The accuracy and bias of single-step genomic prediction for populations under selection

Wan-Ling Hsu*, Dorian J. Garrick*†, and Rohan L. Fernando*†

*Department of Animal Science, Iowa State University, 50011 Ames, Iowa, USA, †Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand

ABSTRACT In single-step analyses, missing genotypes are explicitly or implicitly imputed, and this requires centering the observed genotypes, using the means of the unselected founders. If genotypes are only available on selected individuals, centering on the unselected founder mean is not straightforward. Here, computer simulation is used to study an alternative analysis that does not require centering genotypes but fits the mean $\mu_g$ of unselected individuals as a fixed effect. Starting with observed diplotypes from 721 cattle, a 5 generation population was simulated with sire selection to produce 40,000 individuals with phenotypes of which the 1,000 sires had genotypes. The next generation of 8,000 genotyped individuals was used for validation. Evaluations were undertaken: with (J) or without (N) $\mu_g$ when marker covariates were not centered; and with (JC) or without (C) $\mu_g$ when all observed and imputed marker covariates were centered. Centering did not influence accuracy of genomic prediction, but fitting $\mu_g$ did. Accuracies were improved when the panel comprised only QTL, models JC and J had accuracies of 99.4%; whereas models C and N had accuracies of 90.2%. When only markers were in the panel, the 4 models had accuracies of 80.4%. In panels that included QTL, fitting $\mu_g$ in the model improved accuracy, but had little impact when the panel contained only markers. In populations undergoing selection, fitting $\mu_g$ in the model is recommended to avoid bias and reduction in prediction accuracy due to selection.

KEYWORDS centering genotype covariates; estimated breeding value; genomic prediction; selection; single-step

In pedigree based analyses, the expected value of breeding values is zero. In order to achieve similar properties in whole-genome analyses, marker genotype covariates are often transformed. When all individuals are genotyped, it has been shown that inference on genotype effects does not depend on how the covariates are transformed (Strandén and Christensen 2010). However, when data includes genotyped and non-genotyped individuals, inference on marker effects from single-step analyses may depend on how the covariates are transformed (Fernando et al. 2014). In single-step analyses using marker effects models, the breeding values of non-genotyped individuals are partitioned into components representing the prediction of non-genotyped individuals conditional on their genotyped relatives and an independent deviation (Fernando et al. 2014). The prediction of non-genotyped individuals conditional on their genotyped relatives is done based on best linear prediction, which requires the first moments to be known without error. This is straightforward if the mean of the genomic breeding value is zero in the absence of selection. Centering the observed genotype covariates using what their means would be in the absence of selection would result in genomic breeding values with null means. However, such genotype covariate means are typically unavailable. Fernando et al. (Fernando et al. 2014) proposed a solution for the marker effects model that involves fitting an additional fixed covariate that estimates the mean $\mu_g$ of the linear component of the genotypic value, which is denoted by $a_i$ in (1) below, in a population where selection is absent. Using that approach, even when there is selection, the selection process can be ignored, because the analysis is conditional on the data used for selection (Goffinet 1983; Gianola and Fernando 1986; Im et al. 1989; Sorensen et al. 2001). In Markov chain Monte Carlo analyses, centering results in better mixing (Strandén and Christensen 2010), reducing the number of iterations required to obtain converged genomic predictions. In practice, centering the entire matrix of genotype covariates, including the observed
and imputed genotypes, using their sample means can be done in addition to the Fernando et al. (Fernando et al. 2014) approach of fitting \( \mu_g \). This type of centering of the entire genotype matrix does not affect inference on marker effects.

The same issue with centering of the observed genotype covariates that we described above for the marker effects model is also implicit for the single-step breeding value model (single step GBLUP), and a similar solution was proposed by Vitezica et al. (Vitezica et al. 2011). In their proposed solution, the observed genotype covariates are centered using their means, and in addition the genomic covariate matrix is corrected for the change in the mean breeding value of the genotyped individuals (Vitezica et al. 2011). It was shown in that paper that this is equivalent to fitting the change in breeding value due to selection as a random effect.

Here we use simulated data to compare the accuracy and bias in genomic prediction applied to populations under selection with and without centering the entire matrix of genotype covariates, and with and without fitting \( \mu_g \) as a fixed effect. Further, we will show that when the observed genotype covariates are centered using means calculated from selected individuals rather than means from all individuals, the meaning of \( \mu_g \) changes from the mean of unselected individuals to become the mean breeding value in selected individuals as claimed by Vitezica et al. (Vitezica et al. 2011).

### Materials and Methods

#### Theory

To simplify the presentation of the genetic model, without loss of generality, we will assume that the unconditional expectation of the phenotypic value for all individuals is the same. Let \( \mathbf{m}_i \) denote the row vector of genotypes for individual \( i \), which are often coded as -1,0,1. Then, under additive gene action, the genotypic value, \( g_i \), which is the expected phenotypic value of an individual with genotypes \( \mathbf{m}_i \) can be written as

\[
    g_i = \beta + \mathbf{m}_i' \alpha \tag{1}
\]

where \( \beta \) is the value of \( g_i \) when \( \mathbf{m}_i = 0' \) and \( \alpha \) is the vector of substitution effects. Recognize the scalar \( \beta \) and the vector \( \alpha \) are constants, but \( g_i \) will be a random variable because of randomness in \( \alpha \). Note that the expected value of the linear component \( \alpha \) of the genotypic value in (1) is \( E(\alpha) = E(\mathbf{m}_i' \alpha) = \mathbf{k}' \alpha \), where \( \mathbf{k}' = E(\mathbf{m}_i) \), which may not be equal to zero. Thus, it is customary to write the model for the genotypic value, as can be derived from (1), as follows:

\[
    g_i = (\beta + \mu_g) + a_i - \mu_g = (\beta + \mu_g) + u_i = (\beta + \mu_g) + (\mathbf{m}_i' - \mathbf{k}') \alpha, \tag{2}
\]

where \((\beta + \mu_g)\) is a constant, representing the expected phenotypic value of an individual with genotype \( \mathbf{m}_i = 0' \), and \( u_i = (\mathbf{m}_i' - \mathbf{k}') \alpha \) is a random variable that has null expectation, which is the breeding value predicted in a pedigree-based BLUP evaluation. When genotypes are observed and used in a genomic analysis, they may be transformed or coded by subtracting their expectations, \( \mathbf{k}' \), from the observed values, \( \mathbf{m}_i' \). In both (1) and (2), \( a_i \) is the same substitution effect for locus \( j \). The intercepts in these models, however, are different. In (1), the intercept is \( \beta \) and in the value of \( g_i \) when \( \mathbf{m}_i = 0' \). In (2), on the other hand, the intercept is \((\beta + \mu_g)\), and it is the value of \( g_i \) when \( \mathbf{m}_i' = \mathbf{k}' \).

More generally, \( \mathbf{k}' \) is not known, so genotypes are coded by subtracting a different vector \( \mathbf{v}' \) from the observed genotypes as \( \mathbf{m}_i' - \mathbf{v}' \). Still, \( a_i \) is the substitution effect for locus \( j \), but the intercept will change to become \((\beta + \mathbf{v}' \alpha)\), which is the value of \( g_i \) when \( \mathbf{m}_i' = \mathbf{v}' \). Thus, as more rigorously shown in (Stranden and Christensen 2010), inference about \( \alpha \) does not depend on how the genotypes are coded. A simpler but rigorous proof is given in the appendix of this paper.

In single-step analyses, where some individuals are not genotyped, the missing genotypes are imputed either implicitly (Legarra et al. 2009) or explicitly (Fernando et al. 2014) using best linear prediction. Let \( \mathbf{M}_g \) denote the matrix of genotypes for individuals that were genotyped. Then, the genotypes of the individuals with missing genotypes are imputed as

\[
    \mathbf{M}_n = \mathbf{1} \mathbf{k}' + \mathbf{A}_{ng} \mathbf{A}_{gg}^{-1} (\mathbf{M}_g - \mathbf{1} \mathbf{k}'), \tag{4}
\]

where \( \mathbf{A}_{gg} \) is the matrix of the pedigree-based additive relationships between the non-genotyped and genotyped individual and \( \mathbf{A}_{gg} \) is the matrix of additive relationships among genotyped individuals. Now, the model for the genotypic values, when genotypes are coded as in (2), becomes

\[
    g_n = 1 (\beta + \mu_g) + \mathbf{A}_{ng} \mathbf{A}_{gg}^{-1} (\mathbf{M}_g - \mathbf{1} \mathbf{k}') \alpha + \epsilon \tag{5}
\]

where \( \epsilon \) is that part of \( g_n \) that cannot be imputed from knowledge of the breeding values of genotyped relatives. In practice, the true value of \( \mathbf{k}' \) is not known, and data for its estimation may not be available. Rearranging these equations in terms of the uncentered \( \mathbf{M}_g \) rather than the centered matrix of genotype covariates \( (\mathbf{M}_g - \mathbf{1} \mathbf{k}') \), results in

\[
    g_n = 1 (\beta + \mu_g) - \mathbf{A}_{ng} \mathbf{A}_{gg}^{-1} \mathbf{1} \mathbf{k}' \alpha + \mathbf{A}_{ng} \mathbf{A}_{gg}^{-1} \mathbf{M}_g \alpha + \epsilon \tag{6}
\]

and substituting \( \mu_g = \mathbf{k}' \alpha \), as previously defined, and letting \( \mathbf{M}_n = \mathbf{A}_{ng} \mathbf{A}_{gg}^{-1} \mathbf{M}_g \), results in

\[
    g_n = 1 (\beta + \mu_g) - \mathbf{A}_{ng} \mathbf{A}_{gg}^{-1} \mathbf{1} \mathbf{k}' \mu_g + \mathbf{M}_n \alpha \tag{7}
\]

where \( \epsilon = \mathbf{g}_n - \mathbf{g}_g \) is the value of \( g_n \). This type of centering of the entire genotype matrix in the last term of (4) is required because \( g_n \) does not appear in the model for genotyped individuals.
When the vector $\mathbf{a}$ represents the substitution effects of a large number of loci containing positive and negative effects, $\mu_g = \mathbf{K}a$ will tend to have a value close to zero. Accordingly, we have simulated some scenarios with positive $\mu_g = E(a_i)$ so that the entire $\mathbf{a}$ vector is positive to exacerbate the impact of $\mu_g = \mathbf{K}a$. Nevertheless, when marker rather than causal alleles are fitted in the model, the sign of the substitution effects depends on the phase relationship between marker and causal allele, which may equally likely be positive or negative.

Even if $\mu_g = E(a_i) = 0$, in a population undergoing selection, it is expected that $E(a_i) = E(m_i) \mathbf{a} \neq 0$ in non-founders. Suppose $\nu^\prime$ is the mean of the observed genotype covariates in such a population undergoing selection and these means are used to center the matrix $\mathbf{M}_g$ of observed genotypes. Then, the model for the genotypic values can be written in terms of the matrix $\mathbf{M}_g^\prime = \mathbf{M}_g - 1\nu^\prime$ of centered covariates as

$$\begin{align*}
\mathbf{g}_n &= \mathbf{1}(\beta^* + \mu_g^*) - \mathbf{A}_{ng} \mathbf{A}_{gg}^{-1} \mathbf{1} \mu_g^* + \mathbf{A}_{ng} \mathbf{A}_{gg}^{-1} \mathbf{M}_g^\prime \mathbf{a} + \epsilon \\
\mathbf{g}_g &= \mathbf{1}(\beta^* + \mu_g^*) - 1 \mu_g^* + \mathbf{M}_g \mathbf{a},
\end{align*}$$

and using $\mathbf{J}$ for the covariate corresponding to $\mu_g^*$, (5) can be written as

$$\begin{align*}
\mathbf{g}_n &= \mathbf{1}(\beta^* + \mu_g^*) + \mathbf{J} \mu_g^* + \mathbf{A}_{ng} \mathbf{A}_{gg}^{-1} (\mathbf{M}_g - 1\nu^\prime) \mathbf{a} + \epsilon \\
\mathbf{g}_g &= \mathbf{1}(\beta^* + \mu_g^*) + \mathbf{J} \mu_g^* + \mathbf{M}_g \mathbf{a} + \mathbf{J} \nu^\prime \mathbf{a} + \epsilon \\
\end{align*}$$

which can be combined as

$$\mathbf{g} = \mathbf{1}(\beta^* + \mu_g^*) + \mathbf{J}(\mu_g^* + \nu^\prime \mathbf{a}) + \mathbf{M} \mathbf{a} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \epsilon. \quad (6)$$

Note that the regression coefficients for $\mathbf{J}$, $\mu_g^*$ in (4) and $\mu_g^* + \nu^\prime \mathbf{a}$ in (6) must be equal. This implies that $\mu_g^* = \mu_g - \nu^\prime \mathbf{a}$. Similarly, the intercepts of these two models must be equal too, and this implies $\beta^* = \beta + \nu^\prime \mathbf{a}$.

### Simulations

Phenotypic and genotypic data were simulated, using XSIM (Cheng et al. 2015), based on haplotypes from ten genomic regions of 721 US Hereford beef cattle that were genotyped with the Illumina 770K BovineHD BeadChip and reported in terms of the number of copies of the A allele at each locus. The selected regions came from choosing every 10th single nucleotide polymorphism (SNP), starting at SNP 5,001, to get a low density panel of 200 SNPs from each of chromosomes 1 to 10 (BTA1-BTA10), after eliminating SNP with MAF $< 0.01$. These 2,000 SNPs represent ten 0.1M chromosomes. Average LD between adjacent SNPs was 0.300. Fifty SNPs from each chromosome were randomly chosen to represent QTL. The QTL effects were sampled from a Normal distribution with mean $\mu_g = 0.2$ and multiplied by the number of copies of the A allele to produce the true breeding value (TBV). The TBV were added to a Normally distributed residual term scaled by the sample variance of the TBV to simulate a trait with a heritability of 0.5. The first 20 SNPs from each of the ten chromosomes were also used to simulate a smaller panel, with 5 QTL and 15 markers per chromosome. Average LD between adjacent SNPs was 0.289. TBV were simulated in the same way as for the 2,000 SNP scenario, then scaled to simulate traits with heritabilities ($h^2$) of 0.1, 0.3 or 0.5. An additional scenario with $\mu_a = 0$ was used to simulate TBV for a trait with heritability 0.5.

Half the observed diplotypes from US Hereford cattle were assigned to represent males and the remainder to represent females. Those 360 males and 361 females were sampled in pairs, with replacement, to produce 4,000 male and 4,000 female offspring representing generation G-4. There were no mutations. Four more non-overlapping generations of random mating were carried out with one male and one female offspring per dam to randomly chosen sires to produce the founder population (G0).

The G1 generation was produced by mass phenotypic selection of the top 200 G0 males, and this was repeated for 5 more generations. Each female was randomly mated twice to selected males to produce 1 offspring of each sex each generation. Across non-overlapping generation G0 to G5, a total of 48,000 individuals with phenotypes, genotypes and TBV were simulated for each scenario.

The training data included phenotypes from all individuals in G0 to G4 ($n = 40,000$), and genotypes from all 1,000 sires and all 8,000 G5 animals. Fixed loci, if any, were filtered from the panel before genomic prediction analyses. The genetic and residual variances used in genomic prediction were the sample variance of the TBV in G0 and the corresponding residual variance used to define the desired heritability in the founder population.

### Models

Five statistical models were compared for differences in accuracy and bias of prediction. These include models with or without $\mu_g$ and with or without centering of observed and imputed marker covariates, and a model that used pedigree relationships but not marker covariates.

**1. Mixed Linear Model:**

Accuracy of pedigree-based best linear unbiased prediction (PBLUP) was quantified using the correlation of TBV and estimated breeding values (EBV), where TBV was as simulated and EBV were obtained by fitting the mixed linear model (Henderson 1973, 1984):

$$y = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \epsilon,$$

where $y$ is a vector of phenotypic observations, $\mathbf{1}$ is a vector of 1s, $\mu = \beta + \mu_g$, is a general mean, $\mathbf{u}$ is a vector of random additive genetic effects, $\epsilon$ is a vector of random residual effects, and $\mathbf{Z}$ is a known incidence matrix relating observations to $\mathbf{u}$. In this model, $E(\mathbf{u}) = 0$, $E(\epsilon) = 0$, so that $E(y) = \mathbf{1}\mu$. Further, $\text{Var}(\mathbf{u}) = \mathbf{A}\sigma^2_g$, for $\mathbf{A}$ being the numerator relationship matrix, and $\text{Var}(\epsilon) = \mathbf{I}\sigma^2_e$, so that $\text{Var}(y) = \mathbf{Z}\mathbf{A}\sigma^2_g + \mathbf{I}\sigma^2_e$.

**2. Single-Step Bayesian Regression Model:**

All genomic EBVs (GEBV) were obtained by Single-Step Bayesian regression (Fernando et al. 2014), using BayesC priors for marker effects with $\pi = 0$, which gives predictions that are identical to those from single-step GBLUP. The model was implemented in Julia (http://julialang.org) based on the SSBP package (http://QTL.rocks) to construct an MCMC chain of 50,000 samples. Individuals were separated into 2 groups designated with subscripts $g$ or $n$ according to whether or not simulated genotypes were assumed to be observed or missing. The single-step bayesian regression model including a covariate $\mathbf{J}$ for $\mu_g$ (Model J) was:
were centered, respectively (Table 1).

were centered using means of the observed and imputed genotypes of the non-genotyped and genotyped individuals, \( y_{\mu_g} \) and \( y_{J} \), respectively.

This model was used for four variants of the single-step Bayesian regression analysis depending on whether or not the covariate \( J \) corresponding to the mean \( \mu_g \) was in the model, and whether or not the columns in the marker covariate matrix \( M \) were centered using means of the observed and imputed genotype covariates. The analyses with \( J \) or without \( J \) are denoted as \( J \) or \( N \) when covariates were not centered, and as JC or C, when the entire matrix of imputed and observed genotype covariates were centered, respectively (Table 1).

Accuracy of genomic prediction was quantified using the correlation of TBV and GEBV \((r_{g,g})\), where GEBV were obtained from each of the 4 analyses described above. Bias of genomic prediction was quantified using the deviation from unity of the coefficient of regression of TBV on GEBV \((b_{g,g})\). In models JC and J, GEBV are obtained using equation (24) in (Fernando et al. 2014):
These genotypes and the methodology described previously are sufficient to reproduce the simulations used in this study.

Results and Discussion

Effect of fitting a genotypic mean and centering of observed and imputed marker covariates

1. Accuracy:

The accuracies of genomic prediction as assessed by validation in G5 after training using G0-G4 for a trait with 50 QTL whose effects were sampled from a Normal distribution with \( \mu_\beta = 0.2 \) and \( h^2 = 0.5 \) are in Table 2. The accuracy of PBLUP in predicting breeding values for individuals without phenotypes was 41.5%, accounting for less than 20% of genetic variance. All analyses using genotypes resulted in more accurate predictions than using pedigree alone. Centering the entire matrix of imputed and observed genotype covariates had no affect on the accuracy of the genomic analyses regardless of the nature of the marker panel. This is because, as shown in the Appendix, this type of centering is equivalent to adding and subtracting \( 1\mathbf{m}'\mathbf{a} \) from the model equation and this has no effect on the mixed model solutions for \( \mathbf{a} \).

In this study, selection resulted in successive advance in mean TBV from G0 to G5 being 10.14, 10.63, 11.18, 11.68, 12.15 and 12.57. The mean genotypic value was not zero in G0 because with only markers. This would not be the case in a population were not selected and thus TBV from G0 to G5 being 10.14, 10.63, 11.18, 11.68, 12.15 and 12.57. The mean genotypic value was not zero in G0 because QTL genotypes were not centered and the mean QTL effect was \( \mu_\beta = 0.2 \). Since the QTL effects do not change by selection the advance in TBV reflects changes in the frequencies of the favorable alleles of the QTL. So centering using the allele frequency means of the genotyped sires in G1-G4 and all individuals in G5 does not closely approximate the centering that would have occurred if the allele frequency means had been obtained from the unselected population. In contrast fitting \( \mu_\beta \) in the model estimates the relevant mean from the data.

In panels that included causal variants (QTL), fitting \( \mu_\beta \) in the model substantially improved the accuracy to being near perfect. This is not surprising given there were only 50 QTL, the heritability was 0.5 and there were 40,000 phenotyped ancestors, including 200 genotyped sires per generation in the training. However, in the panel that contained only markers with no causal variants, fitting \( \mu_\beta \) in the model had little impact. This is because in the population that was simulated here the founders were not selected and thus \( \mu_\beta \) was close to zero for the panel with only markers. This would not be the case in a population where even the founders are descendants of selected parents. Thus, for traits that have been subject to selection, fitting \( \mathbf{J} \) in the model is expected to improve accuracy.

Using one replicate as an example, for the panel including both QTL and markers, the estimate of \( \mu \) was about 9.90 for both analyses J and N. The estimate of \( \mu_\beta \) was 4.98 for the analysis using J. For the genotyped individuals, the covariate values in \( \mathbf{J}_\beta \) are all \( -1, \) so \( \mathbf{1}\beta + \mathbf{J}_\beta\beta_\beta \) is a vector of values equal to 9.90 - 4.98 = 4.92. For non-genotyped individuals, the covariate values in \( \mathbf{J}_\beta \) can vary widely but many are close to \(-1\) while others are close to 0. This means that \( \mathbf{1}\beta + \mathbf{J}_\beta\beta_\beta \) will include values that range from 9.90 to 4.92, accounting for variation in accuracy of imputation. When \( \mu_\beta \) is not included in the model, these effects are ignored which can reduce the accuracy of predicting non-genotyped individuals. Failing to account for these effects will propagate errors in \( \hat{\sigma} \) and \( \hat{\mu} \), the latter impacting the accuracy of predicting genotyped individuals. Collectively, these errors reduced accuracy from 97% to 96% for the panel including QTL and markers and from 99% to 90% for the panel including only QTL. However, when the panel comprised only markers, the estimates \( \hat{\mu} \) will include both positive and negative values because the phase of markers and QTL are equally likely to take either sign, in which case \( \hat{\mu}_\beta \) will be close to zero as confirmed in the above mentioned replicate where the estimate was -1.18.

Here, results from the analysis with only QTL on the panel are used to show that \( \mu_\beta \) is the mean of \( a_i \) in the founder population and not the mean of the breeding value \( u_i \) of the genotyped individuals. Recall that the QTL model was used with an intercept of \( \beta = 0.0 \) to simulate the data. Thus, when only QTL are on the panel, the true value of \( \beta \) is zero. In analysis J because \( \mu = (\beta + \mu_\beta) \), both \( \hat{\beta} \) and \( \hat{\mu}_\beta \) are estimates of \( \mu_\beta \), and could be pooled, which for the replicate above would be \((9.91 + 8.19) / 2 = 9.05. \) In that replicate, the actual mean of \( a_i \) in G0 was 10.16, which was estimated in the analysis to be 9.05. On the other hand, the mean of the breeding value \( u_i \) in the 9,000 genotyped individuals was 2.4, which is clearly not near the pooled estimate of 9.05 for \( \mu_\beta \). These genotyped individuals included 1,000 selected sires of which 200 were genotyped in each generation from G0 to G4, and 8,000 offspring from G5. The mean values of \( u_i \) for the selected sires were 1.00, 1.59, 2.08, 2.50, and 2.88, respectively, for G0 through G4, and 2.45 for the offspring in G5. It is apparent that the \( \mu_\beta \) parameter corresponding to the covariate \( J \) is the mean of \( a_i \) in the founder population and not the mean breeding value \( u_i \) in the selected individuals. In analysis JC with the covariates centered, the intercept \( \hat{\beta} \) is the value of \( g_i \) when \((\mathbf{m}'_i - \mathbf{v}_i)\mathbf{a} = 0 \), which is the case when \( \mathbf{m}'_i = \mathbf{v}_i \). The estimate \( \hat{\mu} \) was 16.11 in this analysis, but \( \hat{\mu}_\beta \) remained about the same value, namely 7.80. This shows that \( \hat{\mu}_\beta \) has the same interpretation whether the entire matrix of observed and imputed genotypes is centered or not. In neither case does it represent the mean breeding value of selected individuals.

Accuracy of PBLUP increased with heritability, as expected (Table 4). Further, genomic predictions using panels that include causal mutations were near perfect when \( \mu_\beta \) was included in the model. These high accuracies are a reflection of these phenotypes being influenced by only 50 QTL and there being a large training dataset. Accuracy was reduced when \( \mu_\beta \) was not fitted in the model. There was no advantage in terms of accuracy to centering the covariates but MCMC mixing may have been improved although this was not investigated.

2. Bias:

Table 3 shows the regression coefficients of TBV on (G)EBV for \( h^2 = 0.5 \) and \( \mu_\beta = 0.2 \), the same scenarios represented in Table 2. The regression coefficients of TBV on GEBV for each scenario were close to 1 with very low SE, which indicates the genomic predictions exhibited almost no bias. The differences in regression coefficients between analyses were very small, but the marker panel comprising only markers were biased upwards whereas the marker panels that included causal mutations were biased slightly downwards.

Effect of mean QTL effect (\( \mu_\beta = 0 \) vs \( \mu_\beta = 0.2 \))

We hypothesized that the impact of omitting \( \mu_\beta \) from the model will be greatest when \( \mu_\beta \) departs significantly from 0 which is more likely to occur when \( \mu_\beta \) departs from 0. For that reason our base simulation used \( \mu_\beta = 0.2. \) Results are shown in Table 5 for the panel including QTL and markers with \( h^2 = 0.5 \) for \( \mu_\beta = 0.2 \) compared to \( \mu_\beta = 0 \). These results confirmed the benefit of fitting \( \mu_\beta \) was greatest when \( \mu_\beta = 0.2 \) but there was still an advantage to fitting \( \mu_\beta \) when \( \mu_\beta = 0 \). That advantage is likely to erode as the number of QTL increases. Changing the mean
Table 2 Correlations (%$\pm$SE) between TBV and (G)EBV$^a$ for alternative analyses$^b$

<table>
<thead>
<tr>
<th>Genotype Data$^c$</th>
<th>Analyses</th>
<th>JC</th>
<th>J</th>
<th>C</th>
<th>N</th>
<th>PBLUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 QTL + 150 Markers</td>
<td>97.59 ± .00</td>
<td>97.63 ± .00</td>
<td>96.32 ± .00</td>
<td>96.31 ± .00</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>50 QTL Only</td>
<td>99.44 ± .00</td>
<td>99.45 ± .00</td>
<td>90.20 ± .05</td>
<td>90.18 ± .05</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>150 Markers Only</td>
<td>80.47 ± .03</td>
<td>80.47 ± .03</td>
<td>80.44 ± .03</td>
<td>80.44 ± .03</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>No Genotypes</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>41.56 ± .01</td>
<td></td>
</tr>
</tbody>
</table>

$^a$Average correlation between true breeding value (TBV) and (genomic) estimated breeding values from 10 replications validated in Generation 5, comprising 8,000 individuals with genotypes but no phenotypes. The true QTL effects were sampled from a Normal distribution with mean $\mu_\alpha = 0.2$ and scaled to simulate a trait with a heritability 0.5.

$^b$J: includes a covariate for $\mu_g$ in the model, C: entire matrix of imputed and observed genotype covariates centered, JC: both J and C, N: neither J or C, and PBLUP: pedigree-based BLUP.

$^c$The analyses were based on fitting covariates for only 50 QTL, only 150 markers, or both 50 QTL and 150 markers.

Table 3 Regression coefficients ($\pm$SE) of TBV on (G)EBV$^a$ for alternative analyses$^b$

<table>
<thead>
<tr>
<th>Genotype Data$^b$</th>
<th>Analyses$^c$</th>
<th>JC</th>
<th>J</th>
<th>C</th>
<th>N</th>
<th>PBLUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 QTL + 150 Markers</td>
<td>1.06 ± .02</td>
<td>1.06 ± .02</td>
<td>1.06 ± .02</td>
<td>1.06 ± .02</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>50 QTL Only</td>
<td>1.05 ± .02</td>
<td>1.05 ± .02</td>
<td>1.12 ± .04</td>
<td>1.12 ± .04</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>150 Markers Only</td>
<td>0.95 ± .03</td>
<td>0.95 ± .03</td>
<td>0.95 ± .03</td>
<td>0.95 ± .03</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>No Genotypes</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.95 ± .04</td>
<td></td>
</tr>
</tbody>
</table>

$^a$Average Regression coefficients of true breeding value (TBV) on (genomic) estimated breeding values from 10 replications validated in Generation 5, comprising 8,000 individuals with genotypes but no phenotypes. The true QTL effects were sampled from a Normal distribution with mean $\mu_\alpha = 0.2$ and scaled to simulate a trait with a trait with a heritability 0.5.

$^b$The analyses were based on fitting covariates for only 50 QTL, only 150 markers, or both 50 QTL and 150 markers.

$^c$J: includes a covariate for $\mu_g$ in the model, C: entire matrix of imputed and observed genotype covariates centered, JC: both J and C, N: neither J or C, and PBLUP: pedigree-based BLUP.

Table 4 Correlations (%$\pm$SE) between TBV and (G)EBV$^a$ for alternative analyses$^b$ for different heritabilities

<table>
<thead>
<tr>
<th>Heritabilities</th>
<th>Analyses</th>
<th>JC</th>
<th>J</th>
<th>C</th>
<th>N</th>
<th>PBLUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h^2 = 0.1$</td>
<td>93.68 ± .01</td>
<td>93.71 ± .01</td>
<td>92.42 ± .01</td>
<td>92.42 ± .01</td>
<td>31.33 ± .02</td>
<td></td>
</tr>
<tr>
<td>$h^2 = 0.3$</td>
<td>96.79 ± .01</td>
<td>96.81 ± .00</td>
<td>95.19 ± .01</td>
<td>95.19 ± .01</td>
<td>37.07 ± .02</td>
<td></td>
</tr>
<tr>
<td>$h^2 = 0.5$</td>
<td>97.59 ± .00</td>
<td>97.63 ± .00</td>
<td>96.32 ± .00</td>
<td>96.31 ± .00</td>
<td>41.56 ± .01</td>
<td></td>
</tr>
</tbody>
</table>

$^a$Average correlation between true breeding value (TBV) and (genomic) estimated breeding values from 10 replications validated in Generation 5, comprising 8,000 individuals with genotypes but no phenotypes. The true QTL effects were sampled from a Normal distribution with mean $\mu_\alpha = 0.2$ and scaled to simulate a trait with a trait with heritabilities 0.1, 0.3 or 0.5.

$^b$J: includes a covariate for $\mu_g$ in the model, C: entire matrix of imputed and observed genotype covariates centered, JC: both J and C, N: neither J or C, and PBLUP: pedigree-based BLUP. Covariates were fitted for both 50 QTL and 150 Markers.
QTL effect had no impact on bias, except for a slight influence on PBLUP.

**Effect of more QTL and markers (200 SNP vs 2,000 SNP)**

We had hypothesized that the improvement of accuracy from adding an extra covariate for \( \mu_g \) will reduce as the number of QTL increases because \( \mu_g \) is likely to be closer to zero for a trait that is more polygenic. Table 6 shows that PBLUP was largely unaffected by changes to genetic architecture but the accuracy of genomic prediction declined slightly as the number of QTL increases. This reflects the fact that precision of estimating QTL effects is greater when the effects are large and polygenic traits with more QTL must have on average smaller effects when compared at the same genetic variance. The benefit of fitting \( \mu_g \) in the model was virtually eliminated when the number of substitution effects to estimate increases from 150 to 1,500. In contrast to the results of accuracy, there was no impact of QTL number on bias. Centering had no impact on accuracy or bias.

**Centering using the entire matrix of genotype covariates or only the observed genotype covariates**

Table 7 shows the accuracies and regression coefficients of TBV on (G)EBV for the genotype panel with 150 markers, \( h^2 = 0.5 \) and \( \mu_a = 0 \). The analyses were performed after centering: the entire matrix of imputed and observed genotype covariates \((M^* = M - \frac{1}{n_g + n_q} \cdot \mathbf{11} \cdot M)\); only observed genotype covariates \((M^* = M - \frac{1}{n_g} \cdot \mathbf{11} \cdot M_g)\), which is the type of centering done in single-step genomic best linear unbiased prediction (GBLUP); or not centering the covariates \((M^* = M)\). The accuracy of the genomic analysis with covariates centered as \( M^*_g \) but without \( J \) (model \( C^* \)) was about 17% lower than the other genomic analyses. However, when \( J \) was included in the model with covariates centered as \( M^*_g \), the accuracy of prediction was markedly improved.

As explained previously, \( \mu_g = k' \alpha \), where \( k' \) is the expected value of the covariates in the founders, will tend to zero for the marker panel that does not include QTL even with \( \mu_a \neq 0 \). However, even if \( \mu_a = 0 \), in a population undergoing selection when selected individuals are genotyped, \( \mu'_g = \mu_g - \nu' \alpha \neq 0 \), where \( \nu' \) is the expected value of the observed genotype covariates. In this study, selection was used to increase the mean of the trait. Thus, \( \mu'_g \) is expected to be negative because most of the genotyped individuals were from G5, whereas \( \mu_g \) is expected to be zero. The negative estimate of \( \mu'_g \) from 10 replicates of the JC analysis was -0.75. This explains why fitting \( J \) in the model improved the accuracy of genomic prediction when covariates were centered as in single-step GBLUP.

Fernando et al. (Fernando et al. 2014) found that centering using \( M^*_g \) improves the accuracy without \( J \) in the model when the population was not under selection and the genotyped individuals were unselected. In that study, mating was random with no selection, so the allele frequency means of the genotyped individuals were a reasonable approximation of the allele frequency means in the founder population. In contrast, our simulation here shows that centering using \( M^*_g \) can reduce the accuracy when the population is under selection, unless \( J \) is fitted in the model.

In single-step GBLUP, the observed genotypes are commonly centered by subtracting their mean and used to construct a genomic relationship matrix, such as using the first method proposed by VanRaden (VanRaden 2008). Using that genomic relationship matrix in the single-step GBLUP formula in Aguilar et al. (Aguilar et al. 2010) does not account for \( J \). This was recognised in Vitezica et al. (Vitezica et al. 2011) who proposed a modification for populations under selection that involved adding a constant to all elements of the genomic relationship matrix that they derived by equating the sum of the elements of the genomic relationship matrix to the sum of the elements of the numerator relationship matrix. In the appendix of that paper they showed this modification is equivalent to fitting a covariate \( Q = J \) and treating \( -\mu'_g \) as a random effect. In addition to this modification, Christensen et al. (Christensen et al. 2012) proposed a multiplicative scaling to the genomic relationship matrix such that its diagonals have the same mean as the diagonals of the numerator relationship matrix. Vitezica et al. (Vitezica et al. 2011) claimed that \( -\mu'_g \) represents the mean breeding value of selected individuals, and we have confirmed here that this is true provided the observed genotype covariates are centered by their mean.

Most populations are under natural or artificial selection. In many cases, genotypes are only available on selected individuals. In single-step genomic analysis that combine genotyped and non-genotyped individuals in a joint analysis, the mean of observed genotypes are available for centering. If the observed genotypes include QTL, the accuracy of genomic prediction can be severely compromised, unless the \( J \) covariate is fitted in the model. If the observed genotypes are only markers, the accuracy of genomic prediction may not necessarily be improved by fitting \( J \) in the model, but it doesn't do any harm. However, if centering is applied only to the observed genotypes, which is the type of centering used in single-step GBLUP, accuracy could be severely compromised, unless the \( J \) covariate is fitted in the model or an equivalent approach is adopted.

**Acknowledgements**

The authors are grateful to Bruce L. Golden and Hao Cheng for assistance in the implementation of SSBR models, Jack C. M. Dekkers for his constructive comments on design of the simulation, and to very helpful anonymous reviewers. RLF acknowledges useful discussions with Andres Legarra and Zulma Vitezica. This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive grant no. 2015-67015-22947.

**Literature Cited**


Christensen, O. F. and M. S. Lund, 2010 Genomic prediction when some animals are not genotyped. Genet Sel Evol 42: 2–2.


Table 5 Accuracy\(^a\) and bias\(^b\) of genomic prediction (±SE\(_a\)) for alternative QTL distributions\(^c\) and analyses\(^d\)

<table>
<thead>
<tr>
<th>Substitution Effects</th>
<th>Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JC</td>
</tr>
<tr>
<td>Correlations (%)</td>
<td></td>
</tr>
<tr>
<td>(\mu_a = 0)</td>
<td>97.91 ± .00</td>
</tr>
<tr>
<td>(\mu_a = 0.2)</td>
<td>97.59 ± .00</td>
</tr>
<tr>
<td>Regression Coefficient</td>
<td></td>
</tr>
<tr>
<td>(\mu_a = 0)</td>
<td>1.05 ± .04</td>
</tr>
<tr>
<td>(\mu_a = 0.2)</td>
<td>1.06 ± .02</td>
</tr>
</tbody>
</table>

\(^a\)Accuracy was quantified using the average correlation between true breeding value and (genomic) estimated breeding values from 10 replications validated in Generation 5, comprising 8,000 individuals with genotypes but no phenotypes.

\(^b\)Bias was quantified using the average regression coefficients of true breeding value on (genomic) estimated breeding values from 10 replications.

\(^c\)The true QTL effects were sampled from Normal distributions with mean \(\mu_a = 0\) or \(\mu_a = 0.2\) and scaled to simulate a trait with a heritability 0.5.

\(^d\)J: includes a covariate for \(\mu_g\) in the model, C: entire matrix of imputed and observed genotype covariates centered, JC: both J and C, N: neither J or C, and PBLUP: pedigree-based BLUP. Covariates were fitted for both 50 QTL and 150 markers.

---

Table 6 Accuracy\(^a\) and bias\(^b\) of genomic prediction (±SE\(_a\)) for different numbers of QTL\(^c\) and alternative analyses\(^d\)

<table>
<thead>
<tr>
<th>Numbers of QTL</th>
<th>Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JC</td>
</tr>
<tr>
<td>Correlations (%)</td>
<td></td>
</tr>
<tr>
<td>50 QTL + 150 Markers</td>
<td>97.59 ± .00</td>
</tr>
<tr>
<td>500 QTL + 1500 Markers</td>
<td>90.45 ± .01</td>
</tr>
<tr>
<td>Regression Coefficient</td>
<td></td>
</tr>
<tr>
<td>50 QTL + 150 Markers</td>
<td>1.06 ± .02</td>
</tr>
<tr>
<td>500 QTL + 1500 Markers</td>
<td>1.08 ± .03</td>
</tr>
</tbody>
</table>

\(^a\)Accuracy was quantified using the average correlation between true breeding value and (genomic) estimated breeding values from 10 replications validated in Generation 5, comprising 8,000 individuals with genotypes but no phenotypes.

\(^b\)Bias was quantified using the average regression coefficients of true breeding value on (genomic) estimated breeding values from 10 replications.

\(^c\)The true effects for 50 or 500 QTL were sampled from a Normal distribution with mean \(\mu_a = 0.2\) and scaled to simulate a trait with a heritability 0.5.

\(^d\)J: includes a covariate for \(\mu_g\) in the model, C: entire matrix of imputed and observed genotype covariates centered, JC: both J and C, N: neither J or C, and PBLUP: pedigree-based BLUP. Covariates were fitted for either 50 QTL and 150 markers or 500 QTL and 1500 markers.
Table 7 Accuracy\(^a\) and bias\(^b\) of genomic prediction (±SE\(_g\)) when centering for all genotypes or observed genotypes\(^c\)

<table>
<thead>
<tr>
<th>Analyses</th>
<th>Correlations (%)</th>
<th>Regression Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC</td>
<td>82.08 ± 0.06</td>
<td>0.95 ± 0.07</td>
</tr>
<tr>
<td>J</td>
<td>82.08 ± 0.06</td>
<td>0.95 ± 0.07</td>
</tr>
<tr>
<td>C</td>
<td>82.05 ± 0.06</td>
<td>0.95 ± 0.07</td>
</tr>
<tr>
<td>N</td>
<td>82.05 ± 0.06</td>
<td>0.95 ± 0.07</td>
</tr>
<tr>
<td>JC(^*)</td>
<td>82.09 ± 0.06</td>
<td>0.95 ± 0.07</td>
</tr>
<tr>
<td>C(^*)</td>
<td>65.35 ± 0.06</td>
<td>1.16 ± 0.11</td>
</tr>
<tr>
<td>PBLUP</td>
<td>42.66 ± 0.01</td>
<td>0.97 ± 0.07</td>
</tr>
</tbody>
</table>

\(^a\) Accuracy was quantified using the average correlation between true breeding value and (genomic) estimated breeding values from 10 replications validated in Generation 5, comprising 8,000 individuals with genotypes but no phenotypes. The true QTL effects were sampled from Normal distributions with mean µ\(_e\) = 0 and scaled to simulate a trait with heritability 0.5.

\(^b\) Bias was quantified using the average regression coefficients of true breeding value on (genomic) estimated breeding values from 10 replications.

\(^c\) J: includes a covariate for µ\(_g\) in the model, C: entire matrix of imputed and observed genotype covariates centered, JC: both J and C, N: neither J or C, C*: only observed genotype covariates centered, JC*: both J and C*, and PBLUP: pedigree-based BLUP. Covariates were fitted for 150 markers.


Appendix

Here we show that inference about α does not depend on how the genotypes are coded. The marker effects model can be described by the following general model:

\[ y = 1\mu + M\alpha + e \]  

where \( y \) is a vector of observed phenotypes, \( 1 \) is a vector of 1s, \( \mu \) is a general mean, \( M \) is a matrix of marker covariates, coded 0, 1, 2, which might represent the number of copies of the A allele, \( \alpha \) is a vector of random substitution effects of markers, and \( e \) is a vector of residuals. Henderson’s mixed model equations (MME) that correspond to equation (8) are:

\[
\begin{bmatrix}
1'1 & 1'M \\
0 & M'M - M'M + I_{n^2} / n^2
\end{bmatrix}
\begin{bmatrix}
\hat{\beta} \\
\hat{\alpha}
\end{bmatrix} =
\begin{bmatrix}
1'y \\
M'y - M'M1'y / n
\end{bmatrix}
\]

where \( \hat{\beta} \) is the best linear unbiased estimate of the mean, and \( \hat{\alpha} \) is the best linear unbiased predictor of the vector of random substitution effects of all markers. Now we can eliminate \( \hat{\beta} \) from the equations for \( \hat{\alpha} \), by subtracting from those equations the equation for \( \hat{\beta} \) pre-multiplied by \( M'M / n \). Then, the MME are transformed to:

\[
\begin{bmatrix}
1'1 & 1'M \\
0 & M'M - M'M + I_{n^2} / n
\end{bmatrix}
\begin{bmatrix}
\hat{\beta} \\
\hat{\alpha}
\end{bmatrix} =
\begin{bmatrix}
1'y \\
(M - 1\bar{m}')/y
\end{bmatrix}
\]

and substituting \( 1'1 = n \) and \( 1'M = n\bar{m}' \) and its transpose, the transformed MME become:

\[
\begin{bmatrix}
n & n\bar{m} \\
0 & M'M - n\bar{m}' + I_{n^2} / n
\end{bmatrix}
\begin{bmatrix}
\hat{\mu} \\
\hat{\alpha}
\end{bmatrix} =
\begin{bmatrix}
1'y \\
(M - 1\bar{m}')/y
\end{bmatrix}
\]

where \( \bar{m}' \) is the row vector of column means of \( M \) as in \( \bar{m}' = M'M / n \).

Now, consider the coding obtained by centering the marker genotypes as \( M = 1\bar{m}' \). Then the model can be written as:

\[ y = 1\mu + (M - 1\bar{m}')\alpha + e \]  

where \( \mu^* = \mu + \bar{m}' \alpha \). The MME that correspond to equation (10) are:

\[
\begin{bmatrix}
n & 1'(M - 1\bar{m}') \\
(M - 1\bar{m}')(M - 1\bar{m}') + I_{n^2} / n
\end{bmatrix}
\begin{bmatrix}
\hat{\mu} \\
\hat{\alpha}
\end{bmatrix} =
\begin{bmatrix}
1'y \\
(M - 1\bar{m}')/y
\end{bmatrix}
\]

but \( 1'(M - 1\bar{m}') = n\bar{m}' - n\bar{m}' = 0 \), and, similarly, its transpose is \( (M - 1\bar{m}')/1 = 0 \). Then the MME become:

\[
\begin{bmatrix}
n & 0 \\
0 & (M - 1\bar{m}')(M - 1\bar{m}') + I_{n^2} / n
\end{bmatrix}
\begin{bmatrix}
\hat{\mu} \\
\hat{\alpha}
\end{bmatrix} =
\begin{bmatrix}
1'y \\
(M - 1\bar{m}')/y
\end{bmatrix}
\]

Expanding \( (M - 1\bar{m}')(M - 1\bar{m}') \) gives \( M'M - n\bar{m}'M - M1\bar{m}' + 1\bar{m}'1\bar{m}' \), but because \( 1'M = n\bar{m}' \), \( M1 = n\bar{m}' \), and \( 1'1 = n \), as previously shown, the second term, \( 1\bar{m}'M = n\bar{m}'\bar{m}' \),
which is equal to the last term in the expansion. Thus, the MME become:

\[
\begin{bmatrix}
    n & 0' \\
    0 & M'M - \text{m}\text{m}'n + \frac{1_y^2}{\sigma^2}
\end{bmatrix}
\begin{bmatrix}
    \hat{\beta}' \\
    \hat{\alpha}'
\end{bmatrix} =
\begin{bmatrix}
    1\text{y}' \\
    (M - 1\text{m}')\text{y}'
\end{bmatrix}.
\]

(11)

The equations for \(\hat{\alpha}\) in (9) and (11) are identical, and this proves that centering with \(\text{m}'\) doesn't affect inference about \(\alpha\).

Now suppose an arbitrary vector \(\text{v}'\) is used to transform the genotypes as \((M - 1\text{v}')\). Then the model becomes:

\[
y = 1(\mu + \text{v}'\alpha) + (M - 1\text{v}')\alpha + e.
\]

Adding and subtracting \(1\text{m}'\alpha\), the above equation can be written as:

\[
y = 1[\mu + \text{v}'\alpha + (\text{m}' - \text{v}')\alpha] + (M - 1\text{m}')\alpha + e
\]

\[
= 1(\mu + \text{m}'\alpha) + (M - 1\text{m}')\alpha + e,
\]

\[
= 1\mu^* + (M - 1\text{m}')\alpha + e,
\]

which is identical to equation (10), proving inference about \(\alpha\) does not depend on how the genotypes are coded.