the trajectory of genetic variance in the course of artificial selection. Genet. Res. 77:83–94. https://doi.org/10.1017/s0016672300004845.

Spelman, R. J., J. Arias, M. D. Keehan, V. Obolonkin, A. M. Winkelman, D. L. Johnson, and B. L. Harris. 2010. Application of Genomic Selection in the New Zealand Dairy Cattle Industry. Page 311 in Proc. 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany. Aug. 1 to 6, 2010. Zwonull Media, Leipzig, Germany.

Sullivan, P. G. 2018. Mendelian sampling variance tests with genomic preselection. Proc. 2018 Interbull Technical Workshop. Dubrovink, Croatia. Aug. 25 to 26, 2018. Interbull Bulletin 54. Accessed Oct. 30, 2019. https://journal.interbull.org/index.php/ib/article/view/1463/1525.

Thompson, R. 2001. Statistical validation of genetic models. Livest. Prod. Sci. 72:129–134.

Tyrisevä, A.-M., W. F. Fikse, E. A. Mäntysaari, J. Jakobsen, G. P. Aamand, J. Dürr, and M. H. Lidauer. 2018. Validation of consistency of Mendelian sampling variance. J. Dairy Sci. 101:2187–2198. https://doi.org/10.3168/jds.2017-13255.

VanRaden, P. M., and J. R. O. O'Connell. 2018. Validating genomic reliabilities and gains from phenotypic updates. Page 22 in Proc. 2018 Interbull Meeting. Auckland, New Zealand. Feb. 10 to 21, 2018. Interbull Bulletin. 53:22–26. Accessed Oct. 30, 2019. https://journal.interbull.org/index.php/ib/article/view/1442/1506.

ORCIDS

F. L. Macedo https://orcid.org/0000-0002-1949-9214

A. Legarra https://orcid.org/0000-0001-8893-7620

APPENDIXES

Appendix 1. Example of QMSim Parameter File for Heritability of 0.10

```
/*******************************
** Global parameters **
*******************************/
seed = "./seed.prv";
nthread = 1;
nrep = 20; //Number of replicates
h2 = 0.10; //Heritability
qtlh2 = 0.10; //QTL heritability
```

```
phvar = 1.0; //Phenotypic variance
no_male_rec; //Males have no record
/*******************************
** Historical population **
*******************************/
begin_hp;
hg_size = 100 [0] 100 [100] 100000 [110];
nmlhg = 5000; //Number of males in the last
generation
end_hp;
/*******************************
** Populations **
*******************************/
begin_pop = "p1";
begin_founder;
male [n = 4500, pop = "hp"];
female [n = 45000, pop = "hp"];
end_founder;
ls = 1; //Litter size
pmp = 0.5; //Proportion of male progeny
ng = 10; //Number of generations
md = minf; //Mating design
sr = 0.4; //Replacement ratio for sires
dr = 0.2; //Replacement ratio for dams
sd = ebv /h; //Selection design
cd = ebv /l; //Culling design
ebv_est = external_bv "Sol.sh";
begin_popoutput;
data;
stat;
genotype /snp_code /gen 0 1 2 3 4 5 6 7 8 9
10;
end_popoutput;
end_pop;
/*******************************
** Genome **
*******************************/
begin_genome;
begin_chr = 30;
chrlen = 100; //Chromosome length cm
nmloci = 1500; //Number of markers
mpos = rnd; //Marker positions
nma = all 2; //Number of marker alleles
maf = eql; //Marker allele frequencies
nqloci = 333; //Number of QTL was 10000
qpos = rnd; //QTL positions
nqa = all 2; //Number of QTL alleles
qaf = eql; //QTL allele frequencies
qae = rndg 0.4; //QTL allele effects
end_chr;
mmutr = 2.5e-5 /recurrent; //Marker mutation
rate
qmutr = 0.01 /recurrent; //QTL mutation rate
r_mpos_g; // Randomize marker positions
across genome
r_qpos_g; // Randomize QTL positions across
genome
end_genome;
/*******************************
** Output options **
*******************************/
begin_output;
hp_stat;
monitor_hp_homo /freq 1;
allele_effect;
end_output;
```

Appendix 2. Agreement of Selected Accuracies Computed Using the LR Method and Expected Accuracies from BLUP

Henderson (1975) proved (implicit in the paper and not explicitly shown) that for selection assuming $\boldsymbol{L'y}$ and $\boldsymbol{L'X}=0$, the distribution of variances and covariances is as follows:

$$Var\begin{pmatrix}\boldsymbol{u}\\ \hat{\boldsymbol{u}}\end{pmatrix}=\begin{pmatrix}\boldsymbol{G}-\boldsymbol{B_u}\boldsymbol{H}_0\boldsymbol{B_u}' & \boldsymbol{G}-\boldsymbol{C}_{22}-\boldsymbol{B_u}\boldsymbol{H}_0\boldsymbol{B_u}'\\ \boldsymbol{G}-\boldsymbol{C}_{22}-\boldsymbol{B_u}\boldsymbol{H}_0\boldsymbol{B_u}' & \boldsymbol{G}-\boldsymbol{C}_{22}-\boldsymbol{B_u}\boldsymbol{H}_0\boldsymbol{B_u}'\end{pmatrix},$$

where $\boldsymbol{B_u}$ represents the selection process, $\boldsymbol{H}_0$ represents the decrease in variance under selection, and $\boldsymbol{C}_{22}$ represents the corresponding block of the inverse of the coefficient matrix for animal equations.

In other words, $Var(\boldsymbol{u})=\boldsymbol{G}^*=\ \boldsymbol{G}-\boldsymbol{B_u}\boldsymbol{H}_0\boldsymbol{B_u}'$ describes the reduction in genetic variance due to selection, and then,

$$Var\begin{pmatrix}\boldsymbol{u}\\ \hat{\boldsymbol{u}}\end{pmatrix}=\begin{pmatrix}\boldsymbol{G}^* & \boldsymbol{G}^*-\boldsymbol{C}_{22}\\ \boldsymbol{G}^*-\boldsymbol{C}_{22} & \boldsymbol{G}^*-\boldsymbol{C}_{22}\end{pmatrix},$$

similar to Legarra and Reverter (2018) but with $\boldsymbol{G}$ substituted by $\boldsymbol{G}^*$. For a set of little-related, homogeneous individuals of interest, $\boldsymbol{G}^*\approx\boldsymbol{G}\frac{\sigma^2_{g,i}}{\sigma^2_g}$. The LR method estimates, using $\widehat{acc_p^2}$, selected accuracies (called r^{*2} in Dekkers, 1992), which are $r^{*2}=1-\frac{PEV}{\sigma^2_{g,i}}$. Thus, the way to compare with model-based reliabilities $r^2=1-\frac{PEV}{\sigma^2_g}$ [for instance, from the inverse of the mixed-model equations (MME)] is, considering selection intensities of candidates for selection, to check whether r^{*2} agrees with r^2.

Below, we calculate the expected value of equilibrium parent average reliability [$r^{*2} = \rho^2_{PA,\infty}$ in Bijma (2012)], which is what $\widehat{acc^2_p}$ tries to estimate, given model-based reliabilities and selection intensities. We follow Equation 10 of Bijma (2012) to calculate the parent average reliability at equilibrium with different selection in both sexes:

$$r^{*2} = \rho^2_{PA,\infty} = \frac{1}{2}\,\frac{\rho^2_{m,SC,\infty}\left(1-k_m\right)+\rho^2_{f,SC,\infty}\left(1-k_f\right)}{2},$$

where $\rho^2_{m,SC,\infty}$ and $\rho^2_{f,SC,\infty}$ are the equilibrium reliabilities of the selection criterion for each sex (m = males and f = females, parents of the focal individuals) and k_m and k_f are the proportional reductions in variance for males (m) and females (f) (Robertson, 1977).

The terms $\rho^2_{m,SC,\infty}$ and $\rho^2_{f,SC,\infty}$ could be calculated from

$$\rho^2_{m,SC,\infty} = \rho^2_{m,SC,0}\,\frac{1+\dfrac{1}{2}k_f\left(1-\dfrac{\rho^2_{f,SC,0}}{\rho^2_{m,SC,0}}\right)}{1+\overline{k\left(1-\rho^2_{SC,0}\right)}}$$

and

$$\rho^2_{f,SC,\infty} = \rho^2_{f,SC,0}\,\frac{1+\dfrac{1}{2}k_m\left(1-\dfrac{\rho^2_{m,SC,0}}{\rho^2_{f,SC,0}}\right)}{1+\overline{k\left(1-\rho^2_{SC,0}\right)}},$$

where

$$\overline{k\left(1-\rho^2_{SC,0}\right)} = \frac{k_m\left(1-\rho^2_{m,SC,0}\right)+k_f\left(1-\rho^2_{f,SC,0}\right)}{2},$$

and $\rho^2_{m,SC,0}$ and $\rho^2_{f,SC,0}$ are the unselected reliabilities [r^2 in Dekkers (1992)] of selection criterion of males and females, ignoring selection—or, in other words, the model-based reliability derived from the inverse of the MME.

Application to Simulated Data

In the scenario with a correct genetic model, for both heritabilities, we calculated $\rho^2_{PA,\infty}$ of the focal individuals of generation 7 (males born in generation 7 and used as sires in next generations) from the first replicate, taking the average model-based reliability (from BLUP) of his sires and dams as $\rho^2_{m,SC,0}$ and $\rho^2_{f,SC,0}$. In both cases, the proportion of selected was of 0.08 for males and 0.45 for females, so $k_m = 0.84$ and $k_f = 0.65$.

Case of h^2 = 0.10 (T10)

For the lower heritability, we obtained values of $\rho^2_{m,SC,0} = 0.37$ and $\rho^2_{f,SC,0} = 0.26$. Then, following the equations, $\rho^2_{m,SC,\infty} = 0.27$, $\rho^2_{f,SC,\infty} = 0.14$, and finally, $\rho^2_{PA,\infty} = 0.023$, which represents the equilibrium parent average (PA) reliability for EBV on the partial data set and is the expected value of r^{*2} (true value) and of $\widehat{acc^2_p}$ (estimator), agreeing very well with both (Table 4). In addition, we obtained from the inverse of the MME the model-based (or unselected) reliability EBV using the partial data set ($\hat{u}_p$). The mean reliability obtained from the MME was 0.16, which compares to the equilibrium PA reliability of 0.023. We can see an important deviation from $\rho^2_{PA,\infty}$, respecting the reliability obtained from BLUP evaluation, but this is because they express 2 different reliabilities.

Case of h^2 = 0.30 (T30)

Given $\rho^2_{m,SC,0} = 0.57$ and $\rho^2_{f,SC,0} = 0.49$ from BLUP evaluations, we calculated $\rho^2_{m,SC,\infty} = 0.44$, $\rho^2_{f,SC,\infty} = 0.33$, and $\rho^2_{PA,\infty} = 0.046$. Our result for this case was $acc^2_p = \widehat{acc^2_p} = 0.033$ (Table 4), a value lower than but reasonably close to $\rho^2_{PA,\infty}$. The reason for the difference is perhaps that the reality of selection is not well described by the expressions above. The mean of model-based reliabilities from BLUP was 0.25.

It is necessary to highlight that here we showed examples taking focal males from the seventh generation and only 1 replicate for each heritability. The values of estimations presented in results are the mean across all the replicates, including 5 pairs of partial-whole data sets within each replicate.