the geneticssociety

www.nature.com/hdy

ARTICLE

Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results

Prabina Kumar Meher [1]✉, Sachin Rustgi [2]✉ and Anuj Kumar [3]

© The Author(s), under exclusive licence to The Genetics Society 2022

We evaluated the performances of three BLUP and five Bayesian methods for genomic prediction by using nine actual and 54 simulated datasets. The genomic prediction accuracy was measured using Pearson's correlation coefficient between the genomic estimated breeding value (GEBV) and the observed phenotypic data using a fivefold cross-validation approach with 100 replications. The Bayesian alphabets performed better for the traits governed by a few genes/QTLs with relatively larger effects. On the contrary, the BLUP alphabets (GBLUP and CBLUP) exhibited higher genomic prediction accuracy for the traits controlled by several small-effect QTLs. Additionally, Bayesian methods performed better for the highly heritable traits and, for other traits, performed at par with the BLUP methods. Further, genomic BLUP (GBLUP) was identified as the least biased method for the GEBV estimation. Among the Bayesian methods, the Bayesian ridge regression and Bayesian LASSO were less biased than other Bayesian alphabets. Nonetheless, genomic prediction accuracy increased with an increase in trait heritability, irrespective of the sample size, marker density, and the QTL type (major/minor effect). In sum, this study provides valuable information regarding the choice of the selection method for genomic prediction in different breeding programs.

Heredity (2022) 128:519–530; https://doi.org/10.1038/s41437-022-00539-9

INTRODUCTION

Genotype selection based on genetic merit is an essential component of crop and animal breeding programs (Noshahr et al. 2017). Selection based on best linear unbiased prediction (BLUP; Viana et al. 2011), combined selection (Bhering et al. 2013), and recurrent selection (Ordas et al. 2012) was proposed to improve the selection accuracy. Albeit, marker-assisted selection (MAS) is used widely (Heffner et al. 2009), in which trait-associated markers, after necessary validation, are utilized in the breeding programs (Yáñez et al. 2014), it is less effective for the traits governed by many genes, with minor effects (Heffner et al. 2009; Xu and Crouch 2008).

Genomic selection (GS; Meuwissen et al. 2001) is an alternative to MAS and conventional phenotypic selection. In the GS, the genetic markers (with large and small effects) covering the entire genome are utilized, ensuring selection for not one or a few major effect Quantitative Trait Loci (QTLs) of interest but a combination of them contributing to the phenotype. In the GS model, the marker effects are first estimated based on the genotypic and phenotypic values of the training population. Then, the estimated marker effects are used to compute the genomic estimated breeding value (GEBV) for the selection candidates (test population) having only genotypic information (Heffner et al. 2009; Meuwissen et al. 2001). GS can be employed to predict GEBV at an early growth stage for a candidate, depending on the availability of the genotypic information (Wray et al. 2007; Guo et al. 2011). Thus, GS is advantageous for the traits that express later during life, specifically in the perennial species with long juvenile phases, and also for the traits that are costly to phenotype (Sayfzadeh et al. 2013). In other words, by applying GS, the breeding time and cost could be reduced by selecting candidates at an early growth stage (Tempelman 2015; Wolc et al. 2016; Yu et al. 2016). Hence, the selection based on GEBV could lead to a higher rate of genetic gain in the shorter generation time (Dekkers 2007; Daetwyler et al. 2010). Accuracy of GEBV is key to the success of genomic predictions, which is determined by several factors, including trait heritability (Hayes et al. 2009), marker density, QTL number, LD between QTL and associated marker (Meuwissen et al. 2001; Goddard 2009), size of the reference population, and genetic relationship between the reference and the test population (Zhong et al. 2009; Habier et al. 2007). Low-cost genotyping technologies such as single nucleotide polymorphism (SNP) arrays and genotyping by sequencing (GBS; Elshire et al. 2011; Ganal et al. 2011) offer new possibilities to improve the efficiency of breeding programs by allowing the application of GS (Bassi et al. 2016).

Several GS methods have been developed for predicting GEBV. These can be classified into three groups i.e., parametric, semi-parametric, and non-parametric (Crossa et al. 2014; Li et al. 2018). The BLUP alphabets, such as genomic BLUP (GBLUP; VanRaden 2008), ridge regression BLUP (RR-BLUP; Piepho 2009), SUPER BLUP (SBLUP; Wang et al. 2018), and compressed BLUP (CBLUP; Wang et al. 2018) are the linear parametric methods. The Bayesian alphabets, such as BayesA, BayesB, BayesC, Bayes

[1]Division of Statistical Genetics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi-12, India. [2]Department of Plant and Environmental Sciences, Clemson University Pee Dee Research and Education Center, Darlington, SC, USA. [3]Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi-12, India. Associate editor Chenwu Xu ✉email: prabina.meher@icar.gov.in; srustgi@clemson.edu

Received: 31 October 2021 Revised: 19 April 2022 Accepted: 19 April 2022
Published online: 4 May 2022

Table 1. Summary of the differences between the Bayesian and BLUP alphabets.

Approach	Method type	Marker effect	Marker effect distribution	GEBV estimation	The marker effect variance	Variance distribution of the marker effects	Estimation method
BLUP alphabets (GBLUP, CBLUP, SBLUP)	Linear parametric method	All markers are assumed to have effects on the trait variability	Marker effects are assumed to follow normal distribution.	All marker effects are used for estimating GEBV	Common variance for all marker effects	Normal	Linear mixed model with spectral factorization
Bayesian alphabets (BayesA, BayesB, BayesC, BL and BRR)	Non-linear parametric method	Only a limited number of markers are assumed to have effects on the trait variability	Different prior distributions are considered for different Bayesian models	BayesA, BL, and BRR are shrinkage type models where some of the marker effects are shrunk to zero, and rest of the markers are used for estimating GEBV. In BayesB and BayesC, the markers with non-null effects are used for estimating GEBV	Common variance for BayesC and BRR. Marker specific variances for BayesA, BayesB, and BL	The prior distribution of the variance is an inverse Chi-square	Markov chain Monte Carlo (MCMC) with Gibbs sampling

GBLUP genomic BLUP, *CBLUP* compressed BLUP, *SBLUP* SUPER BLUP, *BL* Bayesian LASSO, *BRR* Bayesian ridge regression.

ridge regression (BRR), and Bayes LASSO (BLASSO or BL), are the non-linear parametric methods. The reproducing kernel Hilbert space (RKHS; Gianola et al. 2006) is the most widely used semi-parametric method (de Los Campos et al. 2013), and the non-parametric method comprises mostly the machine learning techniques. Among the available methods, Bayesian and BLUP alphabets are most commonly used for GEBV prediction. The Bayesian and BLUP alphabets differ in the distribution of marker effects (Meuwissen et al. 2001; Hayes et al. 2009; de Los Campos et al. 2013; Habier et al. 2011; Legarra et al. 2008). In BLUP alphabets, all the markers are assumed to contribute to the trait, whereas a limited number of markers are assumed to have effects on the trait variance in Bayesian methods. Further, equal weights are given to the variance of all the markers in BLUP methods, but the Bayesian methods assign different weights to different markers. In BLUP alphabets, marker effects are assumed to follow a normal distribution (Meuwissen et al. 2001; VanRaden 2008), which means many QTL govern the trait, with each marker exhibiting a small effect (Meuwissen et al. 2001). In BLASSO, it is assumed that a small proportion of markers have large effects and a large proportion of markers have zero effects (Yi and Xu 2008; Hayes and Goddard 2010; Aguilar et al. 2010). In BayesC, it is assumed that there are effects of only a fraction of the markers, with each having a common variance (Habier et al. 2011). In BayesA, all the markers are assumed to have an effect, but each with different variances. In BayesB, it is considered that some of the markers have zero effects and other markers have effects with different variances. In BRR, marker effects follow Gaussian distribution, which induces shrinkage of estimates similar to ridge regression. Also, all marker effects are assumed to have equal variance in BRR. The ridge regression (RR) approach for genomic prediction produces results similar to that of the BLUP approach, given that the genetic covariance between genotypes is proportional to their genotypic similarity (Endelman 2011). The key differences between the Bayesian and BLUP alphabets are summarized in Table 1.

Among the single trait GS methods, the GBLUP model was used extensively on the actual datasets to evaluate the genomic prediction accuracy (Yabe et al. 2018; Tiede and Smith 2018; Fristche-Neto et al. 2018; Rio et al. 2019; Juliana et al. 2019; Michel et al. 2019; Cui et al. 2020). Additionally, the Bayesian alphabets were also deployed to the actual datasets for estimating the GEBV in several earlier studies (Crossa et al. 2010; Heffner et al. 2011; Pérez-Rodríguez et al. 2012; Zhao et al. 2013; Diaz et al. 2021). As far as the comparisons between GBLUP and different Bayesian approaches are concerned, some of the studies were performed on the actual datasets (Yoshida et al. 2018; Wang et al. 2019; Haile et al. 2020; Nsibi et al. 2020; Hong et al. 2020) and others on the simulated datasets (Daetwyler et al. 2010; Habier et al. 2011; Howard et al. 2014; Bhering et al. 2015; Alanoshahr et al. 2018). In a nutshell, the comparisons were mostly performed either with the actual or simulated datasets, where the comparisons were mostly between the GBLUP and specific variants of the Bayesian method. In other words, extensive comparisons between different BLUP variants and the Bayesian alphabets have been lacking hitherto.

Understanding the factors affecting the genomic prediction accuracy may be helpful for a breeder to design an effective genomic breeding strategy (Zhang et al. 2019). Thus, a comprehensive evaluation of the Bayesian and BLUP alphabets is required to assess the genomic prediction accuracy for traits with different genetic architectures. In this study, we evaluated the genomic prediction accuracy of three BLUP methods (GBLUP, CBLUP, and SBLUP) and five Bayesian methods (BayesA, BayesB, BayesC, BLASSO, and Bayes ridge regression) by using both actual and simulated datasets.

Table 2. Summary of the natural datasets.

Crop	Trait	#Genotype	Marker density	Marker type
Wheat	WY1, WY2, WY3, WY4	599	1279	DArT
Maize	MSS, MWW	264	1135	SNP
Barley	BER, BPR, BTW	307	2544	SNP

WY1, WY2, WY3, WY4 wheat yield in four mega environments, *MSS* maize yield in severe water stress condition, *MWW* maize yield in well-watered condition, *BER* ergosterol content, *BPR* protein content, *BTW* test weight, *SNP* single nucleotide polymorphism, *DArT* diversity array technology.

MATERIALS AND METHODS

Actual dataset

Genotypic and phenotypic datasets of three different crops (wheat, maize, and barley) were used. In total, nine quantitative traits were considered for the genomic prediction. For the wheat, traits were grain yield in four mega environments represented as WY1, WY2, WY3, and WY4. The two traits for the maize were grain yield under well-watered (MWW) and severe drought stress (MSS) conditions. The traits considered for the barley were protein content (BPR), test weight (BTW), and ergosterol (BER) content.

Both phenotypic and genotypic datasets of wheat were retrieved from the study of Crossa et al. (2010). There were 599 wheat lines in this dataset, with each line genotyped with 1447 DArT markers. The two marker alleles were denoted by 1 and 0 for presence and absence, respectively. After excluding the markers with <5% minor allele frequency (MAF), the dataset resulted in 1279 markers.

The maize dataset was also obtained from the Crossa et al. (2010) study, which contained 300 tropical maize lines from the CIMMYT's Global Maize Program, with each line genotyped with 1148 SNP markers. However, the grain yield data was only available for 264 lines. After removing the markers with <5% MAF, a dataset of 1135 markers was available for 264 maize lines.

The barley dataset was obtained from Nielsen et al. (2016) study. This dataset consisted of 309 advanced spring barley lines, where each line was genotyped for 7865 SNP markers using a 9 K barley chip. Data for the considered traits (BPR, BTW, and BER) were available only for 307 genotyped lines. After excluding the markers with <5% MAF and missing values, 2544 markers were retained for the analysis. A summary of the actual datasets is provided in Table 2.

Simulated dataset

Simulation of genotypic data. The procedure adopted by Hu and Yang (2014) was followed to simulate the genotypic and phenotypic datasets. We considered the population of n individuals, where each individual was assumed to be genotyped with m markers. By considering the individual with decreasing order of genetic relatedness, the genetic relatedness matrix among n individuals under the AR1 model was obtained as follows:

$$\Theta_{AR1} = \begin{bmatrix} 1 & \theta_a \ldots & \theta_a^{n-1} \\ \theta_a & 1 & \theta_a^{n-2} \\ \ldots & \ldots & \ldots \\ \theta_a^{n-1} & \theta_a^{n-2} & 1 \end{bmatrix}$$

where θ_a^t represents the genetic correlation between the two individuals i and j that are $t = |i - j|$ individuals apart. We also assumed that the three genotypes AA, Aa, and aa are in the 1:2:1 proportion for each individual to avoid any other complex genetic structure. For the generation of markers, we undertook the following steps.

- For 1st individual, generated a vector of random number $\mathbf{z}_1$ from N (0, 1).
- For 2nd individual, generated another vector of random number $\mathbf{z}_2$ from N (0, 1), where ρ ($\mathbf{z}_1$, $\mathbf{z}_2$) = θ_a.
- In general, for the ith individual, generated the random vector $\mathbf{z}_i$ using the recursive relation $\boldsymbol{z}'_i = \theta_a \boldsymbol{z}'_{i-1} + \sqrt{1 - \theta_a}\boldsymbol{k}$, where $\boldsymbol{k}$ is a random vector drawn from N (0, 1).
- Merged the generated random vector of all the n individuals to obtain the genotypic matrix $\boldsymbol{Z} = \{\boldsymbol{z}'_1, \boldsymbol{z}'_2, \ldots, \boldsymbol{z}'_n\}$.
- Converted the genotypic matrix to an indicator genotypic matrix ($\mathbf{Z}$) containing only three values 0, 1, and 2 to mimic the three possible genotypes at each locus, by replacing the normally distributed observations falling in the range ($-\infty$, -0.67449), (-0.67449, 0.67449) and (0.67449, ∞), respectively.

Simulation of phenotypic data. Phenotypic values ($\mathbf{y}$) were generated for all n individuals based on the model

$$\boldsymbol{y} = 1\mu + \boldsymbol{Zu} + \boldsymbol{e} = 1\mu + \boldsymbol{g} + \boldsymbol{e}.$$

Here, μ is the overall mean, $\mathbf{Z}$ is the genotypic matrix containing the values 0, 1, and 2. The $\mathbf{g}$ (= $\mathbf{Zu}$) is the vector of additive genetic effects with $\boldsymbol{g} \sim N\left(0, \boldsymbol{G}\sigma_g^2\right)$ and $\widehat{\boldsymbol{G}} = \frac{\boldsymbol{WW}^T}{2\sum_{j=1}^{m} p_j\left(1 - p_j\right)}$, where $w_{ij} = z_{ij} - 2\widehat{p}_j$ and $\widehat{p}_j$ is frequency of the reference allele for the jth marker. The $\mathbf{e}$ is the vector of random errors with $\boldsymbol{e} \sim N\left(0, \boldsymbol{I}_n\sigma_e^2\right)$. Further, the phenotypic variance $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$ and $\sigma_g^2 = [h^2/(1 - h^2)]\sigma_e^2$, with h^2 being the heritability.

Simulation scenario. The phenotypic datasets were simulated by considering three different heritability ($h^2 = 0.3$, 0.5, and 0.7), three different marker density ($m = 5000$, 20,000, and 50,000), three population size ($n =$ 300, 500, and 700) and two QTL size (5 and 20% markers as causal variants). Thus, a total of 54 datasets were generated to evaluate the performance of genomic prediction methods.

Cross validation

We adopted the repeated five fold cross-validation (CV) approach (González-Camacho et al. 2012; Gianola et al. 2014; Pérez-Rodríguez et al. 2012) to measure the genomic prediction accuracy. In other words, the experiment was repeated 100 times, and a 5-fold CV (Makowsky et al. 2011; Pérez-Cabal et al. 2012; Kramer et al. 2014) procedure was adopted in each experiment. For 5-fold CV, the whole dataset was first divided randomly into five disjoint subsets, with each subset having approximately the same number of individuals. Then, four subsets were used as the training population, and the remaining subset was used as the test population. This process was repeated five times so that each subset was used once as a test population.

Measuring genomic prediction accuracy

The predictive ability of the genomic prediction methods was measured by computing the Pearson's correlation coefficients between the GEBV and the observed phenotypic trait. The Pearson's correlation between the predicted GEBV and phenotypic values was computed in each fold of the five fold CV. The final accuracy was obtained by taking the average over 100 experiments. The slope of the regression of actual phenotype on GEBV was also computed to measure the bias in the GEBV (Echeverri et al. 2014). The regression coefficient close to 1 means no bias, whereas a slope of <1 and >1, respectively, indicate the underestimation and overestimation of GEBV (Resende et al. 2012; Neves et al. 2014).

Genomic prediction methods

We employed three BLUP methods (GBLUP, CBLUP, and SBLUP) and five Bayesian methods (BayesA, BayesB, BayesC, BLASSO, and BRR) for evaluating the genomic prediction accuracy. For this purpose, we assumed that there are n individuals having phenotypic records, with each individual genotyped with m markers. Further, we assumed $\boldsymbol{Z}_{n \times m}$ to be the genotypic matrix, where z_{ij} is the number of chosen alleles at j^{th} ($j = 1,2, \ldots, m$) locus for the i^{th} ($i = 1,2,\ldots,n$) genotyped individual.

BLUP alphabets

GBLUP: The GBLUP model can be written as

$$\boldsymbol{y} = \boldsymbol{1}_n\mu + \boldsymbol{g} + \epsilon$$

Here, $\boldsymbol{y}$ is the n-dimensional vector of phenotypic records, μ denotes the overall mean, $\boldsymbol{g}$ and $\in$ are the random vectors of additive genetic values and errors, respectively. Further, $\boldsymbol{g} \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{G}\sigma_{\boldsymbol{g}}^{2}\right)$ and $\epsilon \sim \boldsymbol{N}\left(\boldsymbol{0}, \boldsymbol{I}\sigma_{\boldsymbol{e}}^{2}\right)$, where $\boldsymbol{G}$ is the genotypic relationship matrix (GRM). We computed the GRM using VanRaden approach (VanRaden 2008), that is, $\boldsymbol{G} = \frac{\boldsymbol{WW}^T}{\boldsymbol{2}\sum_{j=1}^{m} \boldsymbol{p}_j\left(\boldsymbol{1} - \boldsymbol{p}_j\right)}$, where $w_{ij} = z_{ij} - 2p_j$ and p_j denotes the allelic frequency of the j^{th} marker. For the GBLUP model, $\widehat{\boldsymbol{g}}$ is the GEBV.

Table 3. Summary of the prior distribution of the hyper-parameters and the parametric values utilized for genomic prediction with different Bayesian methods.

	BayesA	BayesB	BayesC	BL	BRR
Prior distribution of the marker effects	Scaled-t with degree of freedom df_β and scale S_β	Scaled-t mixture, for the marker with non-zero effects, i.e., proportion π and $1-\pi$ proportion of the total markers are assumed to have null effects	Gaussian mixture, for the marker with non-zero effects, i.e., proportion π and $1-\pi$ proportion of the total markers are assumed to have null effects	Double exponential with parameter λ^2	Gaussian with mean μ_β and variance σ_β^2
Prior distribution of hyper parameters	$S_\beta \sim \Gamma(r,s)$	$S_\beta \sim \Gamma(r,s)$, $\pi \sim Beta\,(p_0,\,\pi_0)$		$\lambda^2 \sim \Gamma(r,s)$	$\sigma_\beta^2 \sim \chi^{-2}(\nu,\,S)$
Prior distribution of the variance of the marker effects and residual		$\sigma_\beta^2 \sim \chi^{-2}(\nu,\,S)$, $\sigma_e^2 \sim \chi^{-2}(\nu,\,S)$, *where* $S \sim \Gamma(r,s)$			
Parametric value considered	$s\,=\,1.1$, $R^2\,=\,0.5$, $df_\beta\,=\,5,\,\nu\,=\,5,\,S_\beta\,=\,\frac{var(y)\times R^2\times\left(df_\beta+2\right)}{MS_x},\,r\,=\,\frac{(s-1)}{S_\beta}$	$s\,=\,1.1,\,R^2\,=\,0.5,\,df_\beta\,=\,5,\,\nu\,=\,5$, $S_\beta\,=\,\frac{var(y)\times R^2\times\left(df_\beta+2\right)}{\left(\frac{MS_x}{\pi}\right)}$, $r\,=\,\frac{(s-1)}{S_\beta},\,\pi_0\,=\,0.5$, $p_0\,=\,10$		$s\,=\,1.1$, $R^2\,=\,0.5$, $v\,=\,5$, $r\,=\,\frac{(s-1)}{\left(\frac{2(1-R^2)}{R^2\times MS_x}\right)}$	$\mu_\beta\,=\,0$, $s\,=\,1.1$, $R^2\,=\,0.5$, $df_\beta\,=\,5,\,\nu\,=\,5$, $S_\beta\,=\,\frac{var(y)\times R^2\times(df_\beta+2)}{MS_x}$, $r\,=\,\frac{(s-1)}{S_\beta}$
Parameter for MCMC	All the five Bayesian models were implemented with 10,000 iterations, burn-in period of 1000 cycles and thin of 15 iterations				

MS_x Sum of the variances of markers under study, *Γ* Gamma, χ^{-2} Inverse Chi-square, *BL* Bayesian LASSO, *BRR* Bayesian ridge regression.

The rrBLUP R-package (Endelman 2011) with default parameters was used to implement the GBLUP model.

CBLUP and SBLUP: To improve the prediction accuracy of low heritable traits, Wang et al. (2018) proposed a variant of the GBLUP model known as compressed BLUP (CBLUP). In this model, kinship (GRM) computed among groups of individuals is employed instead of using the kinship among individuals. The genotypes are clustered into groups based on the GRM, and the optimum number of groups is determined based on the best likelihood. The kinship between any two groups is computed as the average kinship between the individuals of the two groups, and the kinship among individuals is replaced with the kinship of corresponding groups.

The SBLUP is another variant of the GBLUP model developed by Wang et al. (2018), aimed to improve the genomic prediction accuracy for the traits governed by a relatively small number of genes. The associated markers with the trait of interest are first determined through GWAS using the SUPER algorithm (Wang et al. 2014). Then, these associated markers are utilized to derive the kinship matrix, unlike GBLUP, in which all the markers are used to derive the kinship matrix. The CBLUP and SBLUP models were implemented using the GAPIT tool (Tang et al. 2016) with default parameters.

Bayesian alphabets. Consider the genomic prediction model

$$\boldsymbol{y}\,=\,\boldsymbol{1_n}\mu\,+\,\boldsymbol{Z}\beta\,+\,\boldsymbol{\epsilon}$$

where $\boldsymbol{Z}$ is the genotypic matrix and $\boldsymbol{\beta}$ is the vector of marker effects. All other notations are the same as mentioned in the GBLUP model. For this model, the prior density of the marker effects and other hyper-parameters for different Bayesian methods are explained as follows.

In BayesA, prior density of the marker effects follows scaled-t distribution, whereas for BayesB, the prior distribution of the effect of each marker is a mixture of scaled-t distribution with probability π and a distribution of point mass at zero with probability $(1-\pi)$. In BayesA and BayesB, the scaled-t distribution is implemented as finite mixture of scaled-normal densities to avoid computational complexity (Andrews and Mallows 1974). In particular, for BayesA, $\beta_j\,\sim\,N\left(0,\,\sigma_{\beta_j}^2\right)$, whereas for BayesB, $\beta_j\,\sim\,N\left(0,\,\sigma_{\beta_j}^2\right)$ and $\beta_j = 0$ with probability π and $(1-\pi)$ respectively. The prior densities of the marker effects for BayesC model are assumed to be Gaussian mixture, i.e., $\beta_j\,\sim\,N\left(0,\,\sigma_{\beta_j}^2\right)$ with probability π and $\beta_j = 0$ with probability $(1-\pi)$, where π is assumed to follow Beta distribution, i.e., $\pi \sim Beta(p_0,\,\pi_0)$ with $p_0 > 0$ and $\pi_0 \in [0,\,1]$ with $E(\pi) = \pi_0$ and $\mathrm{var}(\pi)\,=\,\frac{\pi_0(1-\pi_0)}{1+p_0}$. In BLASSO, prior densities of the non-zero marker effects are assumed to be double exponential (Park and Casella 2008), and also, the marker effects have locus-specific variance. The double exponential distribution is implemented as independent normal densities, i.e., $f\left(\beta_j|\tau_j^2,\sigma_e^2\right)\,\sim\,N\left(\beta_j|0,\,\tau_j^2\times\sigma_e^2\right)$, where the marker-specific scale parameter are i.i.d exponential distribution with rate parameter $\lambda^2/2$, i.e., $f\left(\tau_j^2\middle|\frac{\lambda^2}{2}\right)\,\sim\,Exp\left(\tau_j^2\middle|\frac{\lambda^2}{2}\right)$ and the rate parameter λ^2 is assigned a gamma prior, i.e., $f\left(\lambda^2|r,s\right)\,\sim\,\Gamma(r,s)$. In BRR model, the marker effects are assigned i.i.d Gaussian prior with same variance for all the effects, i.e., $f\left(\beta_j|\sigma_\beta^2\right)\,\sim\,N\left(\beta_j|0,\,\sigma_\beta^2\right)$. The overall mean μ is assigned a flat prior. The variance of the marker effects and the error variance (σ_e^2) for all the Bayesian models are assumed to follow χ^{-2}(v,S), where *v* and *S* are the degree of freedom and shape parameter, respectively. The prior density of the shape parameter *S* is assumed to follow Gamma distribution, with rate parameter *r* and shape parameter *s*. All the Bayesian methods were implemented using the BGLR R-package (Pérez and de los Campos 2014). The prior distributions of the marker effects, variance of the marker effects, hyper-parameters, and the parametric values utilized in this study are summarized in Table 3.

Heritability estimation

For the BLUP alphabets, the heritability of the trait was computed as $h^2\,=\,\frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$, where σ_g^2 and σ_e^2 represent the additive genetic and residual variances, respectively. For the Bayesian models, heritability was computed as $h^2\,=\,\frac{V_A}{V_A+\sigma_e^2}$, where $\mathrm{V_A}$ is the total additive genetic variance, that is,

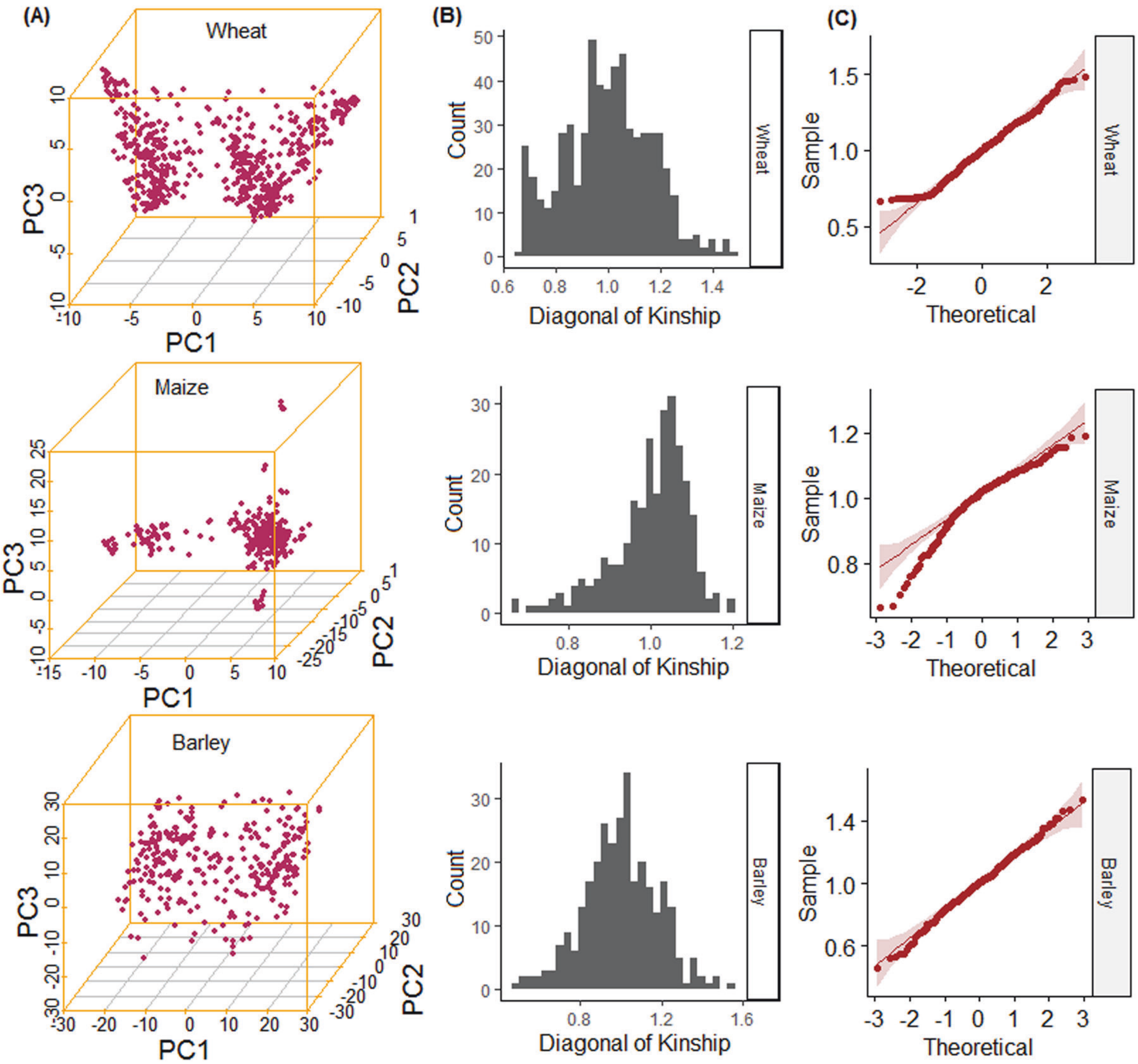

Fig. 1 Plots of the principal components and the distribution of the genomic relationship matrix of the genotypic dataset. A Scatter plots of the first three principal components of the genotypic data. No distinct clusters are observed in any of the three datasets. **B** Histograms of the diagonal values of the genomic relationship matrix. Although distribution is not observed to be perfectly normal, distinct peaks are not seen for any of the three datasets. **C** QQ plots of the diagonal values of the genomic relationship matrix of the actual datasets. From all these plots, it is inferred that the individuals are randomly drawn from a single population without any family structure for a given dataset.

$V_A = \pi \times 2\widehat{\sigma}^2_{SNP} \sum_{j=1}^{m} p_j q_j$, where p_j and q_j are the allelic frequencies of j^{th} locus and π denotes proportion of markers with non-zero effects.

RESULTS

Genomic relationship matrix of the actual data

The GRM for the wheat, maize, and barley were computed using the VanRaden approach (VanRaden 2008). From the distribution of the diagonal values of GRM (Fig. 1B), multiple distinct peaks were not seen, although the distributions were not perfectly normal, as evident from the QQ plots of the diagonal values of GRM (Fig. 1C). Also, no clearly distinguishable genotype clusters were formed in all three datasets (Fig. 1A). Given this knowledge, we could say that each genotype was randomly taken from the population for a given dataset without any family structure, which indicates the absence of any hidden population structure in the datasets. From the plot of the first three principal components (PCs), 15.15, 6.13, and 3.65% variations were explained for wheat (Fig. 1A). Similarly, the first three PCs explained 4.32, 3.01, and 2.34% of total variations in the maize genotypic dataset. And, for the barley, the first three PCs explained 10.86, 5.09, and 4.64% of variations.

Estimate of heritability for the actual dataset

Heritability estimates with different methods for the nine actual datasets are shown in Fig. 2. For the wheat yield trait, higher heritability estimates were observed for the BLUP alphabets, whereas the Bayesian methods resulted in higher heritability estimates for the rest of the five traits. SBLUP predicted higher heritability for WY2 (0.665 ± 0.029) and WY3 (0.589 ± 0.019), whereas the heritability estimates for WY1 (0.694 ± 0.011) and WY4 (0.606 ± 0.011) were higher for CBLUP. For both maize traits, higher heritability estimates were obtained with the BRR model, i.e., 0.604 ± 0.032 for MWW and 0.527 ± 0.034 for MSS (Fig. 2). For BPR and BTW traits of barley, BayesB, respectively, exhibited higher estimates of heritability, 0.54 ± 0.024 and 0.546 ± 0.023, compared to other Bayesian methods (Fig. 2). Further, for the barley BER trait (0.706 ± 0.013), the heritability estimates were a little higher as compared to the other Bayesian alphabets. Among the Bayesian alphabets, the heritability estimate was less with BLASSO for all the traits except WY4. When all the methods were accounted for, the range of heritability for the wheat yield trait was 0.368 ± 0.03 to 0.694 ± 0.011. Similarly, 0.198 ± 0.02 to 0.527 ± 0.034 and 0.228 ± 0.056 to 0.604 ± 0.032 were the ranges of

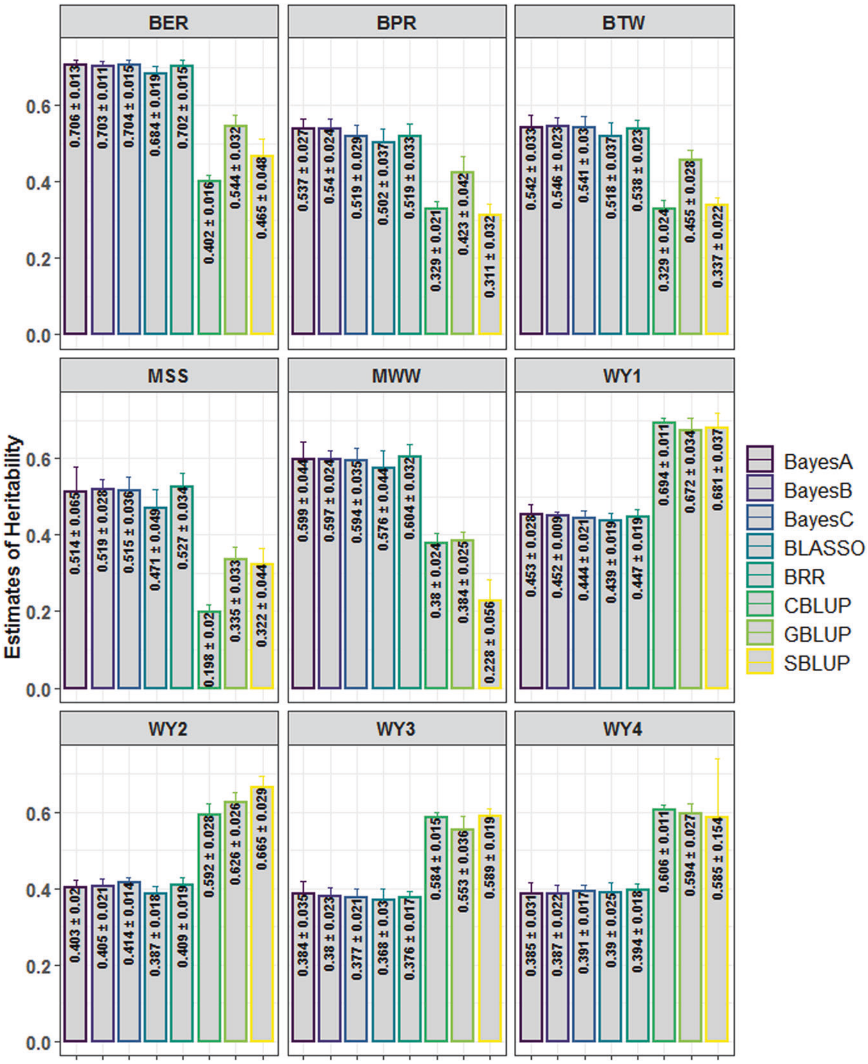

Fig. 2 Heritability estimates for all nine actual datasets with three BLUP and five Bayesian methods. The heritability is estimated by following a repeated fivefold cross-validation approach, where the experiment is repeated 100 times. The heritability estimates for the wheat yield datasets were higher for the BLUP methods. However, the Bayesian methods achieved higher heritability estimates for the rest of the five traits. Heritability is computed as the ratio of the additive genetic variance to the phenotypic variance. For the BLUP alphabets, all markers are used to estimate additive genetic variance, whereas only the markers with non-null effect are used for the Bayesian methods.

heritability estimates for the maize yield trait under severe stress and well-watered condition, respectively (Fig. 2). For barley BER, BPR, and BTW traits, the heritability was found to range, respectively, from 0.402 ± 0.016 to 0.706 ± 0.013, 0.311 ± 0.032 to 0.54 ± 0.024, and 0.329 ± 0.024 to 0.546 ± 0.023 (Fig. 2).

Genomic prediction accuracy for the actual dataset

The genomic prediction accuracies are shown in Fig. 3. Genomic prediction accuracies were observed to increase with the traits' heritability. For instance, lower prediction accuracy was observed for the trait WY3 (0.332 ± 0.022 to 0.401 ± 0.013), for which heritability estimates were low. On the other hand, higher accuracies were observed for the BER trait (0.677 ± 0.027 to 0.761 ± 0.011), for which the heritability estimates were also higher. Among the BLUP methods, SBLUP exhibited the lowest performance, and GBLUP achieved the highest accuracy. In fact, out of nine traits, GBLUP exhibited the highest accuracy for WY2 (0.501 ± 0.012), WY4 (0.468 ± 0.012), MWW (0.561 ± 0.019), and BTW (0.641 ± 0.018). BRR achieved the highest genomic prediction accuracy for WY1 (0.529 ± 0.011) and MSS (0.423 ± 0.023) traits (Fig. 3). The BayesA, BayesB, and CBLUP, respectively, showed the highest accuracy for the BER (0.761 ± 0.011), BPR (0.579 ± 0.017), and WY3 (0.401 ± 0.013) traits (Fig. 3). Among Bayesian methods, BayesA achieved the highest accuracies for three traits, i.e., WY2 (0.497 ± 0.014), BTW (0.578 ± 0.017), and BER (0.761 ± 0.011), whereas BRR achieved the highest accuracies for four traits, i.e., MSS (0.423 ± 0.023), WY1 (0.529 ± 0.011), WY4 (0.459 ± 0.013), and MWW (0.559 ± 0.018). The Bayesian LASSO and BayesB resulted in the highest accuracy for WY3 (0.397 ± 0.013) and BPR (0.579 ± 0.017), respectively.

Genomic prediction accuracy with simulated dataset

The genomic prediction accuracies of the Bayesian and BLUP alphabets for the simulated datasets are shown in Fig. 4. The Bayesian models were found to achieve higher genomic prediction accuracy when the traits were highly heritable (0.5, 0.7) and governed by few QTLs (5% markers as causal variants), with each exhibiting a larger effect (Fig. 5) on the trait genetic variability. When the trait heritability was low and controlled by few QTLs, each having a larger effect, the GBLUP performed at par with Bayesian methods, barring a few exceptions. On the other hand, when many QTLs controlled the trait, the GBLUP

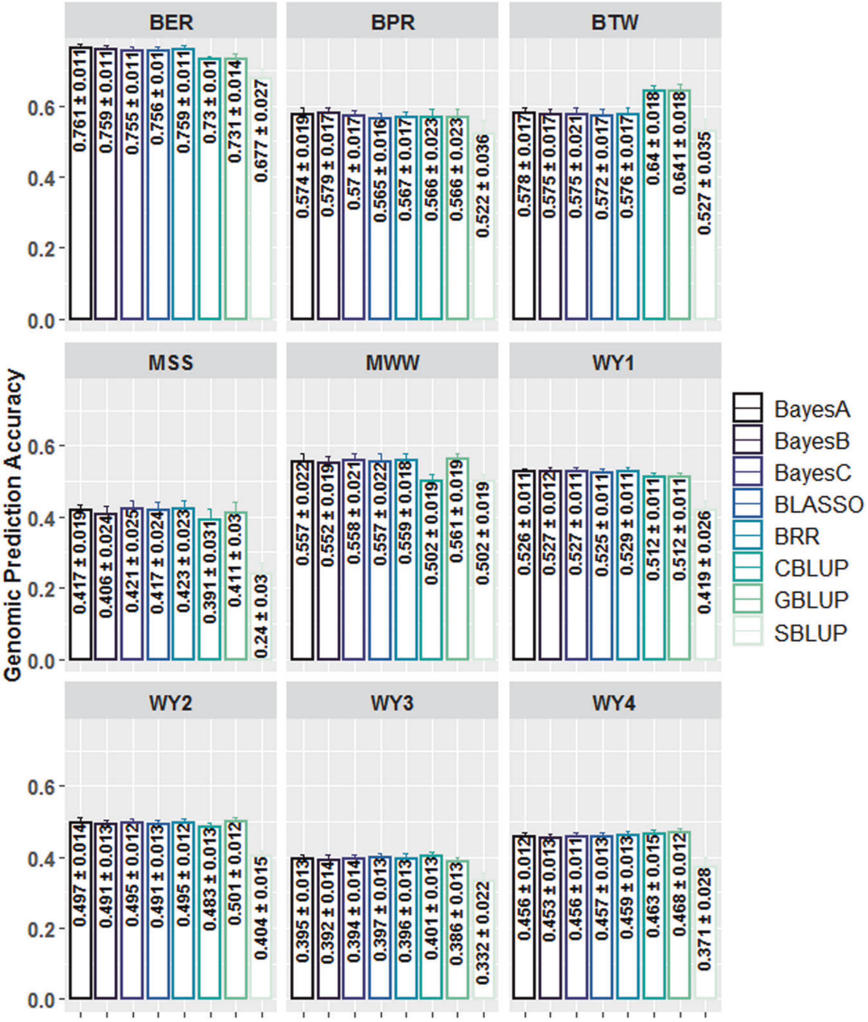

Fig. 3 The estimates of genomic prediction accuracy with actual datasets. The genomic prediction accuracy is measured via the Pearson's correlation coefficient between the genomic estimated breeding value (GEBV) and the observed phenotypic data representing the true breeding value. A repeated fivefold cross-validation approach was adopted to compute the correlation, and the experiment was replicated 100 times. The final accuracy is computed by taking an average over all the fivefolds, and 100 replicates. The SBLUP achieved the lowest accuracy among all the methods. The GBLUP achieved higher accuracy for the yield traits of wheat except for WY1. Except for BTW, the Bayesian methods performed better for the rest of the four traits. The higher accuracy of BLUP and Bayesian methods are seen to be according to their heritability estimates.

method achieved higher accuracy than the Bayesian alphabets, with some exceptions. For the trait with higher heritability and controlled by many QTLs, each having a small effect (Fig. 5) on the trait genetic variability, the Bayesian methods were observed to perform either better than or at par with the GBLUP method. It was also found that the genomic prediction accuracies increase with an increase in the trait heritability, irrespective of the population size, number of QTLs, and marker density. Further, we observed that the genomic prediction accuracy increases with an increase in the population size. For the low heritable (0.3) trait, the BRR achieved the highest accuracy among the considered Bayesian methods, regardless of the genetic architecture of the trait. When the heritability was high (0.5, 0.7), higher genomic prediction accuracy was achieved with either BRR or BLASSO for all combinations of marker density, population size, and QTL percentage. GBLUP achieved higher accuracy for almost all the simulated datasets among the BLUP alphabets. On the other hand, the SBLUP achieved the lowest accuracy among all the Bayesian and BLUP alphabets. The prediction accuracies are also seen to increase with an increase in the marker density.

Bias in GEBV with simulated datasets

The bias in GEBV predictions for all the simulated datasets is shown in Fig. 6. The bias in the GEBV declined with an increase in the trait heritability, irrespective of the sample size, marker density, and the methods used (Fig. 6). For instance, the magnitude of bias in GEBV is closer to 1 for trait heritability 0.5 compared to 0.3, and heritability 0.7 compared to 0.5. The SBLUP was found to mostly under-predict GEBVs irrespective of the marker density, sample size, and trait heritability. Further, it was observed that the GBLUP is the least biased method among all the Bayesian and BLUP alphabets, irrespective of the trait genetic architecture. Among Bayesian alphabets, both BLASSO and BRR were found to be less biased in GEBV predictions. However, GEBV was found to be over-predicted (bias >1) with BLASSO for the highly heritable trait (0.7). Generally, the GEBV was observed to be mostly under-predicted for BayesA, followed by BayesB and BayesC. With an increase in population size, the effect of heritability was seen to decline on the bias estimation. For the given population size, the difference in the bias in the GEBV prediction was found to be more between the differently heritable traits. However, such a difference was

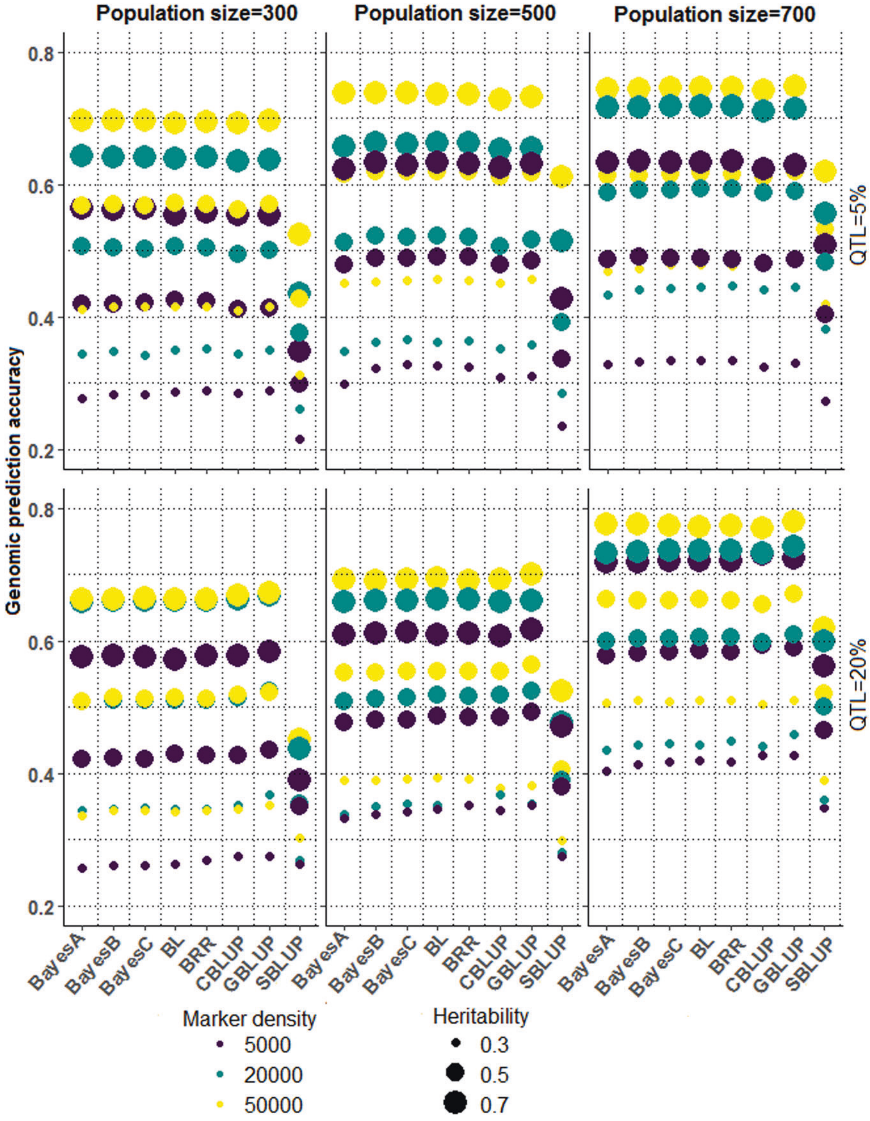

Fig. 4 The genomic prediction accuracy of the Bayesian and BLUP alphabets for different combinations of sample size, marker density, heritability, and QTL size (5 and 20% markers as causal variants). It can be seen that with an increase in heritability, genomic prediction accuracy also increased irrespective of sample size, marker density, and QTL size. The accuracy is also increased with an increase in the sample size and the marker density. The genomic prediction accuracies are seen to be lowest for SBLUP, irrespective of the trait genetic architecture. The GBLUP secured higher accuracy for the traits controlled by many QTLs, each having a small effect. On the other hand, the Bayesian methods produced higher accuracy when a few QTLs governed the trait, each with a larger effect on the genotypic variability.

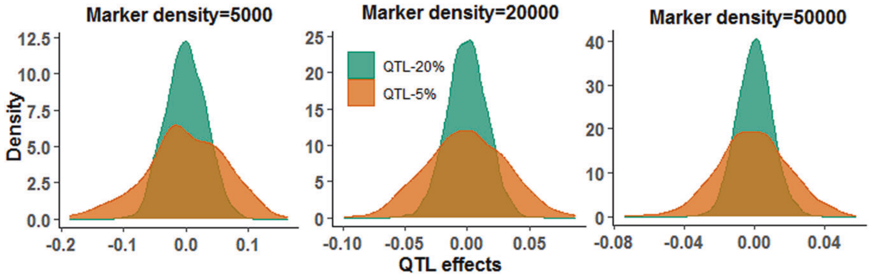

Fig. 5 Density graph of the effects considering 5 and 20% marker as causal variants (QTLs). For each marker density, the effects of the 5% QTLs are larger than that of 20% QTLs.

observed to decline with an increase in the population size. Bias in the GEBV was found to be at par for both GBLUP and CBLUP, as GBLUP is a special case of the CBLUP method (Wang et al. 2018).

DISCUSSION

Genomic selection is an important tool that offers ample opportunities to enhance the genetic gain for the complex traits in plants and animals (Bhat et al. 2016). With the availability of

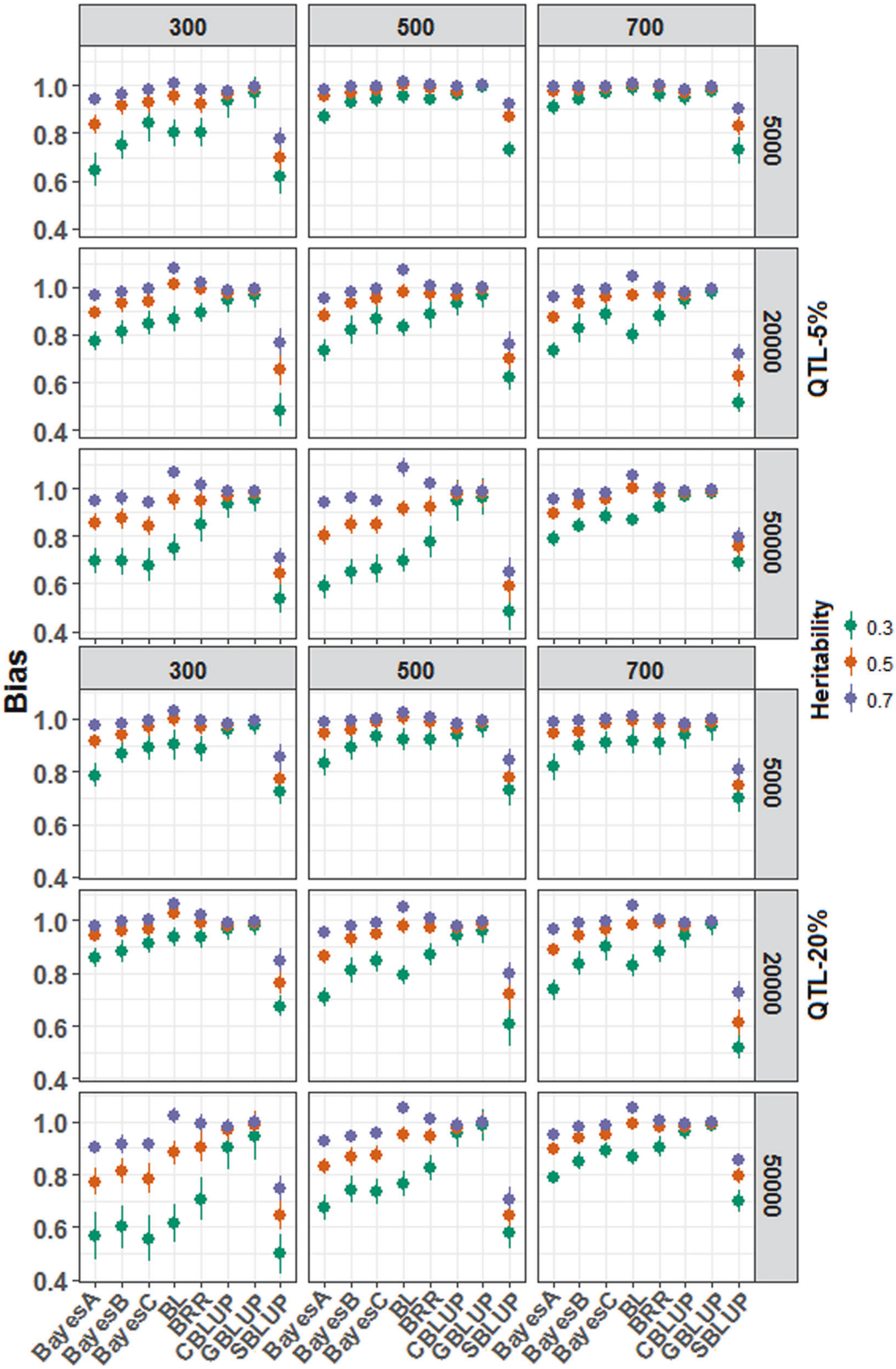

Fig. 6 The bias estimation in genomic estimated breeding value (GEBV) with the simulated dataset. The bias is measured as the regression coefficients obtained by regressing the observed phenotypic values upon the predicted phenotypic values. Further, a repeated (100 times) fivefold cross-validation approach was adopted to compute the coefficients. The final coefficient estimates were computed by taking the average of the five folds and 100 replicates. The GBLUP is the least biased method for the GEBV prediction. GBLUP and CBLUP were identified as robust methods among the BLUP alphabets, irrespective of the trait genetic architecture. The BRR and BLASSO were less biased among the Bayesian models than the other variants.

cost-effective and high-throughput genotypic platforms, such as GBS, the application of GS has become more feasible in breeding programs (Poland et al. 2012). The GS accounts for all genetic variation and hence allows the selection of individuals with higher breeding values. Moreover, the genetic gain per generation cycle by applying GS is much higher than the conventional selection methods (Bassi et al. 2016).

Several factors determine GS accuracy. Among these factors, the choice of the GS model is an important one. Many statistical models are developed for GS, where the Bayesian and BLUP alphabets are the most frequently used methods (Pais de Arruda et al. 2016; Zhang et al. 2015; Hoffstetter et al. 2016). However, a comprehensive comparative analysis of the Bayesian and BLUP alphabets was lacking. Therefore, the present study focused on the comparative analysis of these methods with attention to genomic prediction accuracy.

The BLUP alphabets may be more accurate in predicting heritability with the DArT markers, which may be the probable reason for the higher heritability estimates for the wheat dataset with BLUP methods. The variation in the heritability estimates for the four wheat yield traits corresponding to four different mega environments may be attributed to the interaction of genotype with the environment. For a given trait, heritability estimates were observed to vary with the estimation methods. In the case of BLUP alphabets, the effects of all the markers are accounted for computing the additive genetic variance, whereas the selected markers with non-null effects were only used for computing the additive genetic variance in the case of Bayesian alphabets. It may be one of the probable reasons for different heritability estimates for Bayesian and BLUP alphabets. The variability in the heritability estimates among the Bayesian methods may also be attributed to the different assumptions regarding the distribution of the marker effects, which as a result, affects the estimation of variance components. In the case of BLUP alphabets, the approach for estimating the GRM is different, which may be the possible reason for different heritability estimates.

For the wheat yield traits, GBLUP achieved the highest accuracy for WY2 and WY4, whereas CBLUP achieved the highest accuracy for WY3. In other words, the BLUP alphabets achieved higher accuracy than other methods, and it corresponded with the higher heritability estimates observed with the BLUP methods. In four out of the rest five traits, Bayesian methods achieved higher genomic prediction accuracies, which is also in accordance with the higher heritability estimates with the Bayesian alphabets. Further, accuracies were observed to be less stable (high standard error) for MSS, BTW, BPR, and MWW and more stable (less standard error) for the yield traits and BER. The higher stability in prediction accuracy in the wheat yield trait may be due to the large sample size, whereas higher stability for BER may be attributed to the higher trait heritability. The accuracy of GBLUP was observed to be either higher or at par with the accuracy of CBLUP. Further, GBLUP and CBLUP methods achieved higher accuracy than the SBLUP method. Accuracies of SBLUP were observed to be lowest and least stable than the other methods.

Besides the actual datasets, the genomic prediction accuracy was also evaluated with the simulated datasets. From the analysis, no single method was found to perform better for all the simulated traits. Indeed, different methods performed better with different trait genetic architectures. For the highly polygenic trait, i.e., the trait controlled by many QTLs with each exhibiting a small effect on the genetic variance, BLUP methods (except SBLUP) were found to perform better as compared to the Bayesian methods. However, for a trait governed by few QTLs with each having a larger effect, the Bayesian methods were often found to outperform BLUP methods. It was also noticed that for the highly heritable trait, the Bayesian alphabets either achieved higher accuracy than BLUP models or performed at par with BLUP methods. Thus, it can be said that the suitable method for genomic prediction largely depends on the genetic architecture of a trait. Nonetheless, the genomic prediction accuracies were observed to be increased with an increase in the heritability of the trait, population size, and marker density. An increase in the genomic prediction accuracy with an increase in the marker density may be attributed to the tendency of more QTL being in LD with a marker (Heffner et al. 2009; Desta and Ortiz 2014). However the prediction accuracy may eventually attain a plateau with an increase in the marker density depending upon the within-population genetic diversity as well as the genetic relatedness between the training population and the selection candidates (de los Campos et al. 2013). An increase in the marker density may also adversely affect the prediction accuracy of the Bayesian methods because of slow or no convergence of Markov chain Monte Carlo (MCMC) iterations (Zhang et al. 2019). Besides, the inclusion of many markers without any effect on the trait variance may result in the overfitting of the model.

The GEBV was under-estimated with SBLUP irrespective of heritability, marker density, and sample size. At the same time, the bias was more stable with GBLUP and CBLUP irrespective of heritability, sample size, marker density, and QTL number. It was also noticed that the GBLUP method was the least biased among all the methods as far as the GEBV prediction is concerned. The bias in the GEBV prediction was generally found to decline for the high heritability traits. Among Bayesian methods, BLASSO and BRR were seen to be less biased. However, for the highly heritable traits, the GEBV was overestimated with BLAASO. On the other hand, the BayesA, BayesB, and BayesC methods highly underestimate the GEBV for the low heritability traits.

Though computational time depends upon the population size and analysis method, BLUP methods take less time than Bayesian counterparts. It is expected as Bayesian methods are dependent upon the size of MCMC through simulated Markov chain (Meuwissen et al. 2001) and Gibb sampler for the posterior marginal distributions of the model parameters.

CONCLUSION

In this study, we evaluated the performance of three BLUP and five Bayesian methods using actual (three datasets, nine traits) and simulated (three marker density × three heritability × three sample size × two QTL size) data. The performance of the models was found to reflect the trait genetic architecture. The SBLUP method was the least performer among all the methods. The other two BLUP alphabets achieved higher accuracy for the traits governed by many QTLs, each with small effects. On the other hand, the Bayesian methods produced higher accuracy for the traits governed by few QTLs, with each having a larger effect on the genotypic variability. Concerning the bias in genomic prediction accuracy, the GBLUP method was the least biased approach among all the considered models. The knowledge generated from this study is believed to supplement the existing knowledge on the choice of genomic prediction methods for breeding programs.

REFERENCES

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. J Dairy Sci 93:743–752

Alanoshahr F, Rafat SA, Imany-Nabiyyi R, Alijani S, Robert-Granie C (2018) The impact of different genetic architectures on accuracy of genomic selection using three Bayesian methods. Iran J Appl Anim Sci 8:53–59

Andrews DF, Mallows CL (1974) Scale mixtures of normal distributions. J R Stat Soc Ser B 36:99–102

Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J (2016) Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Sci 242:23–36

Bhat JA, Ali S, Salgotra RK, Mir ZA, Dutta S, Jadon V et al. (2016) Genomic selection in the era of next generation sequencing for complex traits in plant breeding. Front Genet 7:221

Bhering LL, Barrera CF, Ortega D, Laviola BG, Alves AA, Rosado TB et al. (2013) Differential response of Jatropha genotypes to different selection methods indicates that combined selection is more suited than other methods for rapid improvement of the species. Ind Crops Products 41:260–265

Bhering LL, Junqueira VS, Peixoto LA, Cruz CD, Laviola BG (2015) Comparison of methods used to identify superior individuals in genomic selection in plant breeding. Genet Mol Res 14:10888–10896

de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. PLOS Genet 9:e1003608

Crossa J, Campos G, de L, Pérez P, Gianola D, Burgueño J, Araus JL et al. (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J et al. (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112:48–60

Cui Z, Dong H, Zhang A, Ruan Y, He Y, Zhang Z (2020) Assessment of the potential for genomic selection to improve husk traits in maize. G3 10:3741–3749

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031

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

Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet 124:331–341

Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19:592–601

Diaz S, Ariza-Suarez D, Ramdeen R, Aparicio J, Arunachalam N, Hernandez C, Diaz H et al. (2021) Genetic architecture and genomic prediction of cooking time in common bean (*Phaseolus vulgaris* L.). Front Plant Sci 11:2257

Echeverri J, Zambrano J, Herrera AL (2014) Genomic evaluation of Holstein cattle in Antioquia (Colombia): a case study. Rev Colomb Cienc Pecu 27:306–314

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLOS ONE 6:e19379

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4, https://doi.org/10.3835/plantgenome2011.08.0024

Fristche-Neto R, Akdemir D, Jannink JL (2018) Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. Theor Appl Genet 131:1153–1162

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A et al. (2011) A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLOS ONE 6:e28334

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776

Gianola D, Weigel KA, Krämer N, Stella A, Schön C-C (2014) Enhancing genome-enabled prediction by bagging genomic BLUP. PLOS ONE 9:e91693

Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245–257

González-Camacho JM, de Los Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G et al. (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet 125:759–771

Guo G, Zhou Z, Wang Y, Zhao K, Zhu L, Lust G et al. (2011) Canine hip dysplasia is predictable by genotyping. Osteoarthr Cartil 19:420–429

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

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

Haile TA, Heidecker T, Wright D, Neupane S, Ramsay L, Vandenberg A et al. (2020) Genomic selection for lentil breeding: empirical evidence. Plant Genome 13: e20002

Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal breeding. Genome 53:876–883

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

Heffner EL, Jannink JL, Sorrells ME (2011) Genomic selection accuracy using multi-family prediction models in a wheat breeding program. Plant Genome 4, https://doi.org/10.3835/plantgenome2010.12.0029

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

Hoffstetter A, Cabrera A, Huang M, Sneller C (2016) Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat. G3 6:2919–2928

Hong JP, Ro N, Lee HY, Kim GW, Kwon JK, Yamamoto E et al. (2020) Genomic selection for prediction of fruit-related traits in pepper (*Capsicum* spp.). Front Plant Sci 11:570871

Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 4:1027–1046

Hu Z, Yang R-C (2014) Marker-based estimation of genetic parameters in genomics. PLOS ONE 9:e102715

Juliana P, Poland J, Huerta-Espino J, Shrestha S, Crossa J, Crespo-Herrera L et al. (2019) Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nat Genet 51:1530–1539

Kramer M, Erbe M, Seefried FR, Gredler B, Bapst B, Bieber A et al. (2014) Accuracy of direct genomic values for functional traits in Brown Swiss cattle. J Dairy Sci 97:1774–1781

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

Li B, Zhang N, Wang Y-G, George AW, Reverter A, Li Y (2018) Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet 9:237

Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB et al. (2011) Beyond missing heritability: prediction of complex traits. PLOS Genet 7: e1002051

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Michel S, Löschenberger F, Ametz C, Pachler B, Sparry E, Bürstmayr H (2019) Combining grain yield, protein content and protein quality by multi-trait genomic selection in bread wheat. Theor Appl Genet 132:2767–2780

Neves HH, Carvalheiro R, O'Brien AMP, Utsunomiya YT, do Carmo AS, Schenkel FS et al. (2014) Accuracy of genomic predictions in Bos indicus (Nellore) cattle. Genet Sel Evol 46:17

Nielsen NH, Jahoor A, Jensen JD, Orabi J, Cericola F, Edriss V et al. (2016) Genomic prediction of seed quality traits using advanced barley breeding lines. PLOS ONE 11:e0164494

Noshahr FA, Rafat SA, Imany-Nabiyyi R, Alijani S, Robert-Granie C (2017) Genomic accuracy in different genetic architecture and genomic structure. Ind J Anim Sci 87:324–328

Nsibi M, Gouble B, Bureau S, Flutre T, Sauvage C, Audergon JM et al. (2020) Adoption and optimization of genomic selection to sustain breeding for apricot fruit quality. G3 10:4513–4529

Ordas B, Butron A, Alvarez A, Revilla P, Malvar RA (2012) Comparison of two methods of reciprocal recurrent selection in maize (*Zea mays* L.). Theor Appl Genet 124:1183–1191

Pais de Arruda M, Lipka A, Brown P, Krill A, Thurber C, Brown-Guedira G et al. (2016) Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum* L.). Mol Breed 36:84

Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103:681–686

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495

Pérez-Cabal MA, Vazquez AI, Gianola D, Rosa GJM, Weigel KA (2012) Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. Front Genet 3:27

Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y, Dreisigacker S (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 2:1595–1605

Piepho HP (2009) Ridge regression and extensions for genomewide selection in maize. Crop Sci 49:1165–1176

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y et al. (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5, https://doi.org/10.3835/plantgenome2012.05.0005

Resende MFR, Muñoz P, Resende MDV, Garrick DJ, Fernando RL, Davis JM et al. (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 190:1503–1510

Rio S, Mary-Huard T, Moreau L, Charcosset A (2019) Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. Theor Appl Genet 132:81–96

Sayfzadeh S, Honarvar M, Taheri F, Afshari K (2013) Accuracy of genomic prediction using random regression BLUP: a simulation study. Agrochimica Pisa 57:27–31

Tang Y, Liu X, Wang J, Li M, Wang Q, Tian F et al. (2016) GAPIT version 2: an enhanced integrated tool for genomic association and prediction. Plant Genome 9: plantgenome2015.11.0120

Tempelman RJ (2015) Statistical and computational challenges in whole genome prediction and genome-wide association analyses for plant and animal breeding. JABES 20:442–466

Tiede T, Smith KP (2018) Evaluation and retrospective optimization of genomic selection for yield and disease resistance in spring barley. Mol Breed 38:1–16

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

Viana JMS, Faria VR, Silva FFE, de Resende MDV (2011) Best linear unbiased prediction and family selection in crop species. Crop Sci 51:2371–2381

Wang J, Zhou Z, Zhang Z, Li H, Liu D, Zhang Q et al. (2018) Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. Heredity 121:648–662

Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z (2014) A SUPER powerful method for genome wide association study. PLOS ONE 9:e107684

Wang X, Miao J, Chang T, Xia J, An B, Li Y et al. (2019) Evaluation of GBLUP, BayesB and elastic net for genomic prediction in Chinese Simmental beef cattle. PLOS ONE 14:e0210442

Wolc A, Kranis A, Arango J, Settar P, Fulton JE, O'Sullivan NP et al. (2016) Implementation of genomic selection in the poultry industry. Anim Front 6:23–31

Wray NR, Goddard ME, Visscher PM (2007) Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 17:1520–1528

Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci 48:391–407

Yabe S, Yoshida H, Kajiya-Kanegae H, Yamasaki M, Iwata H, Ebana K et al. (2018) Description of grain weight distribution leading to genomic selection for grain-filling characteristics in rice. PLOS ONE 13:e0207627

Yáñez JM, Houston RD, Newman S (2014) Genetics and genomics of disease resistance in salmonid species. Front Genet 5:415

Yi N, Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics 179:1045–1055

Yoshida GM, Bangera R, Carvalheiro R, Correa K, Figueroa R, Lhorente JP et al. (2018) Genomic prediction accuracy for resistance against Piscirickettsia salmonis in farmed rainbow trout. G3 8:719–726

Yu X, Li X, Guo T, Zhu C, Wu Y, Mitchell SE et al. (2016) Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat Plant 2:16150

Zhang H, Yin L, Wang M, Yuan X, Liu X (2019) Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front Genet 10:189

Zhang X, Pérez-Rodríguez P, Semagn K, Beyene Y, Babu R, López-Cruz MA et al. (2015) Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. Heredity 114:291–299

Zhao Y, Zeng J, Fernando R, Reif JC (2013) Genomic prediction of hybrid wheat performance. Crop Sci 53:802–810

Zhong S, Dekkers JCM, Fernando RL, Jannink J-L (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. Genetics 182:355–364

ACKNOWLEDGEMENTS

The authors acknowledge the Advanced Super Computing Hub for Omics Knowledge in Agriculture (ASHOKA) facility at ICAR-IASRI, New Delhi, India, and the NIFA Hatch/Multi-state grant (S009) for financial support to SR.

AUTHOR CONTRIBUTIONS

PKM was responsible for designing the study, writing the report, conducting the search, screening potentially eligible studies, extracting and analysing data, interpreting results, updating reference lists and creating "Summary of findings" tables and figures. SR was responsible for designing the study, contributed to writing the report, interpreting results. AK contributed to data extraction and analysing data, updating reference lists.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Prabina Kumar Meher or Sachin Rustgi.

Reprints and permission information is available at http://www.nature.com/reprints

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