G3 Genes | Genomes | Genetics

Efficient single-stage estimation of marker effects

Alencar Xavier[*,1]
[*]Corteva Agrisciences. 8305 NW 62nd Ave. Johnston IA, USA., [1]Purdue University. 915 W State St. West Lafayette IN, USA.

ABSTRACT The evaluation of prediction machines is an important step for a successful implementation of genomic-enabled selection in plant breeding. Computation time and predictive ability constitute key metrics to determine the methodology utilized for the consolidation of genomic prediction pipeline. This study introduces two methods designed to couple high prediction accuracy with efficient computational performance: 1) a non-MCMC method to estimate marker effects with a Laplace prior; and 2) an iterative framework that allows solving whole-genome regression within mixed models with replicated observations in a single-stage. The investigation provides an insights on predictive ability and marker effect estimates. The regression method is compared to various genomic prediction techniques based on cross-validations on 20 maize and 40 soybean datasets, assessing predictions across and within family, respectively. Properties of quantitative trait loci detection and single-stage model were evaluated on simulated datasets. Estimation of marker effects by the new model is compared to a genome-wide association analysis and whole-genome regression methods. The single-stage approach is compared to a GBLUP fitted via restricted maximum likelihood, and a two-stages approaches where genetic values fit a whole-genome regression. The proposed framework provided high computational efficiency, robust prediction across datasets, and accurate estimation of marker effects.

KEYWORDS
Mixed model
Laplace prior
Single-stage
Gauss-Seidel
Predictability
Elapsed time

INTRODUCTION

Genome-wide markers are utilized in plant and animal breeding to capture quantitative trait loci (QTL) and relationship among individuals for prediction and selection (Meuwissen *et al.* 2001, Habier *et al.* 2007, VanRaden 2008). Most individuals in the plant breeding pipeline are genotyped, whereas in animal breeding genomic information enhances the pedigree-based relationship (Henryon *et al.* 2014). With the ever increasing volume of genotypic and phenotypic data, various statistical methods have been developed to handle large datasets, enabling better use of genomic information for more accurate selection and better allocation of resources (Heslot *et al.* 2012).

Evaluating the predictive performance of these various methodologies has become an important step for a successful implementation of genomic-enabled selection (de los Campos *et al.* 2013, Heslot *et al.* 2015), since the prediction method utilized to generate breeding values may have major impact on the short-term genetic gain, as well as long-term changes on the germplasm (Daetwyler *et al.* 2015, Hickey *et al.* 2017).

Genomic predictions models are used to estimate breeding values of observed individuals and to predict breeding values of unobserved individuals in early-generations. Accuracy is the most important criterion to define which technique will be used to generate the breeding values. Besides accuracy, the computational efficiency has also become key components of prediction pipelines due to the growing number of genotyped individuals, observations per individuals, traits, and genotyping density (Georges *et al.* 2018). Hence the method of choice must have two desirable features: computational feasibility and accurate prediction across various scenarios (VanRanden 2008, Misztal and Legarra 2017).

In plant breeding, the calibration of such models are typically done in two steps: 1) Estimate the genetic values from phenotypes of replicated trials; 2) Calibrate marker effects upon the genetic values to estimate the breeding values and enable prediction. This approach is referred to as "two-stages" approach. However, performing analysis in a single-stage can benefit genomic evaluation by jointly modeling genotypes and replicated phenotypes (Liu *et al.* 2014).

Literature is scarce of studies attempting to estimate marker effects directly from the replicated trials. Taskinen *et al.* (2017) pro-

Manuscript compiled: Tuesday 25th June, 2019
[*]Corresponding author: Corteva Agrisciences. 8305 NW 62nd Ave. Johnston IA, USA. Purdue University, Department of Agronomy 915 W State St. West Lafayette IN, USA. E-mail: alencar.xavier@corteva.com

posed using pedigree of ungenotyped individuals for imputation and subsequent estimation of marker effects. Da *et al.* (2014) provided two frameworks to fit genomic models to estimate variance components and marker effects, one approach suitable for large number of observations and another for large number of markers, but not for both. However, such methods often translate into poor computational performance or convergence issues (Misztal 2016).

Fernando *et al.* (2014, 2016) provided a framework where marker effects can be estimated from whole-genome regression (WGR) methods via Markov chain Monte Carlo (MCMC), enabling a broader range of prior assumptions for the distribution of marker effects that can provide predictive advantages in single-stage models (Zhou *et al.* 2018).

Flexible models that enable the estimation of marker effects among other parameters are commonly based on MCMC method (Fernando *et al.* 2014), but these techniques can be computationally prohibitive at times (Wang *et al.* 2015) and must be replaced by Gauss–Seidel iterations (Garrick *et al.* 2014).

This study proposes an efficient non-MCMC solver for WGR and mixed models based on conditioning and iterative updates. The idea is to develop a single-stage model by jointly iterating the two steps of the two-stages analysis. Predictive ability and computing time of proposed framework are evaluated through simulations and cross-validation on real data, comparing it to other standard methods.

STATISTICAL MODELS

Iterative conditional modeling enables solving complex models without the computationally demanding operation (Graser *et al.* 1987, Thompson and Shaw 1992, Misztal and Legarra 2017). In these methods, conditional expectations are used to efficiently estimate variance components, fixed effects, breeding values, and marker effects (Cunningham and Henderson 1968, Da *et al.* 2014, Liu *et al.* 2014, Fernando *et al.* 2014, Taskinen *et al.* 2017).

Two statistical approaches are introduced in this section. First, an iterative algorithm for WGR that speeds up the marker calibration. Second, a framework to enables solving WGR into a model with replicated observations using a specific type of conditioning.

Whole-genome model

This section describes the implementation of the fast Laplace model (FLM), an iterative method to fit a WGR using a Laplace prior. Laplace priors are popular in genetic analysis for QTL detection and genomic prediction (Xu 2007, Xu 2010, Cai *et al.* 2011, Legarra *et al.* 2011).

The implementation below is based on iterative conditional expectation (ICE) estimates of regression coefficients alongside their associated parameters, updating one parameter at a time (Meuwissen *et al.* 2009). This type of algorithm is commonly referred to as coordinate descent (Friedman *et al.* 2010).

Consider the following univariate linear model fitting phenotypes as a function of an intercept and genotypic information:

$$y = 1\mu + M\beta + \epsilon \qquad (1)$$

where y corresponds to a vector of phenotypes, μ is the intercept, M is a matrix of parameters where each m_{ij} cell corresponds to j^{th} locus of the i^{th} individual coding $\{AA, Aa, aa\}$ as $\{-1, 0, 1\}$, β refers to the vector of marker effects, ϵ represent the vector of residuals.

The first operation in each iteration is the intercept update as:

$$\mu = n^{-1} \sum_{i=1}^{n} (y_i - M_i\beta) \qquad (2)$$

Marker effects and regularization parameters are updated one at a time until convergence. Conditioning the response to all but the j^{th} marker ($\tilde{y} = y - 1\mu - M_{-j}\beta_{-j}$) provides a simple probabilistic structure:

$$\tilde{y}|m, \beta \sim N(m_j\beta_j, \sigma^2_\epsilon) \qquad (3)$$

$$\beta_j|\tau_j^2 \sim N(0, \tau_j^2\sigma^2_\epsilon) \qquad (4)$$

where m_j is a vector containing the information of the j^{th} marker, τ_j^2 is the parameter that regularizes β_j, as the marker effect associated with the j^{th} marker is estimated as:

$$\beta_j = \frac{m_j'\tilde{y}}{m_j'm_j + \tau_j^{-2}} \qquad (5)$$

Each marker has an independent regularization. The regularization parameter τ_j^{-2}, which shapes the marker effects collectively into a Laplace distribution, is derived from an inverse-Gaussian density with expectation (Park and Casella 2008):

$$\tau_j^{-2} = \sqrt{\lambda^2\sigma^2_\epsilon\sigma^{-2}_{\beta_j}} \qquad (6)$$

The scale parameter λ^2 was adapted from Legarra *et al.* (2011), as the sum of marker variances:

$$\lambda^2 = \sum_{j=1}^{p} \sigma^2_{m_j} \qquad (7)$$

Further description of the Laplace prior is provided in the appendix. Residual variance and full-conditional marker variance are estimated from the maximum likelihood (Patterson and Thompson 1971, Harville 1977, Searle *et al.* 1992):

$$\sigma^2_{\beta_j} = \frac{\beta'\beta + tr(C^{ii})\sigma^2_\epsilon}{q} = \beta_j^2 + \frac{\sigma^2_\epsilon}{m'm + \tau_j^{-2}} \qquad (8)$$

$$\sigma^2_\epsilon = \frac{y'Py}{n - r_X} = \frac{y'\epsilon}{n - r_X} \qquad (9)$$

where n corresponds to the total number of observations, q is the number of parameters ($q = 1$), and r_X represents the rank of the design matrix of fixed effects ($r_X = 1$).

The optimization path consists of iteratively updating μ, β_1, $\sigma^2_{\beta_1}$, $\tau^{-2}_{\beta_1}$, β_2, $\sigma^2_{\beta_2}$, $\tau^{-2}_{\beta_2}$, ... and σ^2_ϵ. The pseudo-code for the implementation is provided below (Algorithm 1) and an implementation for R is provided in the appendix. In this study, the convergence criteria was set as 10^{-8} for marker effects or a maximum of 300 iterations.

Iterative single-stage model

The previous section presented how the algorithm for FLM works in the case where each individual has a single phenotypic value. Now consider the scenario of replicated trials, where genotyped individuals are replicated across multiple environments. This approach is here referred to as fast Laplace model in single-stage (FLM-SS).

Algorithm 1 Fast Laplace model

```
Compute $m_j'm_j$ for each marker
Compute $\lambda^2 = \sum_{j=1}^{p} \sigma^2_{m_j}$
Set $\lambda^2$ as initial value for all $\tau_j^{-2}$
Repeat until convergence:
    1. Update intercept
        $\mu^{t+1} = \mu^t + n^{-1} \sum_{i=1}^{n} \epsilon_i$
        $\epsilon^{t+1} = \epsilon^t - (\mu^{t+1} - \mu^t)$
    2. Loop for $j^{th}$ marker in $1 : p$
        $\beta_j^{t+1} = \frac{m_j'\epsilon^t + \beta_j^t(m_j'm_j)}{m_j'm_j + \tau_j^{-2}}$
        $\epsilon^{t+1} = e^t - m_j'(\beta_j^{t+1} - \beta_j^t)$
        $\sigma^2_{\beta_j} = \beta_j^2 + \frac{\sigma^2_\epsilon}{m_j'm_j + \tau_j^{-2}}$
        $\tau_j^{-2} = \sqrt{\lambda^2 \sigma^2_\epsilon \sigma^{-2}_\beta}$
    3. Update residual variance
        $\sigma^2_\epsilon = \frac{y'e}{n-1}$
```

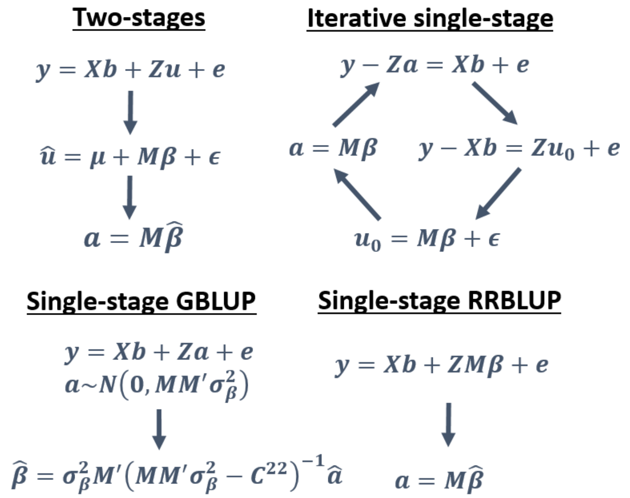

Figure 1 Approaches for modeling breeding values (a) and marker effects (β) from two-stages and different single-stage models.

The term "single-stage" has been used to define the joint modeling of replicated observations with genomic information (Schulz-Streeck *et al.* 2013), which is not to be confused with the "single-step" that elsewhere defines models that combine pedigree and genomic information (Misztal *et al.* 2009).

The following model can illustrate the single-stage procedure:

$$y = Xb + Za + e \tag{10}$$

where y is the vector of phenotypes, X and b represent the design matrix and fixed effect coefficients used to capture nuisance parameters, such as environmental sources of variation. The random terms Z and a correspond to the incidence matrix of individuals and additive genetic effects, hereby estimated from the WGR ($a = M\beta$). For simplicity, residuals ($e = y - Xb - Za$) are assumed to be normally distributed as $e \sim N(0, I\sigma^2_e)$, but the algorithm can be adapted to include residual correlations ($R_e\sigma^2_e$) in order to account for heteroskedasticity.

Fixed effect coefficients are solved via least square, conditioning the response variable to all terms but the fixed effect. This conditioning works by reshaping the linear model into:

$$y - Za = Xb + e \tag{11}$$

Providing the following solution of coefficients:

$$b = (X'X)^{-1}X'(y - Za) \tag{12}$$

In order to avoid building large and dense design matrix of marker effects (ZM), the random effect coefficients are updated using a link function in two steps ($u_0 \rightarrow a$). First, estimate the least-squared genetic values (u_0) as follows:

$$y - Xb = Zu_0 + e \tag{13}$$

Coefficients are solved as:

$$u_0 = (Z'Z)^{-1}Z'(y - Xb) \tag{14}$$

Then, the WGR algorithm introduced in the previous section takes place, solving the following equation to estimate marker effects and breeding values:

$$u_0 = M\beta + \epsilon \tag{15}$$

In this case, the vector of residuals (ϵ) represents genetic signal not captured by the markers. The next step regards the updating of breeding values as:

$$a = M\beta \tag{16}$$

The WGR step can be solved assuming unweighted observations for computational convenience, $a \sim N(Ma, I\sigma^2_\epsilon)$, or weighted according to the number of observations of each genotype, with weights $R_\epsilon = Diag(Z'Z)$, such that $a \sim N(Ma, R_\epsilon\sigma^2_\epsilon)$. Other weights designed for genomic regression in similar settings are described by Garrick *et al.* (2009).

In summary, this single-stage algorithm works through the iterative update of b, u, and a until convergence (Figure 1). Using Gauss-Seidel (Legarra and Misztal 2008) to update regression coefficients, this system of equations mitigates the computational burden of building and inverting large matrices.

Additional random effects

The study has focused on simple mixed models with fixed effects and a single random effect to model genetics. However, the single-stage approach may also include multiple random effects into the model by conditioning the response variable to the fixed effects and genetic term.

Consider a model with one additional random effect:

$$y = Xb + Za + Wg + e \tag{17}$$

Conditioning the response variable to all effects but the additional random effect ($\tilde{y} = y - Xb - Za$), yields:

$$\tilde{y} = Wg + e \tag{18}$$

Assuming $g \sim N(0, I\sigma^2_g)$, the solution for the the random effect coefficients is given by:

$$g = (W'W + kI)^{-1}W'\tilde{y} \tag{19}$$

where $k = \sigma^2_e\sigma^{-2}_g$. The solution for the residual variance is provided in equation (9) replacing ϵ by e. Conditional to other model terms, the variance component associated to this random effect is estimated as (Patterson and Thompson 1971, Harville 1977):

$$\sigma^2_g = \frac{g'g}{n_w - tr(C^{ii})k} = \frac{g'g}{n_w - k\sum_{j=1}^{n_w}(w_j'w_j + k)^{-1}} \tag{20}$$

where n_w is the number of columns of W. For random effects with non-orthogonal design matrices, such as adjacent matrices to model spatial auto-correlation, the variance component can be efficiently approximated as (Schaeffer 1986):

$$\sigma^2_g \cong \frac{(y - Xb)'Wg}{n\sum_{j=1}^{n_w}\sigma^2_{w_j}} \tag{21}$$

MATERIALS AND METHODS

Genomic prediction cross-validation analysis

Soybean dataset. Soybean dataset with 40 bi-parental families available in the R package SoyNAM. Cross-validations were run as 5-fold within family and as leave-family-out. Each family contains approximately 140 individuals genotyped with 4320 markers, and the number of polymorphic markers ranged from 547 to 1262 within family. The soybean trait under evaluation was the best linear unbiased predictors (BLUP) of grain yield collected in as many as 18 environments. BLUPs were generated by modeling grain yield as a function of environment (random effect), genetic merit (random effect) and local check value (fixed effect). More details about the SoyNAM population are described by Diers *et al.* (2018) and Xavier *et al.* (2018).

Maize dataset. Commercial maize general combining ability (GCA) of grain yield, comprising 20 datasets as the combination of two heterotic groups and 10 geographies. For each dataset, the GCA values were computed from 5 years of hybrid phenotypic data (2013-2017) modeled from a classic GCA model (Jacobson *et al.* 2014, Heslot *et al.* 2015) as a function of local environment (fixed effect), target parent (random effect), tester (random effect) and spatial variation using splines. The number of phenotypic records per double-haploids varied from 10 to approximately 200 observations. The average number of individuals per dataset was 4146, ranging from 258 to 12228 double-haploids, each genotyped with 13525 SNP markers. Within dataset the number of segregating SNPs with MAF above 0.05 ranged from 6192 to 12038.

Evaluation criteria. The cross-validation focused on two criteria: 1) the predictive ability measured from 5-fold cross-validations in the maize dataset and within-family soybean dataset, as the correlation between predicted and observed genetic values. Individuals were sampled at random in each cross-validation and this procedure was repeated 20 times. The leave-family-out cross-validation in the soybean dataset works by using 39 families to predict the family left out, and repeating this procedure for all 40 families; and 2) the elapsed time for calibrating the model using the whole data. Elapsed time applies to the maize dataset only, since the computation time was not relevant for the small soybean bi-parental populations.

Prediction methods. FLM was compared to a set of methods designed for high dimensional problems that are implemented and freely available in R (R core team 2019), including: Bayesian alphabet (A, B, C, RR, L) and reproducing kernel Hilbert spaces (RKHS) implemented in BGLR (Perez and de los Campos 2014); BayesCπ and BayesDπ implemented in the R package bWGR; GBLUP with REML variance components implemented in rrBLUP (Endelman 2011); boosting implemented in gbm (Ridgeway 2007); L_1L_2 machines - ridge regression, elastic-net and LASSO implemented in glmnet (Friedman *et al.* 2010); partial least square (PLS) implemented in pls (Mevik and Wehrens 2007); random forest implemented in ranger (Wright and Ziegler 2015); ν and ϵ support vector machines (SVM) implemented in kernlab (Karatzoglou *et al.* 2004); the empirical Bayesian LASSO from Cai *et al.* (2011) implemented in EBglmnet (Huang and Liu 2016); and the extended Bayesian LASSO from Legarra *et al.* (2011) implemented in VIGoR (Onogi and Iwata 2016). Similar to FLM, the latter two methods are efficient implementations based on Laplace prior.

Methods above were deployed with default settings. Tuning parameters for ridge, LASSO and elastic-net were computed through 10-fold cross validation in the training set. To mitigate the computational burden necessary to tune parameters, PLS used 5 components and the empirical Bayes Lasso hyperparameters a-b were set to 0.5. The Gaussian kernel employed for RKHS was computed as $K = exp(-\delta D^2)$ where D^2 is the squared Euclidean distance matrix computed from the marker information and δ is the average value of D^2. The GBLUP model utilized the genomic relationship matrix described by VanRaden (2008).

Detection of QTLs

An experimental population was generated through simulation to evaluate FLM estimates of large effect parameters. This population was generated as F2 bi-parental cross with 1000 individuals. Then, 250 individuals were randomly selected and randomly mated to generate a new population of 1000 individuals. This bottle-necking with subsequent random mating was repeated 5 times. The resulting allele frequency ranged from 0.32 to 0.63. The simulated genome had 10 chromosomes of length 100 cM. The genotyping density was 0.5 marker/cM. A causative marker was assigned to the center of each chromosome with alternating values of positive and negative one.

The response variable was evaluated under heritability of 0.25 and 0.50. The ability of FLM to detect major genes was compared to the Bayesian ridge regression and Bayesian LASSO implemented in the R package BGLR (Perez and de los Campos 2014), and a mixed model association based on P3D algorithm (Zhang *et al.* 2010) implemented in the R package NAM (Xavier *et al.* 2015). Three population sizes were evaluated to estimate the allele effects: 250, 500 and 1000 individuals.

Evaluation of single-stage model

Breeding data is inherently unbalanced. Genotypes are often unreplicated or not equally distributed across environments, and observations from different environments present a variable degree of noise. The single-stage approach was evaluated on simulated datasets that recreated such condition.

Simulated dataset. The simulations were based on assigning the simulated individuals described in the previous section, a genetic pool with 1000 genotypes, to a random set of environments. Each simulated scenario was performed with a combination of number of observations across trials (n = 250, 500, 1000, 2500 and 5000) and genetic architectures (10, 50 and 100 QTL). The number of environments for each simulation was sampled from an uniform distribution between 4 and 10. To simulate heteroscedasticity, each environment had a different heritability sampled from a uniform distribution between 0.25 to 0.75. Individuals were sampled with replacement, such that each environment had an unequal number of entries. Each scenario (combination of size and genetic architecture) was repeated 20x with different seeds to sample the individuals, number of locations, and heritability of the locations.

For the simulated scenarios with less than 1000 observations, the majority of the genotypes were unreplicated, since the observed individuals were sampled from a pool of 1000 genotypes. Selection across unreplicated trials are not unusual when genomic prediction is deployed (Sebastian *et. al.* 2010) since genotypes are connected through the relationship information captured by markers (Habier *et. al.* 2007). Phenotypic values were generate by adding an environmental effect and random noise to the true breeding values. For simplicity, genotype-by-environment interactions, non-additive genetics, and spatial noise were not considered.

Prediction methods. Three methods were evaluated. 1) FLM-SS described in the methods section was implemented in R using the RcppEigen package (Eddelbuettel 2011). 2) Two-stages approach described by Schulz-Streeck *et al.* (2013) based on fitting the best linear unbiased estimators (BLUE) of genetic values without using genomic information (first step), treating environment as random effect, and subsequently fitting a WGR (second step) to estimate breeding values. The first-stage BLUEs were computed with the lme4 package (Bates *et al.* 2015), and markers were fitted with the Bayesian LASSO implemented the BGLR package (Perez and de los Campos 2014), carrying over the covariances from the first-stage (FS) to account for the environmental heteroscedasticity, assuming the second-stage residual covariances to be inherited from the first-stage. 3) GBLUP fitted with the commercial software ASReml (Gilmour *et al.* 2008) using a genomic additive relationship matrix (Zeng *et al.* 2005, Xu 2013). GBLUP is also a single-stage procedure to generate breeding values (Figure 1), however marker effects are not explicitly computed for the prediction on new individuals.

Evaluation criteria. The criteria for comparison was the computation time necessary to fit the model as the elapsed time, and the prediction accuracy as the correlation between estimated breeding values and true breeding values.

Statistical models. The evaluated models aim to estimate breeding values ($a = M\beta$) from phenotypes (y). GBLUP and FLM-SS fit environment (Xb) as fixed effect and genetics as random effect, as Za and $Z(M\beta)$, respectively. The two-stages fits environment (Xb) as random and genetic merit as fixed effect (Zu) in the first stage, followed by modeling the genetic merit (u) as function of intercept (μ) and marker effects ($M\beta$), weighting observations (R_{FS}) as $R_{FS} = Diag(Z'V^{-1}Z)$, where $V = XX'\sigma^2_b + I\sigma^2_e$, which translates into weights $wgts = R^{-1}_{FS}$. The three models can be summarized as follows:

1) Single-stage (FLM-SS):

$$y = Xb + Z(M\beta) + e \tag{22}$$

2) Two-stages:

$$\begin{aligned} y &= Xb + Zu + e \\ u &= \mu + M\beta + \epsilon, \quad \epsilon \sim N(0, R_{FS}\sigma^2_\epsilon) \end{aligned} \tag{23}$$

3) GBLUP:

$$y = Xb + Za + e, \quad a \sim N(0, MM'\sigma^2_\beta) \tag{24}$$

RESULTS

Genomic prediction analysis

The summary of prediction statistics from cross-validation is presented in Figure 2 for the maize dataset, and in Figure 3 for the soybean dataset.

In maize, kernel methods RKHS and ϵ-SVR provided the highest predictive ability for both heterotic groups. FLM provided an average performance for the heterotic group 1 and the third most predictive method for the heterotic group 2. Most methods provided satisfying predictive ability, except for the empirical Bayesian LASSO and boosting, which presented inferior predictive performance. In soybeans, FLM was the most predictive methodology within-family and the second most predictive under leave-family-out cross-validation.

Under this criterion of computation time (Figure 2, bottom), PLS and the three non-MCMC implementations of the Laplace prior provided the lowest computational cost. All four kernel methods provided high computational cost.

Learning properties

The ability of different approaches to correctly estimate major effects through simulation is presented in Figure 4. Marker effect estimated from genome-wide association analysis were the closest to the true simulated values, however it provided an abundance of false positives across the genome.

In most cases, the allele effect estimated by FLM was closer to the true value than its MCMC counterpart, the Bayesian LASSO, and this difference was more evident in the low heritability scenario (Figure 4 bdf). Bayesian ridge regression captured the large effects reasonably well in scenarios with where the heritability was 0.5, but the estimates were not close to the real values in any situation. In general, more realistic values were achieved by all methods as the population size and heritability increased.

Single-stage efficiency

The comparison of accuracy and speed among GBLUP, two-stages approach, and single-stage (FLM-SS) is presented in Figure 5. The accuracy of GBLUP was sensitive to the number of observations when the trait was controlled by a small number of QTLs. As the number of QTL increased, the predictive advantage of FLM-SS and two-stages over GBLUP decreased, and GBLUP outperformed the other two methods under the scenario with the lowest number of observations. However, its computation time was more sensitive to the number of observations.

The two-stages predictive performance was intermediate between GBLUP and FLM-SS for 10 QTLs, and under-performed GBLUP and single-stage for the scenarios with 50 and 100 QTL. In terms of computation time, two-stages was more efficient than to the GBLUP method but less efficient than the FLM-SS. The discrepancy in computation time between single-stage and two-stages can be attributed primarily to the MCMC sampling in the second step, but also to the estimation of variance components in the first step. For most cases, FLM-SS provided the highest predictive and computational performance. GBLUP performed best under small sample size and large number of QTLs.

DISCUSSION

The discussion section frames FLM as a potential method of choice for genomic prediction in plant breeding. The proposed methodology provided accurate prediction across datasets, as well as computational efficiency. Besides the predictive and computational performance, the FLM is an easy-to-implement regression method (Algorithm 1) without the need for complicated prior specifications, tuning or matrix inversion.

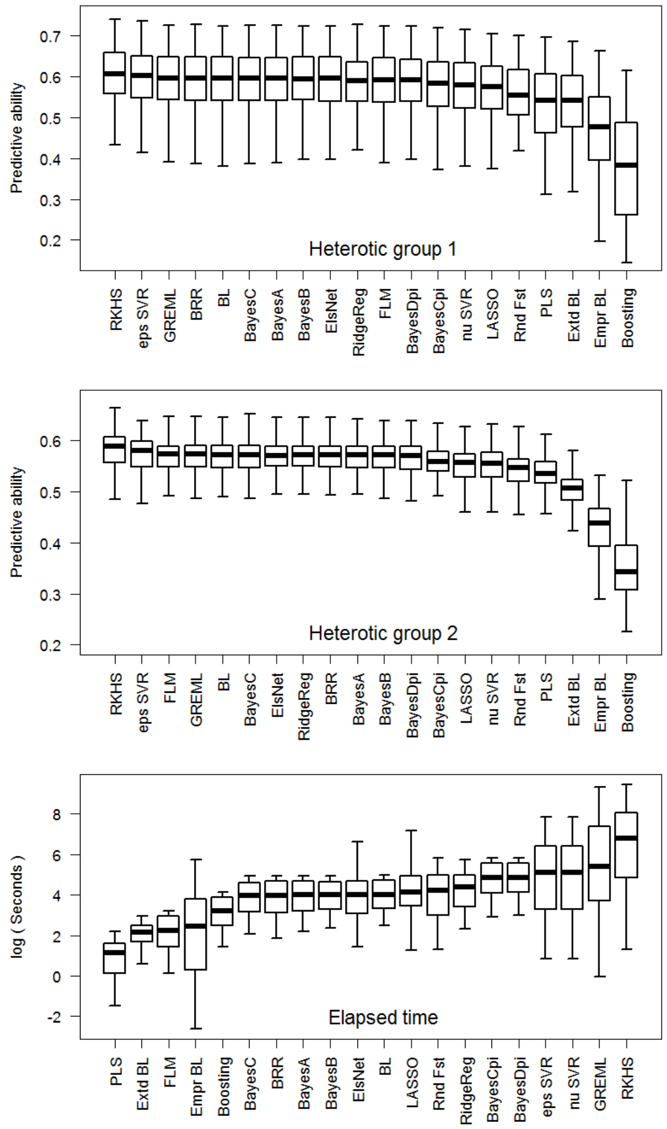

Figure 2 Maize data: Box-and-whiskers plot of predictive ability by heterotic group (**top**, **center**) and computation time to fit the model (**bottom**).

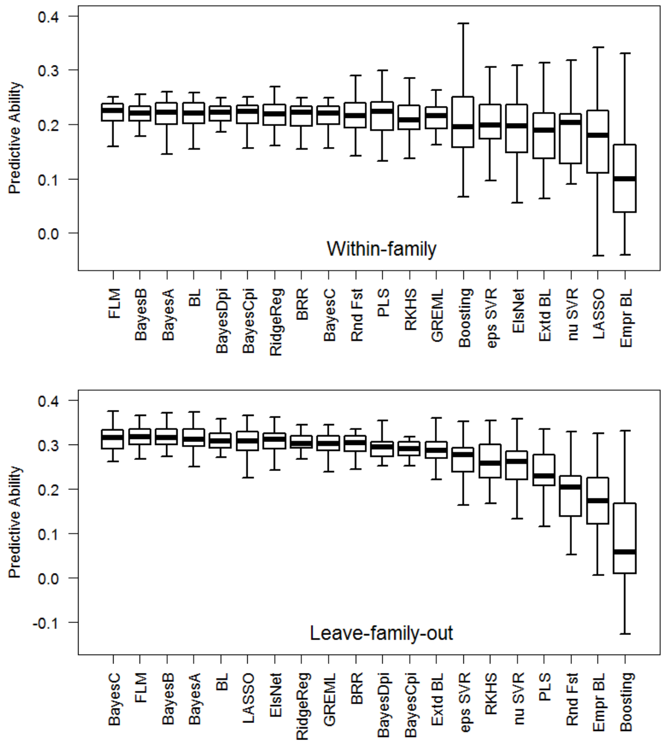

Figure 3 Soybean data: Box-and-whiskers plot displaying the predictive ability of prediction methods across 40 bi-parental family sets. Prediction within-family (**top**) and leave-family-out (**bottom**).

Predictive ability

Most methods provide comparable predictive performance (Perez-Rodriguez *et al.* 2012, Howard *et al.* 2014, Xavier *et al.* 2016). This study compared prediction methods across-family (maize), within-family and family-out predictions (soybean), with predictive ability around 0.2, 0.3 and 0.5, respectively, consistent with literature (Legarra *et al.* 2008, Lian *et al.* 2014, Xavier *et al.* 2016). Within-family predictions rely on modeling the Mendelian segregation between markers and QTLs, whereas across-family predictions are based on capturing the relationship among families (Habier *et al.* 2007, Daetwyler *et al.* 2013, Lehermeier *et al.* 2014). FLM provided competitive values of predictive ability for both maize and soybean datasets. However, the predictive performance of models may vary according to genetic architecture, marker density, trait heritability, and the size of the training set (de los Campos *et al.* 2013, Legarra *et al.* 2015).

Feature selection is a desirable statistical property known to improve the parsimony and predictive ability of WGR models (Wimmer *et al.* 2013). FLM deploys the so-called Laplacian variable selection (O'Hara and Sillanpää 2009), which imposes strong shrinkage without eliminating the parameters from the model. Markers not linked to QTLs often play an important role on prediction by capturing relationship among individuals (Habier *et al.* 2007). In addition, when regression coefficients have priors shaped by heavy tailed distribution, such as Laplace and Student's t, models are suited to capture QTLs because these priors relax the shrinkage of markers with large effect (de los Campos *et al.* 2009, Kärkkäinen and Sillanpää 2012). Other models with similar properties include BayesA, BayesB, BayesC and the Bayesian LASSO (de los Campos *et al.* 2009, Habier *et al.* 2011, Heslot *et al.* 2012, Kärkkäinen and Sillanpää 2012, Legarra *et al.* 2015).

From the signal detection perspective, models able to capture relationship and accurately detect QTLs are deployed for association studies and haplotype analysis (Fernando and Garrick 2013, Hayes 2013, Yang *et al.* 2014, Daetwyler *et al.* 2015, Fernando *et al.* 2017, Goiffon *et al.* 2017). For the scenarios under evaluation, FLM provided a more accurate marker effects estimation than the Bayesian LASSO and ridge regression, with less spurious association than GWA (Figure 5). Both FLM and Bayesian LASSO have a Laplace prior, but with substantial algorithmic differences. Empirical priors have been reported to improve the predictive properties of Laplace models (Xu 2007, Yi and Xu 2008, Xu 2010, Cai *et al.* 2011), thus FLM likely benefits from regularization free of hyper-parameters. Moreover, iterative algorithms often outperform their MCMC counterpart in terms of accuracy (Hayashi and Iwata 2010, Sun *et al.* 2012, Wang *et al.* 2015). The resulting improvement in signal detection translates into higher predictive ability in scenarios

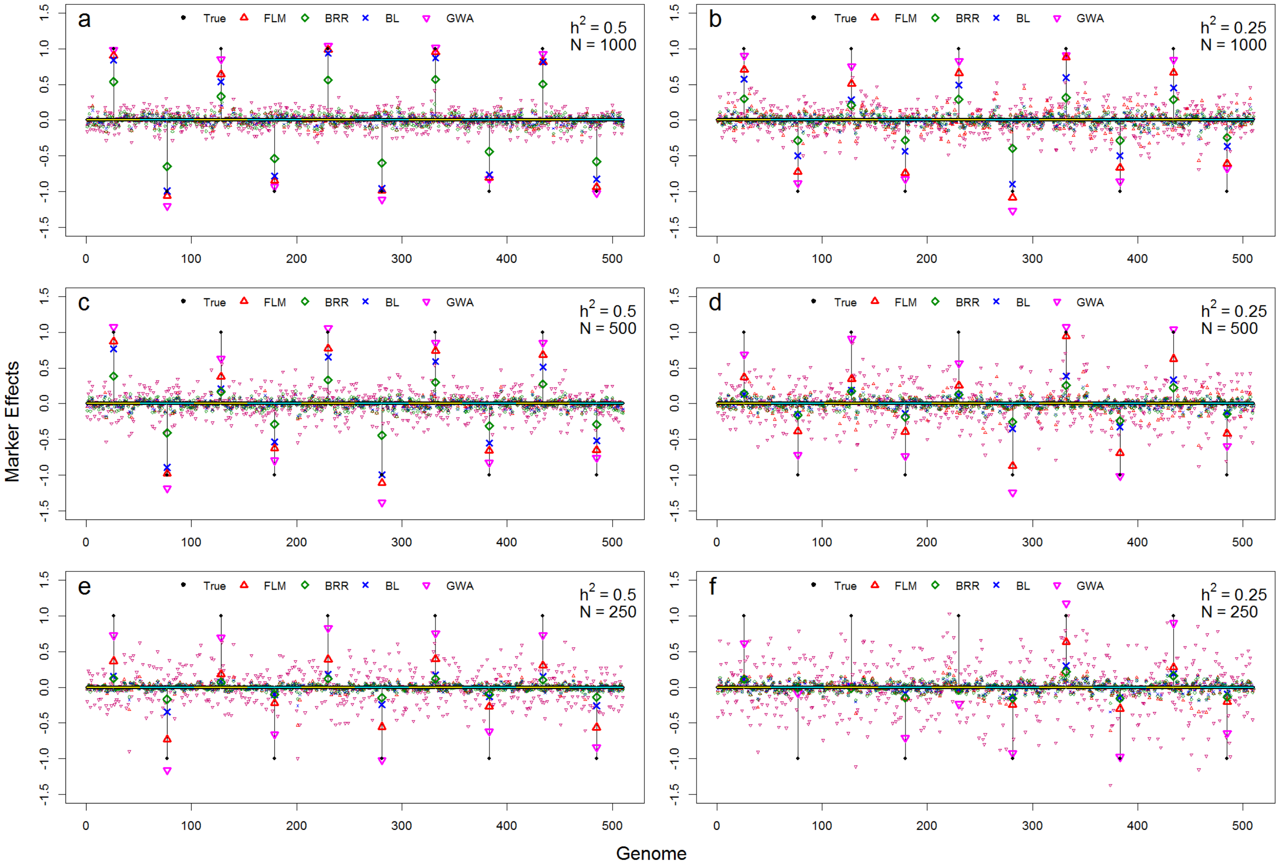

Figure 4 Simulation-based evaluation of marker effect estimation (y-axis) across the genome (x-axis) with varying the heritability and number of individuals, testing: Fast Laplace model (FLM), Bayesian ridge regression (BRR), Bayesian LASSO (BL), genome-wide associations (GWA) analysis, and the true value (True). Effects were plotted larger at the QTL positions and smaller in every other locus.

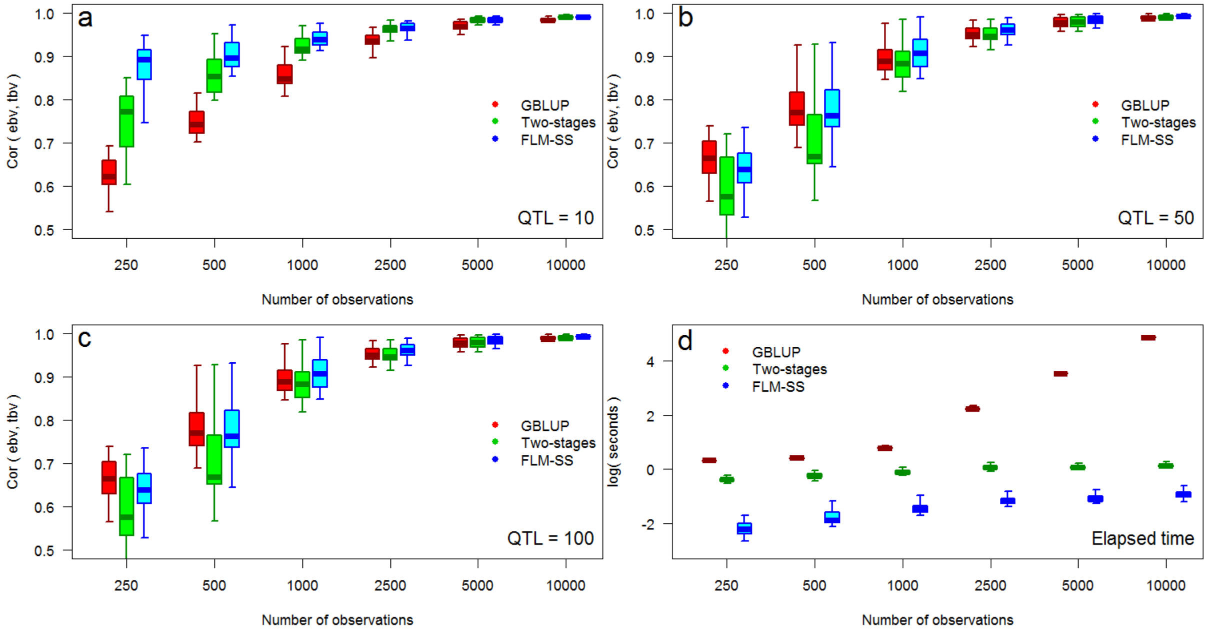

Figure 5 Simulation-based comparison of methods: GBLUP from replicated trials (**GBLUP**); two-stages approach where whole-genome regression fitted on genetic values (**Two-stages**); and an iterative single-stage approach, FLM single-stage (**FLM-SS**). Accuracy under 10 QTL (a), 50 (b) and 100 (c) QTL and elapsed time (d) to fit the model.

where capturing linkage disequilibrium is more important than the relationship among individuals, as depicted in within-family predictions in soybeans (Figure 3). Note that kernel methods, such as RKHS and SVR, were accurate on the maize dataset but not be particularly effective for predictions within bi-parental populations.

The genetic signal captured by WGR methods is solely additive, which is desirable to estimate breeding values but sub-optimal for the prediction of phenotypes. Unlike additive models, semi-parametric methods are capable of capturing non-linear relationship patterns and different levels of epistasis. For this reason, additive models are frequently outperformed by semi-parametric methods, such as RKHS, SVR, random forest and neural networks (Gianola *et al.* 2006, de los Campos *et al.* 2010, Perez-Rodriguez *et al.* 2012, Desta and Ortiz 2014, Howard *et al.* 2014). For the datasets under evaluation, linear models were as predictive as semi-parametric methods, which suggests that most genetic signal was due to additive genetics. However, RKHS and ϵ-SVR were the most accurate methods in the maize data, which supports that some degree of epistasis controls the general combining ability ability of grain yield.

Both RKHS and ϵ-SVR are kernel methods that utilize a Gaussian kernel, but these methods differ with regards to their loss-functions. Whereas RKHS follows a L_2 loss that penalizes square error and coefficients, whereas ϵ-SVR only penalizes the error greater than ϵ (Hastie *et. al.* 2009). Interestingly, ν-SVR did not provide the same degree of predictive ability, despite sharing the same kernel as RKHS and ϵ-SVR.

Computational performance

The time required to calibrate a prediction machine is an important factor to chose a methodology when genomic prediction is utilized for various traits, with often model re-calibration (Meuwissen *et al.* 2009, Hayashi and Iwata 2010, Sun *et al.* 2012, Wang *et al.* 2015). Results indicate a clear discrepancy across methods with regards to the computing time required to fit the prediction models in the maize dataset. Figure 2 shows the most computationally efficient methods were PLS and the non-MCMC implementations of models with Laplace prior. Other regression-type methods provided intermediate efficiency and kernel-type methods were computationally expensive.

Most prediction methods display some computation burden: tuning parameters in machine learning methods; MCMC iterations in Bayesian methods; variance components in GBLUP; and matrix inversion or decomposition in kernel methods. FLM estimates full-conditional variance components, dismissing expensive matrix operations, cross-validation for tuning parameters, or MCMC. The other two prediction methods that efficiently implement Laplace prior, the empirical Bayesian LASSO and the extended Bayesian LASSO, did not provide satisfactory predictive ability. Besides FLM, our results indicate that BayesC is also a cost-effective regression method by providing low computational cost with reasonable predictive ability across datasets.

It is important to point out that kernel methods can be a suitable alternative in high dimensional models, since these rely on the number of individuals rather than the number of parameters. Kernel methods are computationally demanding for two other reasons: it is necessary to 1) build the kernel and 2) compute its inversion or Eigendecomposition. The time needed to build the kernel depends on the number of both individuals and parameters. Many kernels require the computation of distance matrices, which is more computationally demanding. However, the additional computational cost of RKHS pays off in terms of predictive ability (Figure 2).

For the prediction of new observation, kernels must augmented with the genotypes of observed and unobserved individuals, making the inversion or spectral decomposition very challenging. This is particularly cumbersome in plant breeding where the size of the offspring being predicted and selected is much larger than the training set, whereas regression and tree models can be stored and easily employed for prediction of new observations.

Single-stage modeling

The two-step model and FLM-SS were faster than GBLUP by an order of magnitude. This difference can be attributed to the sparse nature of the algorithm and the complexity associated to the estimation of variance components. Whereas GBLUP was fit using AI-REML, a general-purpose algorithm, whereas FLM-SS was specifically designed to provide efficient computation of breeding models. The two-stages model provided an intermediate outcome.

The poor accuracy of the GBLUP model in scenarios with few QTLs can be attributed to the statistical nature based an infinitesimal model. GBLUP works by capturing the relationship among individuals (Habier *et al.* 2007), whereas single-stage and two-steps models enable fitting priors that are suitable to capture both relationship and QTL (Wolc *et al.* 2016). Similar results were reported by Zhou *et al.* (2018), where one-step BayesA and BayesB consistently outperformed the one-step GBLUP under various simulated scenarios. This advantage is also depicted in within-family prediction (Figure 3) where the prediction power comes from detecting LD between markers and QTL, as well as in Figure 4, where it takes a larger number of observations to BRR (counterpart of GBLUP) to identify large effect markers (de los Campos *et al.* 2013, Henryon *et al.* 2014, Hickey *et al.* 2017).

Frameworks where marker effects are estimate alongside all other parameters are not new, but greatly underestimated (Fernando *et al.* 2014, Liu *et al.* 2014, Taskinen *et al.* 2017). Methods and implementations of genomic prediction have been incorporated from animal breeding into plant breeding without much consideration about the large differences in data flow and other statistical properties (Heslot *et al.* 2015, Hickey *et al.* 2017). Two of the major factors that differentiate plant and animal breeding are replicated trials and offspring size. The single-stage framework proposed in this study was design for genomic prediction following the plant breeding data structure, being beneficial from the computational and predictive standpoint.

CONCLUSION

A robust prediction methodology is a key component for a successful genomic-assisted breeding pipeline. This study introduced a fast and accurate algorithm for solving a WGR with Laplace prior, alongside a single-stage methodology that allows to connect WGR into mixed models with replicated observations.

The proposed framework provided more accurate predictions and higher computational efficiency than other methods based on a cross-validation evaluation on maize and soybean datasets. With a simulated dataset, it was shown that the fast Laplace model provided reasonably accurate estimation of QTL effects, being less biased than Bayesian LASSO and ridge regression, and proving less spurious signals than genome-wide association analysis. The algorithm extension to single-stage also presented promising properties, benefiting both computation and prediction.

ACKNOWLEDGEMENTS

Corteva AgriSciences provided the maize dataset. David Habier is acknowledge for revision and useful discussions. Makram Geha contributed to the overall goal the research. Shizhong Xu provided an insight on the coordinate descent methodology.

CONFLICT OF INTEREST

The author declares no conflict of interest.

DATA AVAILABILITY

The soybean data is available in the R package SoyNAM. The maize data can be made available upon request. The implementation of FLM-SS can be made available for research purposes.

LITERATURE CITED

Bates, D., M. Mächler, B. Bolker, and S. Walker, 2015 Fitting linear mixed-effects models using lme4. Journal of Statistical Software **67**: 1–48.

Cai, X., A. Huang, and S. Xu, 2011 Fast empirical bayesian lasso for multiple quantitative trait locus mapping. BMC bioinformatics **12**: 1–12.

Cunningham, E. and C. R. Henderson, 1968 An iterative procedure for estimating fixed effects and variance components in mixed model situations. Biometrics **24**: 13–25.

Da, Y., C. Wang, S. Wang, and G. Hu, 2014 Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using snp markers. PloS one **9**: e87666.

Daetwyler, H. D., M. P. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics **193**: 347–365.

Daetwyler, H. D., M. J. Hayden, G. C. Spangenberg, and B. J. Hayes, 2015 Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. Genetics **200**: 1341–1348.

de Los Campos, G., D. Gianola, G. J. Rosa, K. A. Weigel, and J. Crossa, 2010 Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel hilbert spaces methods. Genetics Research **92**: 295–308.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics **193**: 327–345.

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

Desta, Z. A. and R. Ortiz, 2014 Genomic selection: genome-wide prediction in plant improvement. Trends in plant science **19**: 592–601.

Diers, B. W., J. Specht, K. M. Rainey, P. Cregan, Q. Song, *et al.*, 2018 Genetic architecture of soybean yield and agronomic traits. G3: Genes, Genomes, Genetics **8**: 3367–3375.

Eddelbuettel, D., R. François, J. Allaire, K. Ushey, Q. Kou, *et al.*, 2011 Rcpp: Seamless r and c++ integration. Journal of Statistical Software **40**: 1–18.

Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with r package rrblup. The Plant Genome **4**: 250–255.

Fernando, R., A. Toosi, A. Wolc, D. Garrick, and J. Dekkers, 2017 Application of whole-genome prediction methods for genome-wide association studies: a bayesian approach. Journal of Agricultural, Biological and Environmental Statistics **22**: 172–193.

Fernando, R. L., H. Cheng, B. L. Golden, and D. J. Garrick, 2016 Computational strategies for alternative single-step bayesian regression models with large numbers of genotyped and non-genotyped animals. Genetics Selection Evolution **48**: 1–8.

Fernando, R. L., J. C. Dekkers, and D. J. Garrick, 2014 A class of bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genetics Selection Evolution **46**: 1–13.

Fernando, R. L. and D. Garrick, 2013 Bayesian methods applied to gwas. In *Genome-wide association studies and genomic prediction*, edited by C. van der Werf Gondro and B. Hayes, pp. 237–274, Springer.

Garrick, D., J. Dekkers, and R. Fernando, 2014 The evolution of methodologies for genomic prediction. Livestock Science **166**: 10–18.

Garrick, D. J., J. F. Taylor, and R. L. Fernando, 2009 Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution **41**: 55.

Georges, M., C. Charlier, and B. Hayes, 2018 Harnessing genomic information for livestock improvement. Nature Reviews Genetics p. 1.

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

Gilmour, A., B. Gogel, B. Cullis, R. Thompson, D. Butler, *et al.*, 2008 Asreml user guide release 3.0. VSN Int Ltd .

Goiffon, M., A. Kusmec, L. Wang, G. Hu, and P. Schnable, 2017 Improving response in genomic selection with a population-based selection strategy: optimal population value selection. Genetics **206**: 1675–1682.

Graser, H.-U., S. Smith, and B. Tier, 1987 A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood 1. Journal of animal science **64**: 1362–1370.

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

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. BMC bioinformatics **12**: 1–12.

Harville, D. A., 1977 Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association **72**: 320–338.

Hastie, T., J. Friedman, and R. Tibshirani, 2009 *The elements of statistical learning*, volume 2. Springer series in statistics New York.

Hayashi, T. and H. Iwata, 2010 Em algorithm for bayesian estimation of genomic breeding values. BMC genetics **11**: 1–9.

Hayes, B., 2013 Overview of statistical methods for genome-wide association studies (gwas). In *Genome-wide association studies and genomic prediction*, edited by C. van der Werf Gondro and B. Hayes, pp. 149–169, Springer.

Henryon, M., P. Berg, and A. C. Sørensen, 2014 Animal-breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains. Livestock Science **166**: 38–47.

Heslot, N., J.-L. Jannink, and M. E. Sorrells, 2015 Perspectives for genomic selection applications and research in plants. Crop

Science **55**: 1–12.

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

Hickey, J. M., T. Chiurugwi, I. Mackay, W. Powell, A. Eggen, *et al.*, 2017 Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nature genetics **49**: 1297–1303.

Howard, R., A. L. Carriquiry, and W. D. Beavis, 2014 Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3: Genes, Genomes, Genetics **4**: 1027–1046.

Huang, A. and D. Liu, 2016 Ebglmnet: a comprehensive r package for sparse generalized linear regression models. Bioinformatics .

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo, 2014 General combining ability model for genomewide selection in a biparental cross. Crop Science **54**: 895–905.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis, 2004 kernlab-an s4 package for kernel methods in r. Journal of statistical software **11**: 1–20.

Kärkkäinen, H. P. and M. J. Sillanpää, 2012 Back to basics for bayesian model building in genomic selection. Genetics **191**: 969–987.

Legarra, A., P. Croiseau, M. P. Sanchez, S. Teyssèdre, G. Sallé, *et al.*, 2015 A comparison of methods for whole-genome qtl mapping using dense markers in four livestock species. Genetics Selection Evolution **47**: 1–10.

Legarra, A. and I. Misztal, 2008 Computing strategies in genome-wide selection. Journal of Dairy Science **91**: 360–366.

Legarra, A., C. Robert-Granié, P. Croiseau, F. Guillaume, and S. Fritz, 2011 Improved lasso for genomic selection. Genetics research **93**: 77–87.

Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen, 2008 Performance of genomic selection in mice. Genetics **180**: 611–618.

Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan, *et al.*, 2014 Usefulness of multiparental populations of maize (zea mays l.) for genome-based prediction. Genetics **198**: 3–16.

Lian, L., A. Jacobson, S. Zhong, and R. Bernardo, 2014 Genomewide prediction accuracy within 969 maize biparental populations. Crop Science **54**: 1514–1522.

Liu, Z., M. Goddard, F. Reinhardt, and R. Reents, 2014 A single-step genomic model with direct estimation of marker effects. Journal of dairy science **97**: 5833–5850.

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

Meuwissen, T. H. E., T. R. Solberg, R. Shepherd, and J. A. Woolliams, 2009 A fast algorithm for bayesb type of prediction of genome-wide estimates of genetic value. Genetics Selection Evolution **41**: 1–10.

Mevik, B.-H. and R. Wehrens, 2007 The pls package: principal component and partial least squares regression in r. Journal of Statistical Software **18**: 1–23.

Misztal, I., 2016 Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics **202**: 401–409.

Misztal, I., I. Aguilar, D. Johnson, A. Legarra, S. Tsuruta, *et al.*, 2009 A unified approach to utilize phenotypic, full pedigree and genomic information for a genetic evaluation of holstein final score. Interbull Bulletin pp. 240–244.

Misztal, I. and A. Legarra, 2017 Invited review: efficient computation strategies in genomic selection. animal **11**: 731–736.

O'Hara, R. B., M. J. Sillanpää, and others, 2009 A review of bayesian variable selection methods: what, how and which. Bayesian analysis **4**: 85–117.

Onogi, A. and H. Iwata, 2016 Vigor: variational bayesian inference for genome-wide regression. Journal of Open Research Software **4**.

Park, T. and G. Casella, 2008 The bayesian lasso. Journal of the American Statistical Association **103**: 681–686.

Patterson, H. D. and R. Thompson, 1971 Recovery of inter-block information when block sizes are unequal. Biometrika **58**: 545–554.

Pérez, P. and G. de Los Campos, 2014 Genome-wide regression & prediction with the bglr statistical package. Genetics **198**: 483–495.

Pérez-Rodríguez, P., D. Gianola, J. M. González-Camacho, J. Crossa, Y. Manès, *et al.*, 2012 Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3: Genes, Genomes, Genetics **2**: 1595–1605.

R Core Team, 2019 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ridgeway, G., 2007 Generalized boosted models: A guide to the gbm package. Update **1**: 2007.

Schaeffer, L., 1986 Pseudo expectation approach to variance component estimation. Journal of Dairy Science **69**: 2884–2889.

Schulz-Streeck, T., J. O. Ogutu, and H.-P. Piepho, 2013 Comparisons of single-stage and two-stage approaches to genomic selection. Theoretical and applied genetics **126**: 69–82.

Searle, S. R., G. Casella, and C. E. McCulloch, 1992 Prediction of random variables. In *Variance Components*, pp. 367–377, John Wiley and Sons, Inc.

Sebastian, S., L. Streit, P. Stephens, J. Thompson, B. Hedges, *et al.*, 2010 Context-specific marker-assisted selection for improved grain yield in elite soybean populations. Crop science **50**: 1196–1206.

Sun, X., L. Qu, D. J. Garrick, J. C. Dekkers, and R. L. Fernando, 2012 A fast em algorithm for bayesa-like prediction of genomic breeding values. PLoS One **7**: e49157.

Taskinen, M., E. A. Mäntysaari, and I. Strandén, 2017 Single-step snp-blup with on-the-fly imputed genotypes and residual polygenic effects. Genetics Selection Evolution **49**: 1–15.

Thompson, E. and R. Shaw, 1992 Estimating polygenic models for multivariate data on large pedigrees. Genetics **131**: 971–978.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of dairy science **91**: 4414–4423.

Wang, T., Y.-P. P. Chen, M. E. Goddard, T. H. Meuwissen, K. E. Kemper, *et al.*, 2015 A computationally efficient algorithm for genomic prediction using a bayesian model. Genetics Selection Evolution **47**: 1–16.

Wimmer, V., C. Lehermeier, T. Albrecht, H.-J. Auinger, Y. Wang, *et al.*, 2013 Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics **195**: 573–587.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, *et al.*, 2016 Mixture models detect large effect qtl better than gblup and result in more accurate and persistent predictions. Journal of animal science and biotechnology **7**: 1–6.

Wright, M. N. and A. Ziegler, 2015 Ranger: a fast implementation of random forests for high dimensional data in c++ and r. arXiv preprint arXiv:1508.04409 .

Xavier, A., D. Jarquin, R. Howard, V. Ramasubramanian, J. E.

Specht, *et al.*, 2018 Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. G3: Genes, Genomes, Genetics **8**: 519–529.

Xavier, A., W. M. Muir, and K. M. Rainey, 2016 Assessing predictive properties of genome-wide selection in soybeans. G3: Genes, Genomes, Genetics **6**: 2611–2616.

Xavier, A., S. Xu, W. M. Muir, and K. M. Rainey, 2015 Nam: association studies in multiple populations. Bioinformatics **31**: 3862–3864.

Xu, S., 2007 An empirical bayes method for estimating epistatic effects of quantitative trait loci. Biometrics **63**: 513–521.

Xu, S., 2010 An expectation–maximization algorithm for the lasso estimation of quantitative trait locus effects. Heredity **105**: 483–494.

Xu, S., 2013 Mapping quantitative trait loci by controlling polygenic background effects. Genetics **195**: 1209–1222.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price, 2014 Advantages and pitfalls in the application of mixed-model association methods. Nature genetics **46**: 100–106.

Yi, N. and S. Xu, 2008 Bayesian lasso for qtl mapping. Genetics **179**: 1045–1055.

Zeng, Z.-B., T. Wang, and W. Zou, 2005 Modeling quantitative trait loci and interpretation of models. Genetics **169**: 1711–1725.

Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, *et al.*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nature genetics **42**: 355–360.

Zhou, L., R. Mrode, S. Zhang, Q. Zhang, B. Li, *et al.*, 2018 Factors affecting gebv accuracy with single-step bayesian models. Heredity **120**: 100–109.

APPENDIX 1: RCPP CODE TO IMPLEMENT FLM IN R

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
SEXP FLM(NumericVector y, NumericMatrix X){

  // Convergence settings
  int maxit = 300; double tol = 10e-8;

  // Initial settings and starting values
  int p=X.ncol(), n=X.nrow(), numit=0;
  double b0,b1,eM,Ve,cnv=1,mu=mean(y),Lmb2=0;
  NumericVector e=y-mu,Vb(p),b(p),fit(n);

  // Cross-products and shape parameter
  NumericVector xx(p),sx(p),bc(p);
  for(int k=0; k<p; k++){
    xx[k]=sum(X(_,k)*X(_,k));
    if(xx[k]==0) xx[k]=0.1;
    Lmb2=Lmb2+var(X(_,k));}
  NumericVector iTau2=p+Lmb2;

  // Looping across parameters until convergence
  while(numit<maxit){

    // Updating markers effects
    bc=b+0; for(int j=0; j<p; j++){ b0=b[j];
      b1=(sum(X(_,j)*e)+xx[j]*b0)/(iTau2(j)+xx(j));
      b[j]=b1; e=e-X(_,j)*(b1-b0);}

    // Updating intercept
    eM=mean(e); mu=mu+eM; e=e-eM;

    // Updating variance components
    Ve=sum(e*y)/(n-1);
    Vb=b*b+Ve/(xx+iTau2);
    iTau2=sqrt(Lmb2*Ve/Vb);

    // Check parameters convergence
    ++numit; cnv=sum(abs(bc-b));
    if(cnv<tol){break;}}

  // Fit model
  for(int k=0; k<n; k++){fit[k]=mu+sum(X(k,_)*b);}

  // Return output
  return List::create(Named("mu")=mu,
            Named("b")=b, Named("fit")=fit,
            Named("T2")=1/iTau2, Named("Ve")=Ve);}
```

APPENDIX 2: LAPLACE DENSITY

The fast Laplace model (FLM) is a coordinate descent type algorithm to iteratively solve a variation of the Bayesian LASSO (Park and Casella 2008) with empirical Bayesian priors. Consider the single-marker model:

$$y = m\beta + e \tag{25}$$

with a simple probabilistic structure

$$\begin{aligned} y|m,\beta &\sim N(m\beta, I\sigma_e^2) \\ \beta|\tau^2 &\sim N(0, \tau^2\sigma_e^2) \end{aligned} \tag{26}$$

Strong shrinkage can be provided through the utilization of a Laplace prior to estimation of regression coefficients.

$$p(\beta) = \prod_{j=1}^{P} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \tag{27}$$

Park and Casella (2008) proposed the double-exponential density for the regression coefficient conditional to the residual variance to ensure convergence with a unimodal posterior. The density of regression coefficients is defined as

$$p(\beta|\sigma_e^2) = \prod_{j=1}^{P} \frac{\lambda}{2\sigma_e} e^{-\lambda|\beta_j|\sigma_e^{-1}} \tag{28}$$

where λ is a scale parameter. The regression coefficient solution is given by

$$\beta|\tau^2 = \frac{m'y}{m'm + \tau^{-2}} \tag{29}$$

given the regularization that imposed by the τ^{-2} parameter shapes the regression coefficients as a Laplace distribution, as a mixture of normal with exponential mixing density. The solution of τ^{-2} has inverse-Gaussian density

$$f(x) = \sqrt{\frac{\lambda^2}{2\pi}} x^{-3/2} \exp \frac{\lambda(x - \lambda\sigma_e\sigma_\beta^{-1})}{2x\lambda^2\sigma_e^2\sigma_\beta^{-2}} \tag{30}$$

with expectation

$$E[\tau^{-2}] = \sqrt{\lambda^2\sigma_e^2\sigma_\beta^{-2}} \tag{31}$$

Where the complete-data variance of the regression coefficient (β^2) used in Park and Casella (2008) is replaced by the sample estimator (σ_β^2). The proposed algorithm utilizes the variance components (σ_β^2 and σ_e^2) as presented by Harville (1977, eq 6.3 and 6.4).

In addition, if the genomic heritability is known *a priori*, the scale parameter λ^2 (eq. 7) can incorporate such information and be estimated as:

$$\lambda^2|h^2 = \frac{1-h^2}{h^2} \sum_{j=1}^{p} \sigma_{m_j}^2 \tag{32}$$

APPENDIX 3: AN APPROXIMATION OF VARIANCES

The approximation provided by equation (21) was first derived by Schaeffer (1986) using the pseudo-expectation framework. Below a likelihood-based derivation is presented. Consider the following model with fixed effects and R random effects:

$$y = Xb + \epsilon = Xb + \sum_{i=1}^{R} Z_i u_i + e \tag{33}$$

Where the model variance is defined as:

$$E[\epsilon\epsilon'] = V = \sum_{i=0}^{R} V_i = \sum_{i=1}^{R} Z_i Z_i' \sigma_i^2 + I\sigma_e^2 \tag{34}$$

The likelihood function variance components aim to minimize is:

$$(y - Xb)'V^{-1}(y - Xb) + log|V| \tag{35}$$

Marginally ($\partial L/\partial\sigma_i^2$), each ϵ_i with variance V_i, derives:

$$(y - Xb)'V^{-1}V_iV^{-1}(y - Xb) - tr(V^{-1}V_i) \tag{36}$$

Where $V_i(\sigma_u^2) = ZZ'$ and $V_i(\sigma_e^2) = I$.

Searle *et al.* (1992) showed that the REML is a slight variation that accounts for the estimation error of fixed effects:

$$(y - Xb)'V^{-1}V_iV^{-1}(y - Xb) - tr(SV^{-1}V_i) \tag{37}$$

Where $S = I - X(X'X)^{-1}X'$. This equates into ($\partial L/\partial\sigma_i^2 = 0$):

$$\begin{aligned} 0 &= (y - Xb)'V^{-1}V_iV^{-1}(y - Xb) - tr(SV^{-1}V_i) \\ tr(SV^{-1}V_i) &= (y - Xb)'V^{-1}V_iV^{-1}(y - Xb) \\ tr(SV^{-1}V_i) &= (y - Xb)'V^{-1}\epsilon_i\sigma_{\epsilon_i}^{-2} \\ tr(SV^{-1}V_i)\sigma_{\epsilon_i}^2 &= (y - Xb)'V^{-1}\epsilon_i \end{aligned} \tag{38}$$

The approximation that yields the same as Schaeffer (1986) is based on removing V^{-1} from both sides of the equations. Thus:

$$tr(SV_i)\sigma_{\epsilon_i}^{-2} = (y - Xb)'\epsilon_i \tag{39}$$

From this solution, the variance component estimator for the i^{th} random term ($\epsilon_i = Z_iu_i$) and residuals ($\epsilon_i = e$) are:

$$\sigma_i^2 = \frac{(y - Xb)'Z_iu_i}{tr(Z_iSZ_i')} \tag{40}$$

$$\sigma_e^2 = \frac{(y - Xb)'e}{n - r_X} \tag{41}$$

For computational convenience, equations (21) provides the equation denominator $q_i \sum \sigma_{z_i}^2$ as a replacement for $tr(Z_iSZ_i')$.