Euphytica (2022) 218:54
https://doi.org/10.1007/s10681-022-03006-y

psBLUP: incorporating marker proximity for improving genomic prediction accuracy

Georgios Bartzis · Carel F. W. Peeters · Fred van Eeuwijk

Received: 13 December 2021 / Accepted: 16 March 2022

Abstract Genomic selection entails the estimation of phenotypic traits of interest for plants without phenotype based on the association between single-nucleotide polymorphisms (SNPs) and phenotypic traits for plants with phenotype. Typically, the number of SNPs far exceeds the number of samples (high-dimensionality) and, therefore, usage of regularization methods is common. The most common approach to estimate marker-trait associations uses the genomic best linear unbiased predictor (GBLUP) method, where a mixed model is fitted to the data. GBLUP has also been alternatively parameterized as a ridge regression model (RRBLUP). GBLUP/RRBLUP is based on the assumption of independence between predictor variables. However, it is to be expected that variables will be associated due to their genetic proximity. Here, we propose a regularized linear model (namely psBLUP: proximity smoothed BLUP) that explicitly models the dependence between predictor effects. We show that psBLUP can improve accuracy compared to the standard methods on both Arabidopsis thaliana data and Barley data.

Keywords BLUP · Genomic selection · High-dimensional data · Proximity smoothing

G. Bartzis · C. F. W. Peeters (✉) · F. v. Eeuwijk
Mathematical and Statistical Methods group (Biometris),
Wageningen University and Research, Wageningen,
The Netherlands
e-mail: carel.peeters@wur.nl

Introduction

Genomic selection is a tool applied in animal and plant sciences for improving quantitative traits (Heffner et al. 2009; Hayes et al. 2009; Jannink et al. 2010; Goddard et al. 2010; Van Binsbergen et al. 2015). Genomic values of line performance measuring the genetic merit of lines are calculated using markers (e.g., single nucleoteid polymorphisms; *SNPs*) covering the whole genome (Hayes et al. 2009). By using high density SNP panels, it is expected that SNPs in linkage disequilibrium (*LD*) with quantitative trait loci (*QTLs*) contributing to the phenotypic variation (Hayes et al. 2009; Zeng et al. 2018a) are included.

A training panel that has been both genotyped and phenotyped is used to build a prediction model describing a marker-trait relationship. A common approach to do so is by regressing phenotypes on all available markers using a linear model (de Los Campos et al. 2013). With the prediction model, phenotypic values for non-phenotyped plant genotypes are predicted, which are subsequently used for selection (Hunt et al. 2018).

The first attempts to incorporate and simultaneously estimate SNP effects to predict phenotypic values were made by Bernardo (1994), Bernardo (1996). These have been popularized by Whittaker et al. (2000) and Meuwissen et al. (2001) and have been repeatedly used in plant and animal breeding (Bernardo 2008; VanRaden 2008; Crossa et al. 2010). However, the availability of high-density SNP panels,

Published online: 08 April 2022

where the number of markers (p) exceeds the sample size (n), implies that regularization methods are required in order to estimate all effects.

Common regularization approaches

The most common approach is by using the genomic best linear unbiased predictor (*GBLUP*) method, where a mixed model is fitted to the data with the marker effects as random (normally and independently distributed effects with a common variance) (VanRaden 2008; de Los Campos et al. 2009). GBLUP has also been alternatively parameterized as a ridge regression (Hoerl and Kennard 1970) model (referred to as *RRBLUP*) for genomic prediction (Piepho et al. 2012). Therefore, the level of SNP effect shrinkage can be determined with either a grid search over the regularization parameter for RRBLUP, or by using the ratio of variance components in GBLUP (Heslot et al. 2012). Finally, RRBLUP can also be parameterized in a Bayesian setting with a Gaussian prior for the marker effects (de Los Campos et al. 2013). We will use RRBLUP and GBLUP interchangeably in this work.

RRBLUP assumes that all SNP effects have equal variance, an assumption that has often been criticized, since both causal and non-causal SNPs receive the same amount of regularization. Contrarily, most of the SNPs in the genome are assumed to contribute little to the phenotype and therefore should be penalized more (Shen et al. 2013). By assuming that SNP effects have different distributions, additional flexibility is added to the BLUP model. One such approach is *MultiBLUP* and *Adaptive MultiBLUP* (Speed and Balding 2014) assigning different distributions to the effects, based on prior information or data-driven approaches. In these approaches, markers are assigned to groups with different variances expressing whether the markers have large or zero to small contribution to the phenotypic variance. Each group of markers forms a separate genomic relationship matrix.

Another encompassing approach to regularization is by assigning certain prior densities to the marker effects in the Bayesian setting. Using a t-density (which puts more mass at zero and has thicker tails relative to the Gaussian density), for example, implies that small effects receive stronger shrinkage towards zero than strong effects. This approach is colloquially known as *BayesA* (Meuwissen et al. 2001). *BayesB* (Meuwissen et al. 2001) and *BayesC* (Habier et al. 2011) are obtained by assuming that SNP effects are a mixture of a point-mass at zero and a (diffuse) distribution on some finite interval. BayesB uses a t-density as the slab, while BayesC uses a normal density. Both induce a combination of variable selection and shrinkage (de Los Campos et al. 2013). Empirical studies show only small differences between GBLUP, BayesA, BayesB, and BayesC, with variable selection methods having better performance in scenarios with large-effect QTLs. When the number of SNPs is small, no difference in performance is observed (de Los Campos et al. 2013).

All aforementioned methods are based on the assumption of independence between SNP effects. Nonetheless, it is anticipated that SNPs will be correlated due to spatial proximity within the chromosomes (Gianola et al. 2003). For modeling the correlation between the effects *ante-BayesA*, *ante-BayesB*, and *BayesN* have been proposed (Yang and Tempelman 2012; Zeng et al. 2018b). In these approaches the effect of a SNP is estimated with respect to the relative physical distance of its preceding neighbour, i.e., they have a distance-specific *ante-dependence* parameter (Núñez-Antón and Zimmerman 2009). While these are very interesting Bayesian approaches dealing with the spatial proximity of the SNPs, they involve Markov Chain Monte Carlo methods, which become computationally prohibitive for models involving many variables. We offer a simpler alternative method based on penalized regression to account for the spatial proximity.

Contribution

In this article we propose, motivated by the network constrained regularization and variable selection (Li and Li 2008), a regularized linear model: the proximity smoothed BLUP (psBLUP). Li and Li (2008) use a combination of L_1 (Lasso) and L_2 (ridge) penalties. The former is used for variable selection, the latter for encouraging smoothness on neighboring marker effects. psBLUP uses an L_2 instead of an L_1-norm on the coefficients (like RRBLUP), while similarly to Li and Li (2008) it imposes a second L_2-norm to encourage smoothness on neighboring effects. psBLUP explicitly accounts for the dependence between marker effects due to the SNPs' relative spatial proximity within chromosomes. A smooth solution on the differences between adjacent marker effects is employed, since it is expected that neighboring markers are in LD with the same QTLs. One feature of the method is that we do not require a strict definition of the markers' proximity, which can be estimated from the data. For example, the correlation coefficient between markers can be used as a measure of LD (Zaykin et al. 2008). In our applications, we use the squared correlation coefficient for those SNP pairs being equal or less than 10 centimorgan (cM) apart as a measure of proximity and observe that it is sufficient to outperform RRBLUP in terms of accuracy.

Our intention is to present a genomic prediction method that improves the accuracy of the traditional ridge penalty on marker effects in RRBLUP / GBLUP by using additional spatial information on marker locations and forcing marker effects to be more similar when the marker locations are closer. We expect this method to be suitable for genomic prediction of unphenotyped genotypes in homogeneous plant families (F2, RIL, MAGIC) for phenotypic traits with a low genetic signal to noise ratio in combination with a small training set of genotypes (< 100). For homogeneous plant families, a few hundred markers suffice for genomic prediction because linkage disequilibrium extends far (10-20 cM). We did not evaluate our method for diversity panels with fast linkage disequilibrium decay. Computational requirements would be substantial in that case and need further study. For the current applications to homogeneous plant families, we present mixed model implementations in theory and software.

Overview

The remainder is organized as follows. In Sect. 2, we review RRBLUP and propose the psBLUP as a way of incorporating information on the SNPs proximity in genomic prediction. This section also introduces the data with which the two methods (RRBLUP vs psBLUP) are compared in terms of predictive ability: Arabidopsis thaliana data coming from the Seed Lab of Wageningen University and Research, and Barley data from the North American Barley Genome Mapping Project (NABGMP). In Sect. 3 we demonstrate our approach on these two applications and show that psBLUP can lead to a gain in accuracy. We conclude in Sect. 4 by discussing possible extensions for computational efficiency and the advantages of the method in settings with limited sample sizes or low heritability phenotypes.

Materials and methods

Phenotyped and genotyped datasets

Population 1: Arabidopsis thaliana data from Wageningen

The first population is a Recombinant Inbred Line (RIL) population created from a cross between two natural Arabidopsis accessions, i.e., Bayreuth (*Bay-0*) and Shahdara (*Sha*). The data come from the Seed Lab of Wageningen University and Research (Netherlands). Seeds of 164 RILs were divided into four subpopulations (41 lines each) representing four important developmental stages of seed germination. The concentration levels of 161 metabolites were determined for all 164 lines. Finally, 64 metabolites were retained to be used for further analysis as phenotypes. Concentration levels of the metabolites were log -transformed and adjusted for the four developmental seed stages by subtracting the mean levels from each group. Finally, information on $p = 1059$ markers (5 chromosomes) was available. More information on the study design and data can be found in Joosen (2013) and Joosen et al. (2013).

Population 2: Barley data from NABGMP

The second population concerns the well-known *Steptoe* × *Morex* doubled haploid (DH) population developed by the NABGMP (https://wheat.pw.usda.gov/ggpages/SxM/). This DH population was developed between 1991 and 1992 at several locations in North America. It consists of $n = 150$ DH lines of Barley that were evaluated in different environments. We retained five traits for further analysis, i.e., yield (measured in 16 environments), percentage of grain protein (measured in 9 environments), percentage of malt extract (measured in 9 environments), line's height (measured in 16 environments), and the degree of α-amylase activity (measured in 9 environments). A total of 148 lines were genetically characterized by $p = 794$ markers covering the seven barley chromosomes. More information on the study design and data can be found in Hayes et al. (1993) and Malosetti et al. (2004).

Methods for genomic prediction

Let, for n samples, $\boldsymbol{y} = [y_1, \dots, y_n]^\top$ be a $n \times 1$ centered response vector representing a phenotype of interest ($\sum_i y_i = 0$). Also, let $\boldsymbol{X}$ be a $n \times p$ matrix containing scaled SNPs ($\sum_i x_{ij} = 0$, $\sum_i x_{ij}^2 = n$ for all $j = 1, \dots, p$). In order to build a genomic prediction model and establish a genotype-phenotype relationship, a vector of SNP effects needs to be estimated. We first present the standard RRBLUP model, before extending to psBLUP.

RRBLUP

In RRBLUP the vector of SNP effects is obtained by minimizing the penalized least squares with respect to $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{RR} := \arg\min{}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X\beta})^\top (\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda_1 \boldsymbol{\beta}^\top \boldsymbol{I}_p \boldsymbol{\beta} \right\}, \tag{1}$$

where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix and where $\lambda_1 \geq 0$ represents the shrinkage parameter controlling the amount of regularization. Since $\hat{\boldsymbol{\beta}}$ depends on λ_1, a cross-validation criterion is typically used to select λ_1 from a grid of possible values.

Another way to select λ_1 is by estimating the variance components of a mixed model with SNP effects as random, since the two models are equivalent (Habier et al. 2007; Piepho et al. 2012; de Los Campos et al. 2013; de Vlaming and Groenen 2015). The linear mixed model can be written as:

$$\boldsymbol{y} = \boldsymbol{Xu} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\varepsilon}$ are the residuals distributed as $N(0, \sigma_\varepsilon^2 \boldsymbol{I}_n)$ and $\boldsymbol{u}$ are the random effects distributed as $N(0, \sigma_{\boldsymbol{u}}^2 \boldsymbol{I}_p)$. The ridge regression model with $\lambda_1 = \sigma_\varepsilon^2 / \sigma_{\boldsymbol{u}}^2$ gives the same estimated SNP effects as (2) (i.e., $\hat{\boldsymbol{\beta}}_{RR} = \hat{\boldsymbol{u}}$). Selecting λ_1 and calculating the SNP effects based on the mixed model is often preferred due to its computational efficiency (Clark and van der Werf 2013).

SNP proximity matrix

Before presenting the penalized least squares for obtaining psBLUPs, we briefly introduce the proximity between the SNPs, represented as a matrix. Let $\boldsymbol{W}$ be a matrix containing information on the spatial relationship between SNPs. For example, the matrix element $w_{jj'}$ could contain the LD between the jth and j'th SNPs or the relative (physical/genetic) distance between them. Here, $\boldsymbol{W}$ is calculated using the square of markers' pairwise Pearson correlation coefficient (VanLiere and Rosenberg 2008) if they are close. We deem markers whose genetic distance is equal or less than 10cM to be close. A genetic distance of 10cM concurs with a recombination rate of at most .1 (Hartl 2011) which translates to a Pearson correlation of at least .6 (Warrens 2008). Let j and j' be two SNP indices, let g_j and $g_{j'}$ be the physical/genetic position of the two corresponding SNPs on the chromosome, and let $\boldsymbol{x}_j$ and $\boldsymbol{x}_{j'}$ be two vectors containing genetic information on n samples for those SNPs. The matrix element $w_{jj'}$ is then defined as:

$$w_{jj'} = w_{j'j} = \begin{cases} \rho(\boldsymbol{x}_j, \boldsymbol{x}_{j'})^2, & \text{if } |g_j - g_{j'}| \leq 10 \text{ cM}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $\rho(\boldsymbol{x}_j, \boldsymbol{x}_{j'})$ is the Pearson correlation between SNPs j and j'. By that definition, each SNP can be viewed as the center of a local network of SNPs, and is connected to SNPs up to 10cM away. Essentially, for these connections, the squared correlation coefficient is calculated.

Figure 1 contains a toy example illustrating how chromosomal spatial information is translated to network information that is explicitly used in psBLUP. On the top panel (chromosomal representation), six SNPs are marked on a segment of a chromosome. The distances between SNPs equal or less than 10cM have been shown with dashed lines. On the center panel, the same SNPs are represented as nodes in a network where an edge is connecting a pair of SNPs if their distance is less than or equal to 10cM. The width of the edges is analogous to the proximity between two SNPs. Finally, the network is represented as a matrix (bottom panel), where the similarity between connected SNP pairs is coded in grey-colored circles. A darker color indicates a stronger similarity. Empty cells imply that the distance between two SNPs is larger than 10cM and they do not share a connection in the network representation.

To estimate the SNP effects using psBLUP we need to calculate the normalized Laplacian matrix $\boldsymbol{L}$ (Chung and Graham 1997) of $\boldsymbol{W}$ with elements:

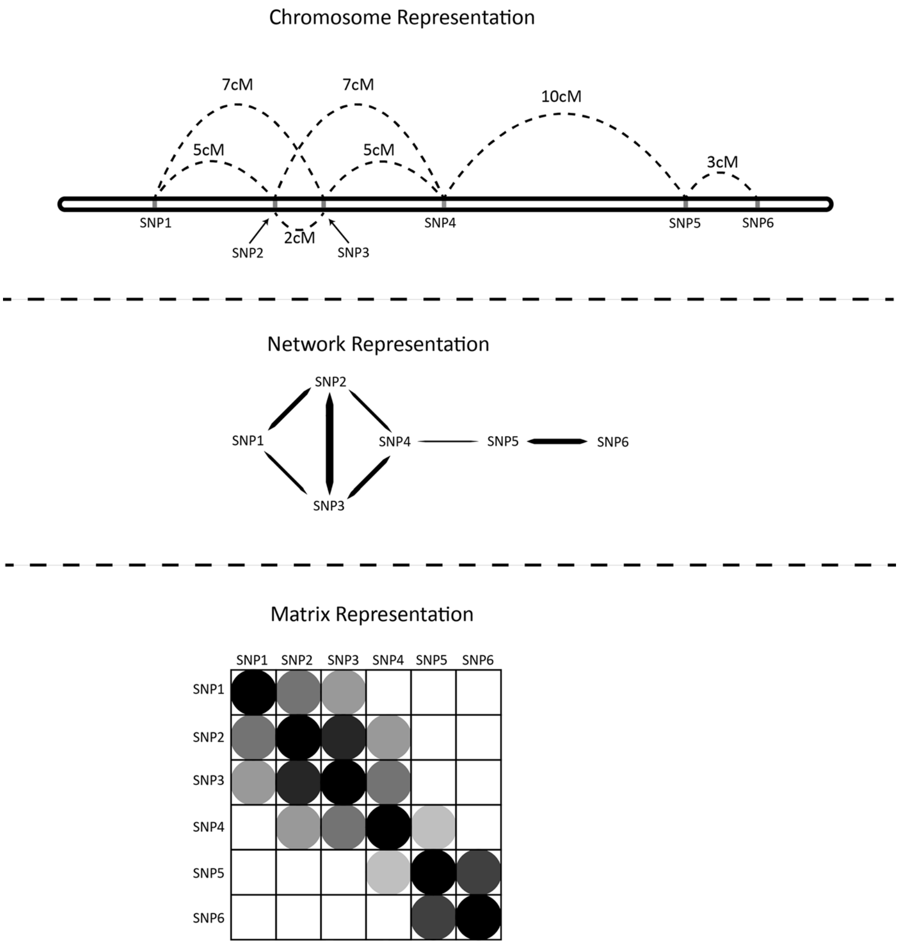


Fig. 1 **Chromosomal representation**: six SNPs are marked on a part of a chromosome. Dashed lines indicate the distance between pairs of SNPs. **Network representation**: the six SNPs are represented as nodes in a network with edges connecting only SNPs with distance equal or less than 10cM. SNPs proximity is encoded as edges' width (SNPs with low distance have wider edges). **Matrix representation**: the similarity of all pairs of SNPs is coded using grey-colored circles. Higher similarity is encoded with darker grey. Empty cells indicate that a pair of SNPs does not share an edge in the network representation

$$l_{jj'} = l_{j'j} = \begin{cases} 1 - w_{jj'}/s_j, & \text{if } j = j' \text{ and } s_j \neq 0, \\ -w_{jj'}/\sqrt{s_j s_{j'}}, & \text{if } j \neq j' \text{ and } w_{jj'} \neq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $s_j = \sum_{j'} w_{jj'}$ is the weighted total connectivity of SNP j.

psBLUP

The SNP effects are obtained by minimizing the proximity-penalized least squares with respect to $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{ps} := \arg\min{}_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X\beta})^{\top}(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda_1 \boldsymbol{\beta}^{\top}\boldsymbol{I}_p\boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^{\top}\boldsymbol{L\beta} \right\}, \tag{5}$$

where $\boldsymbol{L}$ is the normalized Laplacian matrix obtained with expression 4 and $\lambda_2 \geq 0$ is the parameter inducing shrinkage on the differences between SNP effects analogous to their proximity. Finally, as in expression 1, the term $\boldsymbol{\beta}^{\top}\boldsymbol{I}_p\boldsymbol{\beta}$ is the L_2-norm shrinking the SNP coefficients.

The term $\boldsymbol{\beta}^{\top}\boldsymbol{L\beta}$ can also be written as (Li and Li 2008):

$$\boldsymbol{\beta}^{\top}\boldsymbol{L\beta} = \sum_{j=1}^{p}\sum_{j'=1}^{p} \left(\frac{\beta_j}{\sqrt{s_j}} - \frac{\beta_{j'}}{\sqrt{s_{j'}}} \right)^2 w_{jj'}. \tag{6}$$

This implies that the psBLUPs are smoothed by penalizing the sum of weighted squares of the differences between them. Therefore, when SNPs j and j' are close on the chromosome, they are expected to have almost equivalent association to $\boldsymbol{y}$ and thus similar effects, translating in a small difference in coefficients.

Solving psBLUP

Following Zou and Hastie (2005) and Li and Li (2008), we reduce the problem in (5) to a ridge regression using the augmented data solution. Let, $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\top}$ be the eigendecomposition of the $p \times p$ normalized Laplacian matrix $\boldsymbol{L}$, with $\boldsymbol{Q}$ the $p \times p$ matrix of eigenvectors and $\boldsymbol{\Lambda}$ the diagonal matrix with the eigenvalues. Define $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}$, $\gamma = \lambda_1/\sqrt{1+\lambda_2}$, and $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$. The new $(n+p)$-dimensional vector of responses $\boldsymbol{y}^*_{(n+p)}$ and $(n+p) \times p$ matrix of predictors $\boldsymbol{X}^*_{(n+p)\times p}$ are then defined as:

$$\boldsymbol{y}^* = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}, \qquad \boldsymbol{X}^* = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{T}^{\top} \end{pmatrix}.$$

Using $\boldsymbol{y}^*$ and $\boldsymbol{X}^*$, expression (5) is rewritten as:

$$\hat{\boldsymbol{\beta}}^*_{ps} := \arg\min_{\boldsymbol{\beta}^*} \left\{ (\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)^{\top}(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*) + \gamma \boldsymbol{\beta}^{*\top}\boldsymbol{I}_p\boldsymbol{\beta}^* \right\}, \tag{7}$$

which is a conventional ridge regression model in the augmented data $\boldsymbol{y}^*$ and $\boldsymbol{X}^*$.

Fitting a mixed model is less computationally demanding than the search for an optimal penalty-value for ridge regression. We select the psBLUPs and the regularization parameter γ using the following model:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{u}^* + \boldsymbol{\varepsilon}^* \tag{8}$$

where $\boldsymbol{\varepsilon}^*$ is the vector of residuals distributed as $N(0, \sigma^2_{\boldsymbol{\varepsilon}^*}\boldsymbol{I}_{(p+n)})$ and $\boldsymbol{u}^*$ is distributed as $N(0, \sigma^2_{\boldsymbol{u}^*}\boldsymbol{I}_p)$. As the accuracy in terms of correlation is not sensitive to its value, λ_2 was assessed along a crude grid of equidistant values (ranging from 1 to 75). Finally, $\gamma = \sigma^2_{\boldsymbol{\varepsilon}^*}/\sigma^2_{\boldsymbol{u}^*}$ and therefore, $\lambda_1 = (\sqrt{1+\lambda_2})\sigma^2_{\boldsymbol{\varepsilon}^*}/\sigma^2_{\boldsymbol{u}^*}$. Fitting a ridge regression model was done by using the augmented design matrix as input to the *rrBLUP* R-package (Endelman 2011). The solution to (5) is then obtained as $\hat{\boldsymbol{\beta}}_{ps} = (1+\lambda_2)^{-1/2}\hat{\boldsymbol{\beta}}^*_{ps}$.

Evaluation

We evaluate RRBLUP and psBLUP using the following approach. We split the data in training and test sets based on three scenarios:

(1) Use 25% of the data for training and 75% for testing,
(2) Use 50% of the data for training and 50% for testing,
(3) Use 75% of the data for training and 25% for testing.

For each case, RRBLUPs and psBLUPs are estimated. The correlation between the fitted and observed values is used to assess the accuracy of each method. We repeat the process 100 times for computing a mean gain/loss of psBLUP compared to RRBLUP. For each iteration, we calculate the difference in accuracy between psBLUP and RRBLUP. Then, the mean accuracy gain/loss is calculated as the average of the accuracy difference, over the 100 runs.

The selection of scenarios is justified as follows: by using 25-75 training-test split, we investigate how good the model performs when there is little information for estimating SNP-phenotypic relationships, and how in such cases having proximity information can help improve accuracy when generalizing to a much larger population. Inversely, selecting a 75-25 training-test split can show two things: (i) that when having more power and most SNP-phenotypic relationship is explained, spatial information may not add information; (ii) nevertheless, if the sample size is still not an important aspect because studying low heritability traits, spatial information on SNPs can still improve accuracy. Finally, the 50-50 training-test split uses the same number of samples for training and testing.

Results

Application 1: Wageningen Arabidopsis thaliana data

Here, we want to assess the gain in predictive accuracy when using information on the spatial proximity of the markers, by comparing psBLUP to RRBLUP for 64 metabolites. The markers' proximity was measured using expression (3).

The mean accuracy, for each of the three (sample size) scenarios and for each of the two models, was determined as the mean correlation coefficient across all 100 realizations between the predicted genotypic values and observed phenotypes of the test data. A summary of the results is presented in Fig. 2 for the scenario using 50% of the data for training and the rest for testing (the results for all scenarios can

Table 1 The predictive ability of RRBLUP vs psBLUP together with their observed difference using the Arabidopsis metabolite data from Wageningen University Seed Lab. psBLUP and RRBLUP were fitted 100 times under random subsampling for different scenarios: (i) 25% of the samples used for training and 75% for testing, (ii) 50% of the samples used for training and 50% for testing, and (iii) 75% of the samples used for training and 25% for testing. The accuracy is calculated over all iterations of the process. The parentheses contain the 5*th* and 95*th* percentile of the point estimate

Training set	RRBLUP accuracy	psBLUP accuracy	Gain in accuracy	% of times psBLUP > RRBLUP
25%	20.99%	24.45%	3.46%	86.6%
	(3.06, 57.84)	(6.56, 60.77)	(1.50, 6.85)	
50%	26.49%	29.40%	2.91%	86.3%
	(4.61, 65.20)	(9.08, 66.39)	(0.77, 6.98)	
75%	29.55%	33.09%	3.54%	86.9%
	(5.00, 69.47)	(12.61, 70.21)	(1.17, 8.41)	
Mean	25.68%	28.98%	3.30%	86.6%
	(4.16, 66.10)	(7.93, 67.43)	(1.87, 7.17)	

be found in the Supplementary Material). It can be seen that on average, psBLUP gives higher accuracy than RRBLUP, since the gain in accuracy is positive. The mean difference between psBLUP and RRBLUP was 3.3%. The results have also been summarized in Table 1.

In Fig. 2 we observe that the differences in predictive ability between psBLUP and RRBLUP are consistent. Results indicate that phenotypic information is contained within markers' correlation structure, since using information on the proximity between them yields improved accuracy. In Table 1, the accuracy using RRBLUP and psBLUP has been summarized together with the estimated gain (the 5*th* and 95*th* percentile is displayed in the parentheses). In both cases (RRBLUP and psBLUP), the accuracy increases with larger training sample sizes, as expected. The gain in accuracy when using psBLUP ranges for 2.91% to 3.54% in all training set scenarios. In the last column of Table 1 we see that psBLUP yields superior accuracy from RRBLUP in more than 86% of the cases for any scenario.

Interestingly, when the predictive accuracy using RRBLUP is high, the gain using psBLUP is small. Inversely, the gain using marker proximity is higher when the genomic prediction model is not so

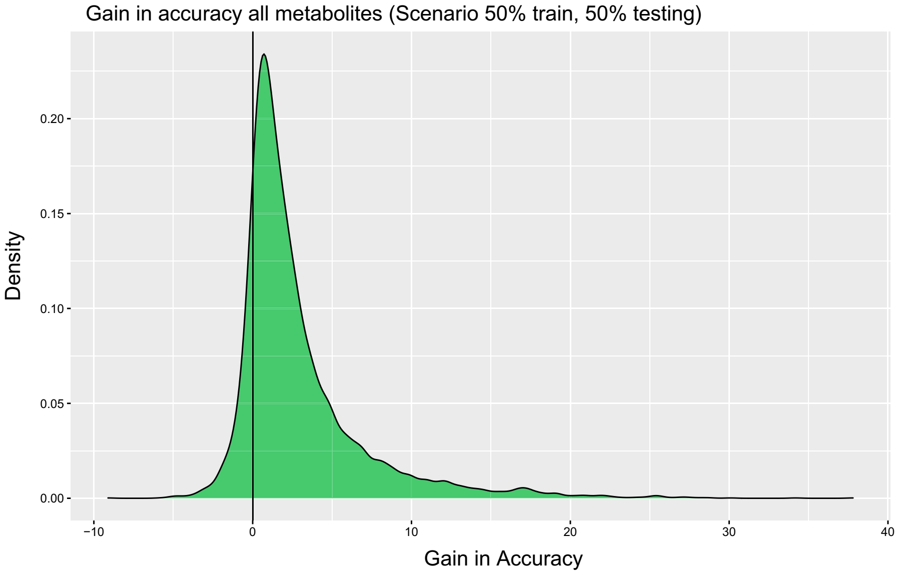


Fig. 2 The gain in accuracy when using psBLUP vs RRBLUP for 64 metabolites. The x-axis is expressed in percentages. For every metabolite, psBLUP and RRBLUP was fitted 100 times by randomly sub-sampling 50% of the samples to be used for training the models and 50% for testing

informative. This result has been visualized in Fig. 3. Each dot represents the mean accuracy using RRBLUP and mean gain in accuracy when psBLUP is used, over 100 runs. For metabolites with high predictive accuracy using RRBLUP, the gain in psBLUP is small, while the highest gains using psBLUP have been observed for metabolites with very low predictive accuracy using RRBLUP. We will return to this observation in the discussion.

Application 2: NABGMP barley data

In this application we assess the gain in predictive accuracy when using information on the spatial proximity of the markers, by comparing psBLUP to RRBLUP for 59 trait-environmental combinations (Barley data from NABGMP). The markers proximity was measured using expression (3).

As in the first application, the mean accuracy of the models was determined using the mean correlation coefficient between the predicted and observed phenotypes of the test data for each of the three (sample size) scenarios over 100 runs. A summary of the results is presented in Fig. 4 for the scenario with half the samples used for training and the rest for testing. The results have also been summarized in Table 2.

In Fig. 4 we see that the mean difference in predictive ability between psBLUP and RRBLUP is positive in some cases. In Table 2 the results have also been summarized. Across all traits, the accuracy increases for larger sample sizes using either genomic prediction method (RRBLUP or psBLUP). The 5*th* and 95*th* percentiles are displayed in the parentheses for each trait-subsampling scenario. In the last column of Table 2 the percentage of times psBLUP yields greater accuracy than RRBLUP is shown.

As in the metabolite data application, the gain in predictive accuracy is greater when the accuracy using RRBLUP is lower. The scenario with 50% of the data used as training and the rest as testing (for all five phenotypes) has been visualized in Fig. 5 were a downward trend can be seen. Each dot shows the mean RRBLUP accuracy and gain in accuracy when using psBLUP over 100 runs. With regard to the traits, we see that plant height has overall the highest accuracy using RRBLUP and subsequently the lowest gain when using psBLUP. The scenarios using a 25-75 and 75-25 split for training and testing can be found in the Supplementary Material.

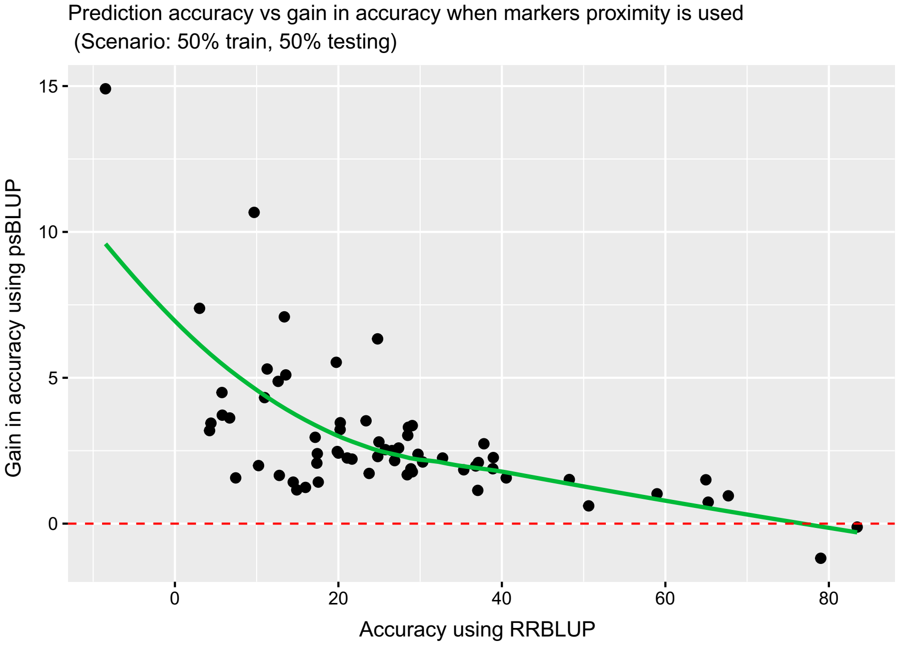


Fig. 3 Prediction accuracy vs gain in accuracy for the 64 metabolites used (points in the plot) when markers proximity is used. The y and x-axes are expressed in percentages. The y-axis shows the percentage gain in prediction accuracy when psBLUP is used instead of RRBLUP. The x-axis shows the percentage accuracy for a metabolite. Each dot represents the mean accuracy using RRBLUP and mean gain in accuracy when psBLUP is used, over 100 runs

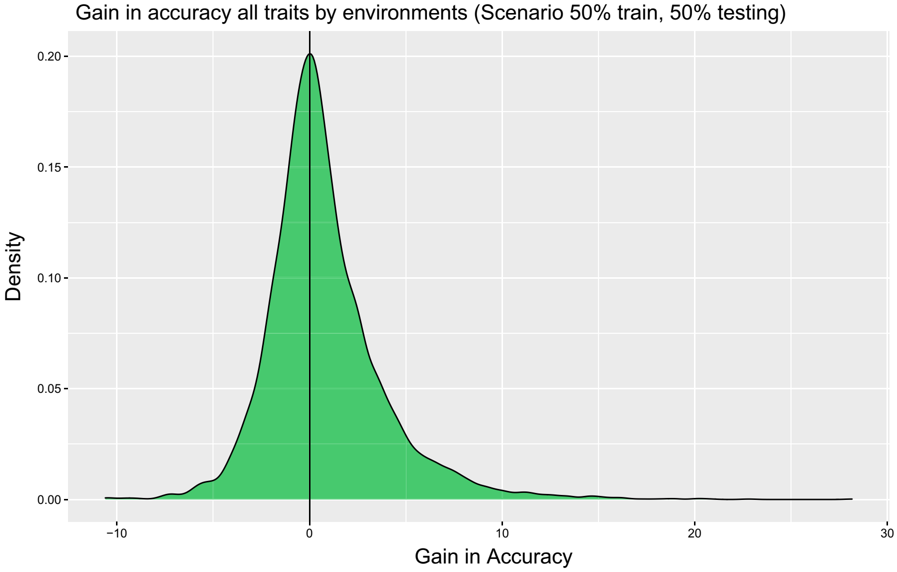


Fig. 4 For every combination of the 59 trait-environments of the NABGMP dataset, 100 psBLUP and RRBLUP models were fitted when using 50% of the data for training and the rest for testing. The x-axis is expressed in percentages and shows the absolute difference in accuracies between the best selected psBLUP model and RRBLUP

Discussion

In this work, we developed a regularized regression model that uses information on the proximity of the explanatory variables in order to increase prediction accuracy. Our model (psBLUP) was used in the context of genomic prediction as an extension of RRBLUP: the spatial proximity between the SNPs was used to improve the predictive ability of RRBLUP. When no penalty is used to account for the dependence between SNP effects, the two methods should be identical by definition.

For demonstrating the proposed approach two applications were considered. In the first application, the data were part of a RIL population of 164 lines with 1059 SNPs, and 64 metabolites. In the second application, the data were part of the Steptoe × Morex DH barley population having 148 lines characterized by 794 SNPs. In both applications we utilized SNP information in order to build a prediction model for the responses, using psBLUP and RRBLUP. The two methods were compared with regard to their prediction accuracy. The gain using marker proximity is highest when the standard genomic prediction model is not so informative.

A few things can be noted for the inverse relationship between accuracy gain and training sample size, i.e., greater gain for smaller training sample sizes. In cases were the training sample size is small, the accuracy of the RRBLUP model is expected to be low. Therefore, the variation margin that can be explained by the SNPs' spatial proximity (psBLUP) is high. Modeling the spatial proximity/accounting for correlation between SNP effects is therefore more important for low heritability and smaller training sets.

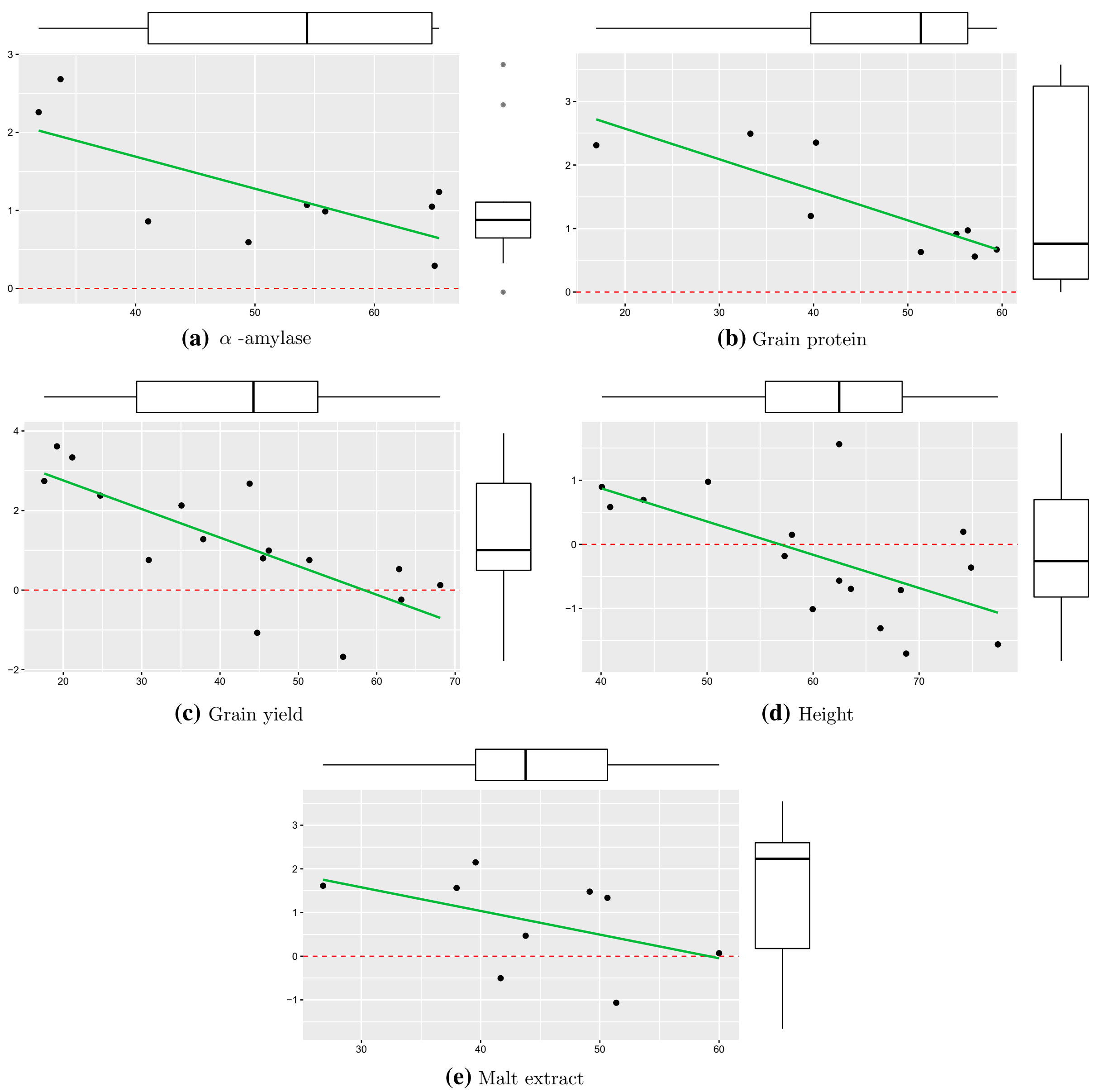
We note that in some cases (e.g., association panel) neighboring markers can have effects with opposite signs. Then they will wrongly tend to cancel out, leading to smaller overall accuracy. In that case, all predictors can be recoded to be positively associated with the response prior to model fitting. Alternatively, the squared scaled absolute differences between the SNP coefficients could be penalized in expression (6).

An advantage of the psBLUP approach is the broad applicability, since it can be used for any continuous outcome and type of predictor variables. Additionally, it can be implemented using standard statistical software that can fit a mixed model, making it easily accessible. Moreover, there is no strict definition for the markers spatial proximity, which can be estimated by the data or by using prior information making the data analysis more flexible.

Some issues still need to be addressed. We utilized the mixed model equivalence to ridge regression for reducing the model tuning to the evaluation of parameters that can be obtained with a single optimization. Even though the speed is greatly improved by solving the mixed model equations on the augmented data, the efficiency needs to be further improved for incorporating high density SNP panels. For estimating

Table 2 The predictive ability of RRBLUP vs psBLUP together with their observed difference when using the DH barley data from NABGMP. psBLUP and RRBLUP were fitted 100 times under random subsampling for 3 scenarios: (i) 25% of the samples used for training and 75% for testing, (ii) 50% of the samples used for training and 50% for testing, and (iii) 75% of the samples used for training and 25% for testing. The accuracy is calculated over all iterations of the process. The parentheses contain the 5*th* and 95*th* percentile of the point estimate

	Training set	RRBLUP accuracy	psBLUP accuracy	Gain in accuracy	% of times psBLUP > RRBLUP
Grain Yield	25%	34.47%	36.93%	2.46%	69.9%
		(11.5, 59.57)	(16.51, 61.6)	(0.03, 5.55)	
	50%	41.75%	42.94%	1.19%	57.4%
		(18.79, 64.38)	(22.19, 64.59)	(-1.23, 3.4)	
	75%	45.38%	46.54%	1.16%	54.2%
		(21.56, 69.35)	(26.67, 69.98)	(-1.93, 5.38)	
Mean		40.53%	42.14%	1.61%	
		(15.68, 66.87)	(19.34, 67.22)	(-1.47, 5.06)	
Grain Protein	25%	38.17%	40.66%	2.49%	82.2%
		(18.28, 51.15)	(21.06, 53.29)	(1.98, 2.92)	
	50%	45.51%	46.86%	1.35%	65.8%
		(23.48, 58.5)	(25.86, 59.13)	(0.59, 2.44)	
	75%	49.54%	51.04%	1.50%	60.8%
		(27.81, 65.86)	(30.4, 66.41)	(0.33, 2.89)	
Mean		44.41%	46.19%	1.78%	
		(18.27, 62.8)	(20.67, 63.11)	(0.57, 2.94)	
Malt extract	25%	36.77%	38.90%	2.13%	73.4%
		(24.76, 48.13)	(26.94, 50.06)	(1.22, 2.85)	
	50%	44.55%	45.34%	0.79%	60.2%
		(31.26, 56.55)	(32.85, 56.83)	(-0.84, 1.93)	
	75%	47.53%	49.12%	1.59%	61.7%
		(34.06, 61.56)	(36.45, 62.22)	(-0.15,3.01)	
Mean		42.95%	44.45%	1.50%	
		(27.63, 58.88)	(29.53, 59.49)	(-0.41, 2.99)	
Height	25%	51.81%	52.98%	1.17%	58.2%
		(32.91, 68.09)	(35.91, 68.38)	(-0.49, 3.07)	
	50%	60.54%	60.35%	-0.19%	39.5%
		(40.66, 75.53)	(41.32, 74.87)	(-1.6, 1.12)	
	75%	64.59%	64.66%	0.06%	42.0%
		(43.83, 79.66)	(44.58, 79.02)	(-0.97, 1.07)	
Mean		58.98%	59.33%	0.35%	
		(36.99, 77.25)	(38.41, 76.45)	(-1.21,2.48)	
α-Amylase	25%	44.74%	47.24%	2.50%	79.8%
		(25.81, 60.77)	(29.59, 62.8)	(1.79,3.78)	
	50%	51.29%	52.52%	1.23%	66.4%
		(32.61, 65.28)	(35.03, 66.35)	(0.41, 2.51)	
	75%	53.29%	54.63%	1.34%	61.1%
		(34.38, 68.57)	(37.28, 69.47)	(0.21, 2.91)	
Mean		49.77%	51.46%	1.69%	
		(28.55, 67.21)	(31.32, 68.23)	(0.33, 3.15)	



(a) α -amylase

(b) Grain protein

(c) Grain yield

(d) Height

(e) Malt extract

Fig. 5 Percentage gain in prediction accuracy (y-axis) when using psBLUP vs percentage prediction accuracy per trait when using RRBLUP (x-axis). Each dot represents a trait-environmental combination showing the mean RRBLUP accuracy and gain in accuracy when using psBLUP over 100 runs. y- and x- marginal boxplots show the psBLUP gain in accuracy and RRBLUP accuracy, respectively. For ease of comparison, fitted regression lines have been added.

the penalized coefficients of the model, the proximity matrix needs to be stored and decomposed. When the number of SNPs is high, the memory needed to store such matrix is sizable. Such problem can partially be solved by encoding the matrices in sparse format. Still, the matrix needs to be decomposed to its eigenvectors and eigenvalues which becomes intensive for big p.

For computational efficiency, when the number of variables far exceeds the number of samples, an alternative parameterization can be used by writing model (8) as a single trait mixed model with subject-specific random effects. Let, $\boldsymbol{G} = \boldsymbol{X}^{*}\boldsymbol{X}^{*\top}$ be the realized additive relationship matrix indicating the relatedness between individuals. By ignoring any fixed effects, the mixed model with subject-specific random effects is written as:

$$\boldsymbol{y}^{*} = \boldsymbol{\alpha}^{*} + \boldsymbol{\varepsilon}^{*} \tag{9}$$

where $\boldsymbol{\alpha}^{*} \sim N(0, \boldsymbol{G}\sigma^2_{\boldsymbol{\alpha}^{*}})$. The information connecting subject-specific effects $\hat{\boldsymbol{\alpha}}^{*}$ to SNP effects $\hat{\boldsymbol{u}}^{*}$ is contained in $\boldsymbol{X}^{*}$ (Shen et al. 2013). After $\hat{\boldsymbol{\alpha}}^{*}$ is obtained, the SNP effects can be acquired as:

$$\hat{\boldsymbol{u}}^{*} = \boldsymbol{X}^{*\top}\boldsymbol{G}^{-1}\hat{\boldsymbol{\alpha}}^{*}. \tag{10}$$

Even though the search grid for the tuning parameter in psBLUP is reduced to one dimension since the mixed model solution is used, the computational time can be demanding for high p and high n by working with the augmented data solution i.e., the predictor data set is a $(n + p) \times p$ matrix. One approach to making the solution more efficient is by estimating the SNP coefficients per chromosome. Since SNPs are considered independent between chromosomes, multiple regularized linear models can be fit. Such approach could potentially yield superior accuracy by estimating chromosome specific regularization parameters and thus making the fit more flexible (by working with much smaller matrices). In addition, a shared λ_1 can also be estimated for each chromosome while λ_2 can vary per chromosome allowing for a better spatial flexibility per chromosome. In that case, the mixed model solution cannot be employed anymore.

Alternatives to psBLUP are the ante-dependence models (Yang and Tempelman 2012; Zeng et al. 2018b). These Bayesian models are based on the idea that SNP coefficients are dependent. A typical shortcoming of Bayesian methods is the computational time needed for estimating all coefficients using MCMC methods. For p SNPs, when only the first neighbor is considered (first order dependence), $2p - 1$ coefficients need to be estimated, making it burdensome for higher order dependencies and more dense SNP panels. Naturally, for every new SNP incorporated to the model, at least two more coefficients need to be estimated, resulting in additional computational time. We feel that psBLUP offers an alternative perspective to the same problem using a simpler set-up. Finally, the choice of connected neighbors in the ante-dependence models is fixed for all SNPs, while psBLUP allows for different number of neighbors per SNP, making it more flexible.

Important future research needs to be done. First, assessing how sensitive the results are to the selection of the proximity matrices. In this paper, we restricted the range within which SNPs were allowed to contribute information to 10cM, which for segregating populations like RILs and DHs is equivalent to a correlation between markers of .6. One could play around with this number to see whether the performance of psBLUP improves. For our choice of 10cM psBLUP often outperformed RRBLUP. Second, a more detailed evaluation of the sample size effect on the estimated accuracy needs to be done. Here, we used 25, 50, and 75% of the data samples as tests. A random subsample (as small as 25% of the original data) can initially be used in any study, to determine what is the maximum potential gain from psBLUP and what are some possible values for the smoothing parameter λ_2.

Finally, the sensitivity to the number of SNP needs to be studied. We would expect that the accuracy gain will be larger when using smaller number of SNPs. Using a big number of SNPs will naturally result in higher RRBLUP accuracy, thus smaller gain.

Acknowledgements The Authors would like to thank Martin Boer, Willem Kruijer, and Wilco Ligterink for constructive comments.

Funding The Authors declare that no financial support was received during the preparation of this manuscript.

Data Availability The Arabidopsis thaliana data are available upon resonable request from the Authors of Joosen et al. (2013). The Barley data is freely available from https://wheat.pw.usda.gov/ggpages/SxM/.

Declarations

Conflict of interest The authors have no relevant interests, financial or otherwise, to disclose.

Code availability An R implementation of the psBLUP function, as well as an R script showing its usage in our data analysis, can be obtained from https://git.wur.nl/Biometris/articles/psBLUP.

References

Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci 34(1):20–25

Bernardo R (1996) Best linear unbiased prediction of maize single-cross performance. Crop Sci 36(1):50–56

Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci 48(5):1649–1664

Chung FR, Graham FC (1997) *Spectral graph theory*. Number 92. American Mathematical Society

Clark SA, van der Werf J (2013) Genomic best linear unbiased prediction (gblup) for the estimation of genomic breeding values. In *Genome-Wide Association Studies and Genomic Prediction*, pages 321–330. Springer

Crossa J, de Los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh R, Dreisigacker S, Yan J et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186(2): 713–724

de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2):327–345

de Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigrees. *Genetics* 182(1): 375–385

de Vlaming R, Groenen PJ (2015) The current and future use of ridge regression for prediction in quantitative genetics. *BioMed Research international*, 2015

Endelman JB (2011) Ridge regression and other kernels for genomic selection with r package rrBLUP. The Plant Genome 4(3):250–255

Gianola D, Perez-Enciso M, Toro MA (2003) On marker-assisted prediction of genetic value: beyond the ridge. Genetics 163(1):347–365

Goddard ME, Hayes BJ, Meuwissen TH (2010) Genomic selection in livestock populations. Genet Res 92(5–6):413–421

Habier D, Fernando R, Dekkers JC (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinform 12(1):186

Hartl D (2011) *Essential genetics: a genomics perspective*. Sudbury, MA: Jones and Bartlett, 5th edition

Hayes BJ, Bowman PJ, Chamberlain A, Goddard M (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92(2):433–443

Hayes P, Liu B, Knapp S, Chen F, Jones B, Blake T, Franckowiak J, Rasmusson D, Sorrells M, Ullrich S et al (1993) Quantitative trait locus effects and environmental interaction in a sample of north american barley germ plasm. Theor Appl Genet 87(3):392–401

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49(1):1–12

Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52(1):146–160

Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1):55–67

Hunt CH, van Eeuwijk FA, Mace ES, Hayes BJ, Jordan DR (2018) Development of genomic prediction in sorghum. *Crop Science* 58(2):690–700

Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Func Genom 9(2):166–177

Joosen RVL (2013) *Imaging genetics of seed performance*. PhD thesis, Wageningen University & Research

Joosen RVL, Arends D, Li Y, Willems LA, Keurentjes JJ, Ligterink W, Jansen RC, Hilhorst HW (2013) Identifying genotype-by-environment interactions in the metabolism of germinating arabidopsis seeds using generalized genetical genomics. Plant Physiol 162(2):553–566

Li C, Li H (2008) Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24(9):1175–1182

Malosetti M, Voltas J, Romagosa I, Ullrich S, Van Eeuwijk F (2004) Mixed models including environmental covariables for studying QTL by environment interaction. Euphytica 137(1):139–145

Meuwissen T, Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Núñez-Antón VA, Zimmerman DL (2009) Antedependence models for longitudinal data. Chapman and Hall/CRC, UK

Piepho H, Ogutu J, Schulz-Streeck T, Estaghvirou B, Gordillo A, Technow F (2012) Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. Crop Sci 52(3):1093–1104

Shen X, Alam M, Fikse F, Rönnegård L (2013) A novel generalized ridge regression method for quantitative genetics. *Genetics*, pages genetics–112

Speed D, Balding DJ (2014) MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research* 24(9): 1550–1557

Van Binsbergen R, Calus MP, Bink MC, Eeuwijk FA, Schrooten C, Veerkamp RF (2015) Genomic prediction using imputed whole-genome sequence data in holstein friesian cattle. Genet Sel Evol 47(1):71

VanLiere JM, Rosenberg NA (2008) Mathematical properties of the r^2 measure of linkage disequilibrium. Theor Popul Biol 74(1):130–137

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423

Warrens M (2008) On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. Psychometrika 73:777–789

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75(2):249–252

Yang W, Tempelman RJ (2012) A bayesian antedependence model for whole genome prediction. Genetics 190(4):1491–1501

Zaykin DV, Pudovkin A, Weir BS (2008) Correlation-based inference for linkage disequilibrium with multiple alleles. Genetics 180(1):533–545

Zeng J, Garrick D, Dekkers J, Fernando R (2018) A nested mixture model for genomic prediction using whole-genome snp genotypes. PloS One 13(3):e0194683

Zeng J, Garrick D, Dekkers J, Fernando R (2018) A nested mixture model for genomic prediction using whole-genome SNP genotypes. PloS One 13(3):e0194683

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Statist Soc: Ser B (Statist Methodol) 67(2):301–320

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.