




# CV- $\alpha$ : designing validations sets to increase the precision and enable multiple comparison tests in genomic prediction

Rafael Massahiro Yassue · Felipe Sabadin · Giovanni Galli · Filipe Couto Alves · Roberto Fritsche-Neto 

Received: 17 April 2020 / Accepted: 16 April 2021 / Published online: 9 May 2021  
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

**Abstract** Usually, the comparison among genomic prediction models is based on validation schemes as Repeated Random Subsampling (RRS) or K-fold cross-validation. Nevertheless, the design of training and validation sets has a high effect on the way and subjectiveness we compare models. Those procedures cited above have an overlap across replicates that might cause an overestimated estimate and lack of residuals independence due to resampling issues and might cause less accurate results. Furthermore, ANOVA and multiple-comparison tests, such as Tukey, are not recommended due to assumptions unfulfilled regarding residuals' independence. Thus, we propose a new way to sample observations to build training and validation sets based on cross-validation alpha-based design (CV- $\alpha$ ). The CV- $\alpha$  was meant to

create several validation scenarios (replicates  $\times$  folds), regardless of the number of genotypes. Using CV- $\alpha$ , the number of genotypes in the same fold across replicates was much lower than K-fold cross-validation, indicating higher residual independence. Therefore, based on the CV- $\alpha$  results, as proof of concept, via ANOVA, we could compare the proposed methodology to RRS and K-fold cross-validation, applying four genomic prediction models with a simulated and real dataset. Concerning the predictive ability and bias, all validation methods showed similar performance. However, regarding the mean squared error and coefficient of variation, the CV- $\alpha$  method presented the best performance under the evaluated scenarios. Moreover, as it has no additional cost or complexity, it is more reliable and allows non-subjective methods to compare models and factors. Therefore, CV- $\alpha$  can be considered a more precise validation methodology for model selection.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10681-021-02831-x>.

R. M. Yassue · F. Sabadin · G. Galli ·  
R. Fritsche-Neto (✉)  
Department of Genetics, “Luiz de Queiroz” College of  
Agriculture, University of São Paulo,  
Piracicaba, São Paulo, Brazil  
e-mail: roberto.neto@usp.br

F. C. Alves  
Departments of Epidemiology and Biostatistics, Statistics  
and Probability and Institute of Quantitative Health  
Science and Engineering, Michigan State University,  
East Lansing, USA

**Keywords** Repeated random subsampling · K-fold cross-validation · Model selection

## Abbreviations

CV	Cross-validation
CV- $\alpha$	Cross-validation alpha-based design
GP	Genomic prediction
RRS	Repeated Random Subsampling
TGV	True genetic value
BIBD	Balanced incomplete block design

$\alpha$ -design	Alpha lattice design
PA	Predictive ability PA
CoV	Coefficient of variation
MSE	Mean squared error

## Introduction

Genomic prediction (GP) proposed by Meuwissen et al. (2001) evolved over the years, but it continues aiming to estimate breeding values of unevaluated genotypes. Hence, it is an important tool for plant breeders to shorten the breeding cycle, increase selection accuracy, and assess genetic variation (Heff et al. 2010; Crossa et al. 2017). Usually, to evaluate the prediction accuracy of the genomic prediction models, the data is divided into training and validation sets. The first set is used to fit the genomic prediction model and estimate the marker effects, whereas the validation set is used to validate the effects estimated in the training set and evaluate the predictions' accuracy (Crossa et al. 2011).

In the genomic prediction context, several methods and parameters have been proposed to compare prediction models (Blondel et al. 2015). Nevertheless, the predictive ability and the bias of the measures are two of the most utilized to evaluate the superiority and goodness of models and scenarios. The former is estimated by Pearson's correlation between the predicted and true breeding values of individuals contained in the validation set. The latter is obtained by regressing the predicted breeding values over the true ones to obtain the regression coefficient, which indicates the shrinkage (compression) between both (Piepho et al. 2008; Luan et al. 2009).

Some studies have shown that the model accuracy is influenced by the training and validation set (Akdemir et al. 2015; Wu et al. 2015; Auinger et al. 2016), being the main schemes to design training and validation sets in GP studies are K-fold cross-validation (Burgueño et al. 2012; Crossa et al. 2014; Fè et al. 2016) and Repeated Random Subsampling (RRS), also called Monte Carlo CV (Würschum et al. 2014; Yu et al. 2016; Zhang et al. 2016). The first consists of splitting the data into  $k$  groups (folds) and fitting a model using each fold as training and validation sets. In this sense, if  $k = 5$ , the model will be fitted five times. The second consists of splitting the dataset

randomly into training and validation sets. Both schemes are generally repeated  $n$  times (see Arlot and Celisse, 2010).

The accuracy estimate obtained by K-fold cross-validation might be affected by the number of folds, fold size, and replicates (Wong 2015). Likewise, in cross-validation schemes, the RRS is influenced by the relation between training and validation sets and the number of replicates (Kohavi 1995). Furthermore, some factors may lead to biased estimates of predictive ability, such as overlapping between the training and validation set and different relatedness between individuals through sets (Runcie and Cheng 2019). The overlap between training and validation sets over replicates may cause biased results due to correlations between predictions and non-independent residuals (Amer and Banos 2010). Therefore, neither validation schemes guarantee independence among replicates due to resampling issues. Thus, researchers cannot use standard and non-subjective methods to compare models and factors, such as ANOVA and other multiple comparison tests, due to assumptions unfulfilled regarding residuals independence. On the other hand, model selection based only on accuracy or error rate are not reliable because without a proper statistical test that considers the assumptions unfulfilled may lead the model selection to rely on uncertain decisions based on the stochastic variation (Hothorn et al. 2005; Boulesteix et al. 2015).

As an alternative, Shao (1993) proposed an extension of cross-validation (CV) schemes for model selection applying balanced incomplete block design (BIBD) to create the cross-validation scenario, considering that each fold is treated as "block" and each genotype as "treatment". The balanced distribution of the genotype across the fold in the BIBD will guarantee an equalized number of pairwise concurrences across genotype pairs ( $\lambda$ ). Therefore, the CV schemes using the incomplete block design may increase the quality of estimates (Fuchs and Krautenbacher 2016), residuals independence, and may allow further multiple comparison analyses. The limitation in using BIBD is that as the number of treatments increases, it becomes challenging to design orthogonal or balanced training and validation sets across the replicates without increasing the number of blocks substantially. This problem is similar to experimental field designs involving a large number of treatments. As an alternative, we are proposing to use a resolvable

partially incomplete block design as alpha lattice design ( $\alpha$ -design), once that it will seek to maintain balance among blocks with the advantage of the flexibility regarding the number of treatments and folds.

Based on described above, in this study, we propose a new method to design the training and validation sets for GP studies based on an alpha-lattice design scheme, called cross-validation alpha-based design (CV- $\alpha$ ), and compare its performance to the methods commonly applied in GP studies for model selection. Also, based on the CV- $\alpha$  results, a case of study, via analysis of variance (ANOVA), we could compare the proposed methodology to RRS and K-fold cross-validation, applying four genomic prediction models with a simulated and real dataset.

## Material and methods

In order to demonstrate the properties of the proposed cross-validation scheme, we aimed to mimic a standard genomic prediction study, for instance, comparing kernels and statistical methods. Thus, our aim is not to compare genomic matrices or Bayesian and frequentists approaches but simply show that our cross-validation scheme allows multiple comparison tests. For that, we create a simulated population (knowing the true parameters) and also use a well-known real dataset.

### Simulated dataset

We simulated a population of maize single-crosses from inbred parents to perform genomic prediction studies. For this, we used the *AlphaSimR* package (Gaynor et al. 2020). A founder population of 1000 individuals was simulated with ten chromosomes containing 30,000 segregating loci (SNPs). The individuals were inbred and diploid. Forty-nine individuals were randomly sampled and crossed to compose a partial diallel to obtain 906 hybrids. The phenotypic value (adjusted mean based on heritability) was simulated by randomly sampling 500 QTN from the segregating loci with a mean of 100 and variance 50. The narrow and broad-sense heritabilities were set to be equal to 0.30 and 0.50, respectively. Finally, to understand the validation methods' effect on the predictive ability and bias of the true genetic (TGV)

and phenotypic value, we performed genomics prediction using both metrics. We repeated the simulations 25 times and averaged the estimates above.

### An empirical case of study: USP maize dataset

We used a dataset of 906 maize single-crosses from a full diallel among 49 tropical inbred lines, according to Griffing's method 4 (Griffing 1956). The experiments were evaluated in two locations, two years, and under two nitrogen levels. The genotypic information from the 49 tropical inbred lines was obtained from Affymetrix® Axiom® Maize Genotyping Array, containing about 614,000 SNPs (Unterseer et al. 2014). For more details about the phenotypic and genotypic data, see (Fristche-Neto et al. 2018).

The markers with a lower call rate ( $< 95\%$ ), heterozygous loci on at least one individual, and linkage disequilibrium ( $> 0.90$ ) were removed. The missing markers were imputed using the *Beagle 4.0* algorithm (Browning & Browning 2009) from the *synbreed* R package (Wimmer et al. 2012). Later, each hybrid's genotype was built by combining the genotypes of its parental lines, and hybrids with minor allele frequency (MAF  $< 0.05$ ) were removed. After quality control, a total of 32,207 SNPs was available for further analysis.

We evaluated the grain yield (GY, Mg ha<sup>-1</sup>), corrected it to 13% moisture, and stand across the eight environments to perform the genomic prediction studies. It was used to estimate the BLUP for hybrids following the model:

$$y = Sl + Xb + Wc + Tg + Ui + \varepsilon$$

where  $y$  is the vector of the phenotypic value of hybrids;  $l$  is the vector of fixed effects of the environment (the combination of site  $\times$  year  $\times N$  level);  $b$  is the vector of fixed effects of blocks within an environment;  $c$  is the vector of fixed effects of checks;  $g$  is genotypic values, where  $g \sim N(0, I\sigma_g^2)$ ;  $i$  is the interaction between environments and checks, where  $i \sim N(0, I\sigma_{ge}^2)$ ;  $\varepsilon$  is the vector of random residuals from checks and hybrid by environments effects, where  $\varepsilon \sim N(0, I\sigma_e^2)$ .  $\sigma_e^2$  was jointly estimated based on  $e$  environments with  $r$  replicated checks in each site.  $S$ ,  $X$ ,  $W$ ,  $T$ , and  $U$  are the incidence matrices for  $l$ ,  $b$ ,  $c$ ,  $g$ , and  $i$  (Fristche-Neto et al. 2018).

## Genomic prediction

To perform the genomic prediction, we used the additive GBLUP model and the Reproducing Kernel Hilbert Spaces regression (RKHS). The following model equation is the general form of these two approaches:

$$\hat{g} = 1\mu + Za + \varepsilon$$

where  $\hat{g}$  is the vector of BLUP;  $\mu$  is the intercept;  $a$  is the vector of additive genetic effects with  $a \sim N(0, I\sigma_a^2)$ ; and  $\varepsilon$  is the vector of random residuals with  $\varepsilon \sim N(0, I\sigma_e^2)$ .  $1$  is the incidence vector of  $\mu$ , and  $Z$  is the incidence matrix for  $a$ .  $G$  is the genomic relationship matrix ( $G_a$  – additive genomic relationship matrix, and  $K$  – for Gaussian kernel), and  $I$  is the identity matrix.  $\sigma_a^2$  is the additive genetic variance for  $G_a$  or genetic variance for  $K$ , and  $\sigma_e^2$  is the residual variance. The additive genomic relationship matrix ( $G_a$ ) was calculated as  $G_a = \frac{WW'}{2\sum_{i=1}^n p_i(1-p_i)}$ , where  $W$  is the centered matrix of SNPs, and  $p_i$  is the frequency of the allele  $p$  in locus  $i$  (VanRaden 2008). The Gaussian kernel ( $K$ ) was calculated as  $K(x_i, x_j) = \exp\left(\frac{-hd_{ij}^2}{q_{0.05}}\right)$ , where  $x_i$  and  $x_j$  are the marker vectors for the  $i$ th and  $j$ th individuals, respectively, and  $q_{0.05}$  is the fifth percentile for the squared Euclidean distance  $d_{ij}^2 = \sum_k (x_{ik} - x_{jk})^2$  (Pérez-Elizalde et al., 2015). The  $h$  value was considered equal to 1.

We fitted two prediction models considering two statistical approaches (frequentist and Bayesian Genomic Best Linear Unbiased Predictor—GBLUP), resulting in four scenarios: 1) GBLUP with  $G_a$  kernel (GA\_MM); 2) GBLUP with  $K$  kernel (GK\_MM); 3) Bayesian GBLUP with  $G_a$  kernel (GA\_Bayes), and 4) Bayesian GBLUP with  $K$  kernel (GK\_Bayes).

The analyses were performed using *ASReml-R* (Gilmour et al. 2009) and *BGLR* (Pérez and de los Campos 2014) packages for R. For Bayesian GBLUP models were performed using 10,000 iterations, 3000 burn-in, and 5 thinning values. The convergence checks for Bayesian models are available in Supplemental Figure S1 and S2.

## Cross-validation alpha-based design (CV- $\alpha$ )

K-fold cross-validation can be seen as a field experiment using resolvable incomplete block design with the blocks represented by the folds and the genotype as treatments. However, during the sorting process, the congruence between pairs of genotypes in the fold across replication is not considered, neglecting the number of pairwise concurrences across genotype pairs. As an alternative, the cross-validation alpha-based design (CV- $\alpha$ ) is an extension of the methodology presented by (Shao 1993) and consists of assigning genotypes to folds in each replication by applying the alpha-lattice design. The CV- $\alpha$  was intended to create scenarios with two, three, or four replicates, regardless of the number of genotypes. Each replicate is split into folds, and the number of folds will determine the percentage of training and validation sets. Each fold across replicates is based on the  $\alpha(0,1)$  lattice design aiming to reduce the concurrences of any two genotypes (associate class) in the same fold (block) across the replicates. Hence, each concurrence between any pair of genotypes across replicates in the same fold will be 0 or 1 (Patterson & Williams 1976).

However, the  $\alpha(0,1)$ -lattice design assumptions involve the number of blocks ( $s$ ) and block size ( $k$ ) (number of the folds and fold size, in our context) to determine the number of genotypes (Patterson & Williams 1976). As the number of genotypes is variable in a real scenario, we computed the nearest smallest number to attend to the assumptions above, and the remaining genotypes were randomly allocated into the folds. The alpha lattice design was created using the *agricolae* package (Mendiburu 2019), and the scripts are available on Github (<https://github.com/allogamous/CV-Alpha>).

In order to compare CV- $\alpha$  with the other two benchmarks schemes, we simulated two scenarios: 5-folds with four replicates and 10-folds with two replicates. First, we simulated a scenario with the number of genotypes varying from 200 to 2,000 and computed the percentage of remaining genotypes that were randomly assigned into folds for each scenario. After, we compared the same two benchmarks schemes according to the mean and standard deviation of the concurrence of any two genotypes, i.e., the number of folds containing both genotypes. The simulations were replicated ten times.

### Cross-validation methods comparison

To evaluate the cross-validation alpha-based design (CV- $\alpha$ ) performance, we compared it to benchmark validation schemes: repeated random subsampling (RRS) and K-fold cross-validation using real and simulated datasets for genomic prediction. For RRS, we used 100 replicates, each with 80% of the data for the training set and the remaining 20% of the data for the validation set, whereas CV- $\alpha$  and K-fold cross-validation were used five-fold and four replicates. The number of replicates or folds for each method considers the most common values for genome prediction studies using Bayesian and frequentist approaches (Zhao et al. 2013; Zhang et al. 2015, 2016; Yu et al. 2016).

From those, we obtained the predictive ability of each statistical model for the different CV methods. The predictive ability was estimated as Pearson's correlations between the predicted and observed phenotypes. For each CV method, we estimated the slope coefficient for the regression of the predicted values of the validation sets on its phenotypes. Thus, the regression coefficient between predicted and genetic value was considered the prediction bias, measuring the degree of inflation/deflation of prediction genomics. Nonbiased models are expected to have a regression coefficient equal to 1. For CV- $\alpha$  and K-fold cross-validation, the level of averaging considered was at replicates. After, we compute the mean and standard deviation from the predictive ability and bias. Although RRS and K-fold cross-validation schemes do not have independence between replicates, ANOVA has been used to compare the predictive abilities from different models, even breaking the independence assumption. To verify how variance components of the models are affected by these methods and estimate the coefficient of variation, we performed the ANOVA test considering the following model:

$$l = 1\mu + X_1m + X_2n + X_3o + \varepsilon$$

where  $l$  is the vector of Pearson correlation transformed by Fisher  $z$ -transformation using the R package *DescTools* (Signorell 2021);  $\mu$  is the overall mean;  $m$  is the vector of statistical approach effect;  $n$  is the vector of relationship kernel;  $o$  is the vector of interaction between statistical approach and kernel; and  $\varepsilon$  is the vector of residuals.  $X_1$ ,  $X_2$  e  $X_3$  are

incidence matrices for  $m$ ,  $n$ , and  $o$ , respectively. Quadratic components were estimated by the method of moments based on mean square expectation.

## Results

### Cross-validation alpha-based design (CV- $\alpha$ )

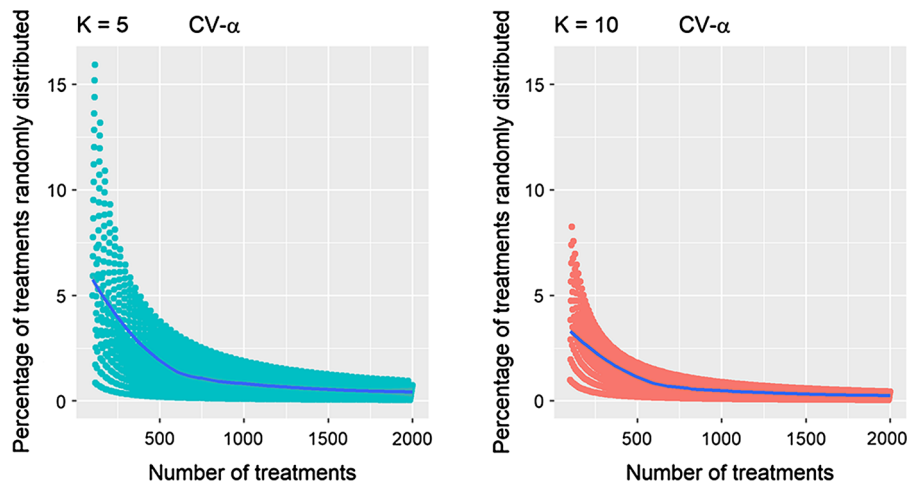
We performed several analyses to evaluate cross-validation alpha-based design (CV- $\alpha$ ) performance. For this, we computed the number of randomly assigned treatments among folds and the concurrence between pairs of treatments in the same fold across replicates (Fig. 1). The results revealed that the proportion of genotypes randomly assigned among folds in order to fulfill the alpha lattice design premises reduced as the number of treatments increased and tended to converge to 0.38% and 0.27% for five-folds with four replicates and ten folds with two replicates, respectively (Fig. 1). Considering the concurrence between pairs of treatments in the same fold across replicates, the CV- $\alpha$  revealed a lower mean and standard deviation in both evaluated scenarios when compared with the K-fold cross-validation (Fig. 2).

### Simulated dataset

To understand the effects of validation schemes on genomic prediction, we simulated populations to obtain true genetic values (TGV) and phenotypic values. The validation methods did not significantly influence the average prediction ability of TGV and phenotypic values. Nevertheless, the RRS had several "extreme" values when compared to K-fold cross-validation and CV- $\alpha$ . Besides, RRS showed a more substantial variation for bias, with several values overtaking 0.5 and 1.5 for phenotypic and TGV. (Fig. 3).

For predictive ability (PA), bias, TGV, and phenotypic value, in terms of mean and standard deviation, the three validation methods did not differ among them (Table 1), except for phenotypic bias for RRS. On the other hand, when we considered the mean squared error (MSE) and the coefficient of variation (CV), CV- $\alpha$  showed the lowest CV for all scenarios evaluated compared with RRS and K-fold cross-validation.





**Fig. 1** The proportion of treatments randomly distributed into folds to attend the alpha-design presupposition using CV- $\alpha$  with 5-folds with four replicates (a), and 10-folds with two replicates (b), based on simulated data

### Maize dataset

For the maize dataset, PA and bias showed similar mean values for all validation methods. In terms of SD, K-fold cross-validation and CV- $\alpha$  presented similar performance and were lower than RRS (Table 2). For mean squared error and coefficient of variation, CV- $\alpha$  presented lower values than K-fold cross-validation and RRS. The coefficient of variation for K-fold cross-validation was 34.70% and 10% higher than CV- $\alpha$  for PA and bias, respectively (Table 2).

We applied the CV- $\alpha$  to validate two statistical approaches (Bayesian and Mixed models) and two types of kernels (Additive and Gaussian kernel) for genomic prediction models (Table 3). For this, we applied a two-way ANOVA, and observed significant effects for types of the kernel for PA and bias. The Gaussian kernel (**K**) presented higher PA (0.44) than **G<sub>a</sub>** (0.42) and lower bias (1.00 and 0.98, for **K** and **G<sub>a</sub>**, respectively). For the type of two statistical approaches, the Bayesian revealed a more biased estimation (0.98) when compared with GBLUP (1.01) (Table 3).

We noted that the proportion of phenotypic variance explained variation by each source of variation varied across validation schemes (Fig. 3). PA and bias had similar performance across CV schemes for residual variance but varied for other variances. The RRS presented higher residual variance and lower variances due to model effects. For the interaction,

K-fold cross-validation showed higher values for PA. CV- $\alpha$  presented lower proportions of residual variances and higher variance due to the kernel and statistical approaches effects (Fig. 4).

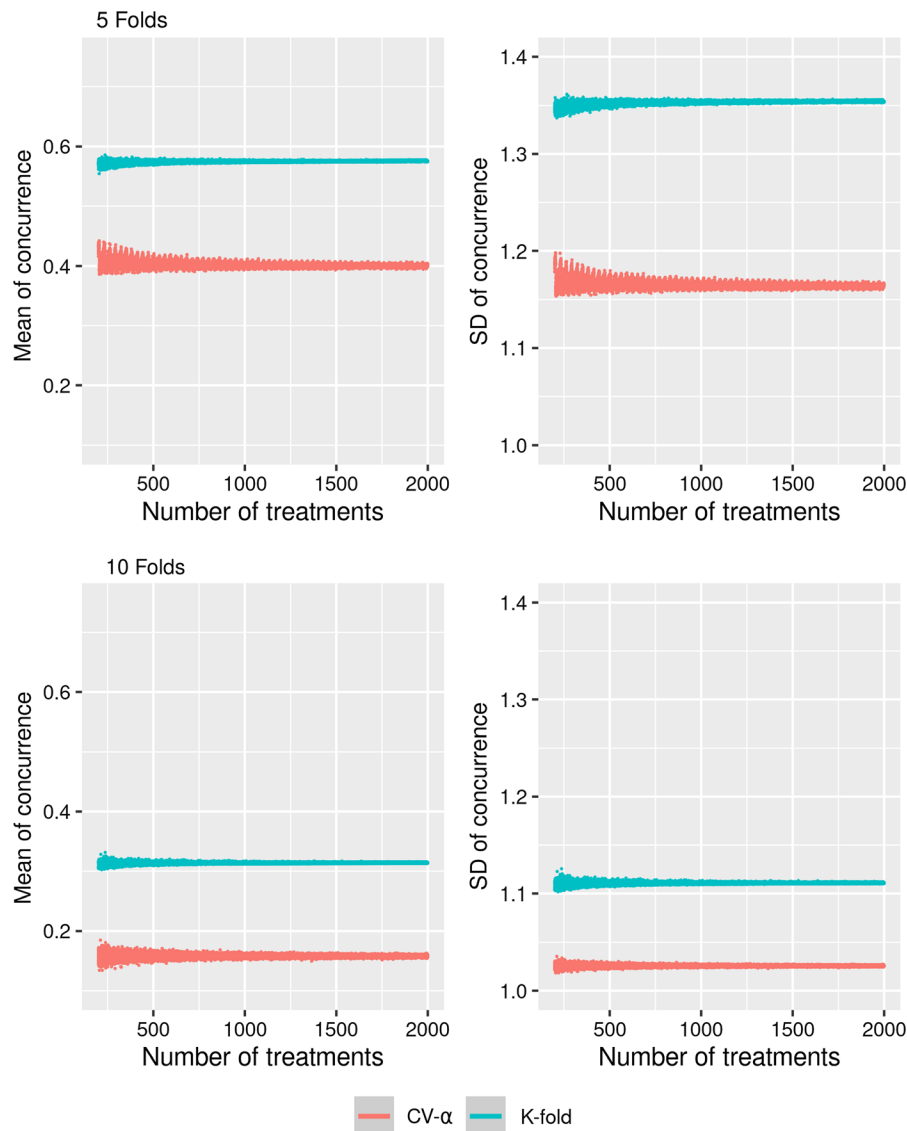
### Discussion

Considering the CV scheme proposed by Shao (1998), to obtain the number of replication needed to fulfill the condition to have a BIBD, we used the following equation:

$$\lambda(t-1) = r(k-1)$$

where  $t$  = number of genotypes;  $k$  = number of units per block (block size);  $r$  = number of replication of each genotype;  $\lambda$  = number of times that genotypes occur together in the same block (John 1998). In a fictitious dataset with  $t = 900$  with five-folds,  $k = 180$ , the minimum values of  $\lambda$  and  $r$  needed are 179 and 899, respectively. This high number of replications indicates that Shao's (1993) cross-validation is not suitable for model selection in the genomic prediction context. It will need much more computational resources, especially if covariance-variance structure complexes and several model scenarios are considered, common in this kind of study.

The main advantages of considering the  $\alpha$ -design instead of the BIBD and K-fold cross-validation are the flexibility regarding the number of treatments and folds (Singh & Bhatia 2017), reduce the concurrence

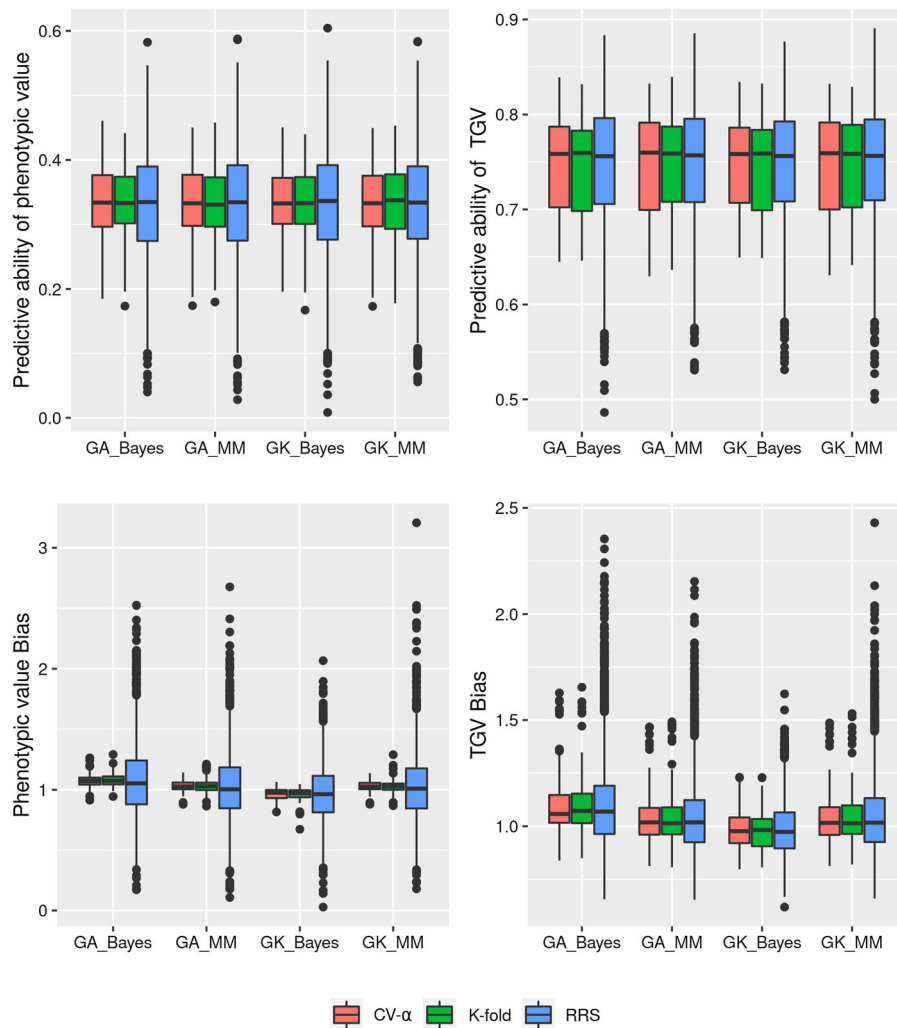


**Fig. 2** Concurrence (number of times that a pair of treatments appear together in the same fold) mean and standard deviation between treatments pairs in the same fold across replicates using

CV- $\alpha$  and  $K$ -fold with 5 and 10 folds with 4 and 2 replicates, respectively, based on simulated data

between pairs of treatments, increase the quality of estimates (Fuchs & Krautenbacher 2016), and residuals independence, allowing further multiple comparison analyses. The  $\alpha$ -design is widely used in plant breeding experiments as well as ANOVA (Alam et al. 2017; Ta et al. 2018; Galic et al. 2019). Based on this, in the context of genomic prediction, the flexibility of the CV- $\alpha$  is a worthy alternative to compare genomic selection models.

Our results revealed that CV- $\alpha$  reduced the concurrence between pairs of treatments (genotypes) in the same fold across replicates and its standard deviation compared with the  $K$ -fold cross-validation scheme (Fig. 2). The concurrence of any two treatments caused dependence among folds and comparative tests to become less precise. Thus, the CV- $\alpha$  designs fold and replicate with few or non-concurrence across folds, generating a more independent and



**Fig. 3** Predictive ability (PA) and bias for TGV and phenotypic value for three validation schemes (CV- $\alpha$ , K-fold, and RRS) and four genomic prediction models (scenarios)

better scheme for composing training and validation sets in a genomic prediction context.

Comparison between CV- $\alpha$ , K-fold cross-validation, and RRS must be pondered since they have a different level of averaging and different numbers of replicates than RRS, although CV- $\alpha$  and K-fold cross-validation are equivalent (Wong 2015). RRS showed a higher number of outliers, probability as results of the different levels of average. However, it is an internal procedure for the method. The strategy to divide folds and replicates according to the alpha-lattice design, as we suggested in CV- $\alpha$ , permits considering as replicate level mean, similar to replicating the alpha-lattice design effect.

Moreover, the RRS showed a large variation in PA estimates and, especially, for prediction bias. We expected values for bias around 1.0. However, the RRS showed several values overtaking 0.5 and 1.5, which shows a considerable inflation/deflation on the estimates. These results indicate that RRS is a less accurate method, mainly when we use few replicates.

Estimates more accurate combined with few replicates to run a CV scheme is desirable, especially when we consider a large number of genotypes, which is common in plant and animal breeding. In these cases, computing the inverse matrix of the genomic relationship matrix is a challenge, and several studies have been to handle this issue (Misztal et al. 2014; Misztal



**Table 1** Averaged of 25 simulated datasets for mean, standard deviation (SD), mean squared error (MSE), and coefficient of variation (CoV) for predictive ability (PA) and bias for three CV schemes (CV- $\alpha$ , K-fold, and RRS)

Scheme	Parameter	Mean	SD	MSE	CoV (%)
CV- $\alpha$	PA of phenotypic value	0.331	0.063	0.00017	3.78
	PA of TGV	0.748	0.053	0.00022	1.50
	Phenotypic Bias	1.024	0.064	0.00217	4.26
	TGV Bias	1.049	0.149	0.00040	1.68
K-Fold	PA of phenotypic value	0.331	0.062	0.00016	3.84
	PA of TGV	0.748	0.053	0.00023	1.52
	Phenotypic Bias	1.027	0.072	0.00251	4.36
	TGV Bias	1.050	0.151	0.00049	1.80
RRS	PA of phenotypic value	0.331	0.084	0.00414	19.41
	PA of TGV	0.748	0.062	0.00616	8.10
	Phenotypic Bias	1.024	0.274	0.07283	25.47
	TGV Bias	1.050	0.192	0.01366	10.26

**Table 2** Summary of ANOVA, mean, standard deviation (SD), and coefficient of variation (CoV) for three validation schemes (CV- $\alpha$ , K-fold, and RRS) for predictive ability (PA) and bias

Model	CV- $\alpha$			K-fold			RRS		
	Df	PA MS	Bias	Df	PA MS	Bias	Df	PA MS	Bias
St.Approaches	1	0.0002	0.0049 *	1	0.0002	0.0048 *	1	0.0035	0.1278
Kernel	1	0.0016 **	0.0025	1	0.0007	0.0005	1	0.1292 **	0.4748 **
St.Approaches:Kernel	1	0.0001	0.0001	1	0.0005	0.0002	1	0.0002	0.0123
Residuals	12	0.0001	0.0007	12	0.0002	0.0009	396	0.0046	0.0340
CV (%)		2.16	2.70		2.91	2.97		14.61	18.38
Mean		0.433	0.99		0.440	1.02		0.433	1.00
SD		0.014	0.033		0.016	0.033		0.070	0.188

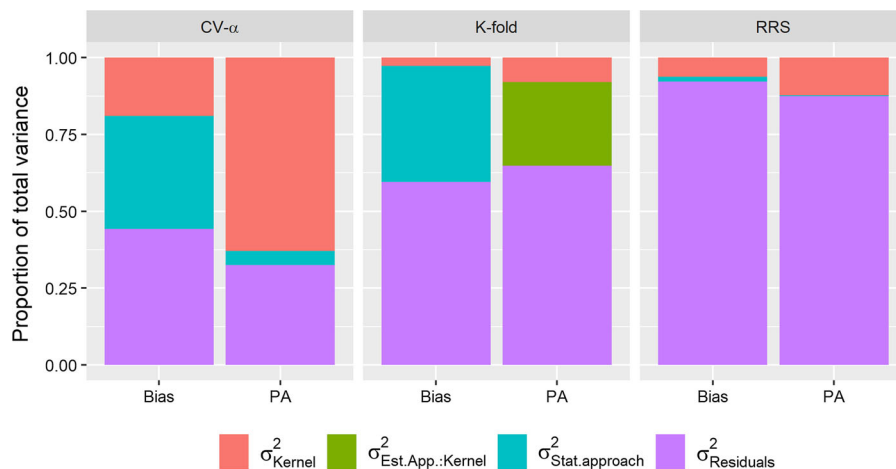
\*\*, \*, ns: Significant at 1%, 5%, 10% and non-significant of error probability by *F*- test

Statistical approaches (St. Approaches), Degrees of freedom (Df), Predictive ability (PA), Mean Squared (MS)

**Table 3** Means, marginal means, and Tukey's test for the type of kernels and statistical approaches for predictive ability (PA) and bias

	PA			
	G <sub>a</sub>		K	Marginal Means
Bayesian	0.424		0.436	0.430
Mixed models	0.426		0.445	0.436
Marginal Means	0.425	b	0.441	a
	Bias			
	G <sub>a</sub>		K	Marginal Means
Bayesian	0.963		0.992	0.977
Mixed models	1.002		1.023	1.012
Marginal Means	0.982	b*	1.008	a

\*Means followed by the same lowercase letter in the row and uppercase letter in the column do not differ by the Tukey test at 5% and 10%\* probability



**Fig. 4** The proportion of total variance decomposed into effects of the kernel, statistical approach, the interaction between the kernel and statistical approach, and residual for bias and predictive ability (PA) applied in three cross-validation schemes (CV- $\alpha$ , K-fold, RRS)

2016). Based on this, CV- $\alpha$  is a worthy alternative to design cross-validation schemes for GP purposes and has a more precise estimative in a case where the number of replicates is a limitation.

The simulated dataset used the correlation between the predict and the true genetic value in order to estimate the influence of the cross-validation population in the true predictive ability. As expected, no difference was observed among the cross-validation methods, once that, when we have several replications (20 for K-fold cross-validation and CV- $\alpha$  and 100 for RRS), all methods tend to converge to the true predictive ability of the model. The advantage of the CV- $\alpha$  in the context of genomic prediction model selection is related to selecting genomic prediction models using ANOVA and multiple-comparison tests. Indeed, for the simulated and maize dataset, when we consider in terms of MSE and coefficient of variation, CV- $\alpha$  has better performance due to higher independence across replicates.

Traditionally, in GP studies, model comparison and selection are based on subjective methods such as mean and standard deviation without a comparative test. Some studies also considered ANOVA and other statistical tests. Although due to assumption unfulfilled regarding residuals independence and our results, this is not recommended. CV- $\alpha$  reveals the lesser occurrence of pairs of genotypes in the same fold across replicates, causing a more precise estimative. The CV- $\alpha$  methodology consists of applying  $\alpha(0,1)$  lattice design to design the folds across

replicates, and because of this, it allows multiple-comparison tests.

The results above indicate that CV- $\alpha$  had a more precise estimative due to the reduction of coefficient of variation, and the variance components were better discriminated across the factors in the two-way ANOVA. It reveals how folds design across each replicate shifts the proportion of the total variation explained by each model factor reducing the residual variance. Furthermore, the ANOVA test using RRS and K-fold cross-validation to compare the performance of different models can produce mistake conclusions since the variance components estimate load bias. Therefore, CV- $\alpha$  allows determining how much variation each model factor has and compares different genomic selection models based on the ANOVA test and multiple-comparison tests. Furthermore, the use of CV- $\alpha$  does not imply any additional computer cost or complexity in the validation process of model selection.

As proof of concepts, we applied the proposed methodology to exemplify model selection. For the simulated and maize dataset, both do not show considerable differences across approaches (GBLUP and Bayesian) and kernel type (Additive genomic and Gaussian kernel) for predictive ability. Although, for the maize dataset, the use of  $K$  kernel showed higher predictive ability than  $G_a$ . This result is expected since the  $K$  kernel captures additive and non-additive effects (Heslot et al. 2012). For bias, mixed models showed less biased results. Although, comparison among these

models is not the focus of these studies since they have already been extensively studied (Chen et al. 2014; Gota & Gianola 2014; Cuevas et al. 2017).

In the context of genomic prediction studies, there are other ways to design training and validation sets. The CV- $\alpha$  may be expanded for these cases to better design training and test sets across replicates and environments, such as CV1 and CV2 schemes (Burgueño et al. 2012) and other multi-environment and multi-trait studies. Also, the CV- $\alpha$  may be applied in any other cross-validation studies to select models and verify as the model factors behave according to the different sources of variation.

## Conclusion

This study showed that the CV- $\alpha$  method is a worthy alternative to design cross-validations folds and replicates, mainly when researchers want to compare genomic prediction models, increasing precision in the model estimative, and unravel the model factors impact in the total variation. Even though there were no differences in the mean and standard deviation for predictive ability and bias, our proposal was more accurate in terms of the mean squared error and coefficient of variation. Another advantage of CV- $\alpha$  is that it does not require any additional cost regarding computing demand or complexity. Furthermore, CV- $\alpha$  allows using the non-subjective methods to compare models and factors through ANOVA and other multiple comparison tests, such as Tukey and Scott-Knott.

**Acknowledgements** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

## Declaration

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Akdemir D, Sanchez JI, Jannink JL (2015) Optimization of genomic selection training populations with a genetic algorithm. *Genet Sel Evol* 47:1–10. <https://doi.org/10.1186/s12711-015-0116-6>
- Alam MA, Seetharam K, Zaidi PH et al (2017) Dissecting heat stress tolerance in tropical maize (*Zea mays* L.). *F Crop Res* 204:110–119. <https://doi.org/10.1016/j.fcr.2017.01.006>
- Amer PR, Banos G (2010) Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit. *J Dairy Sci* 93:3320–3330. <https://doi.org/10.3168/jds.2009-2845>
- Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. *Stat Surv* 4:40–79. <https://doi.org/10.1214/09-SS054>
- Auinger HJ, Schönleben M, Lehermeier C et al (2016) Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.). *Theor Appl Genet* 129:2043–2053. <https://doi.org/10.1007/s00122-016-2756-5>
- Blondel M, Onogi A, Iwata H, Ueda N (2015) A Ranking Approach to Genomic Selection. *PLoS ONE* 10:e0128570. <https://doi.org/10.1371/journal.pone.0128570>
- Boulesteix A, Hable R, Lauer S, Eugster M (2015) A statistical framework for hypothesis testing in real data comparison studies. *Am Stat* 69:201–212. <https://doi.org/10.5282/ubm/epub.14324>
- Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84:210–223. <https://doi.org/10.1016/j.ajhg.2009.01.005>
- Burgueño J, de los Campos G, Weigel K, Crossa J, (2012) Genomic prediction of breeding values when modeling genotype  $\times$  environment interaction using pedigree and dense molecular markers. *Crop Sci* 52:707–719. <https://doi.org/10.2135/cropsci2011.06.0299>
- Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml User Guide Release 3.0 VSN International Ltd, Hemel Hempstead, HP1 1ES, UK [www.vsn.co.uk](http://www.vsn.co.uk)
- Chen L, Li C, Sargolzaei M, Schenkel F (2014) Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. *PLoS ONE* 9:1–7. <https://doi.org/10.1371/journal.pone.0101544>
- Crossa J, Pérez-Rodríguez P, Cuevas J et al (2017) Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci* 22:961–975. <https://doi.org/10.1016/j.tplants.2017.08.011>
- Crossa J, Pérez P, de los Campos G et al (2011) Genomic selection and prediction in plant breeding. *J Crop Improv* 25:239–261. <https://doi.org/10.1080/15427528.2011.558767>
- Crossa J, Pérez P, Hickey J et al (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* (edinb) 112:48–60. <https://doi.org/10.1038/hdy.2013.16>
- Cuevas J, Crossa J, Montesinos-López OA et al (2017) Bayesian genomic prediction with genotype  $\times$  environment interaction kernel models. *G3 Genes, Genomes, Genet* 7:41–53. <https://doi.org/10.1534/g3.116.035584>
- Fè D, Ashraf BH, Pedersen MG et al (2016) Accuracy of genomic prediction in a commercial perennial ryegrass breeding program. *Plant Genome*. <https://doi.org/10.3835/plantgenome2015.11.0110>

- 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. <https://doi.org/10.1007/s00122-018-3068-8>
- Fuchs M, Krautenbacher N (2016) Minimization and estimation of the variance of prediction errors for cross-validation designs. *J Stat Theory Pract* 10:420–443. <https://doi.org/10.1080/15598608.2016.1158675>
- Galic V, Franic M, Jambrovic A et al (2019) Genetic correlations between photosynthetic and yield performance in maize are different under two heat scenarios during flowering. *Front Plant Sci* 10:1–11. <https://doi.org/10.3389/fpls.2019.00566>
- Gaynor C, Gorjanc G, Hickey JM (2020) AlphaSimR: an R package for breeding program simulations. *G3 Genes Genomes, Genet* 0:1–5. <https://doi.org/10.1093/g3journal/jkaa017>
- Gota M, Gianola D (2014) Kernel-based whole-genome prediction of complex traits: A review. *Front Genet* 5:1–13. <https://doi.org/10.3389/fgene.2014.00363>
- Griffing B (1956) Concept of general and specific combining ability in relation to diallel crossing systems. *Aust J Biol Sci* 9:463–493
- Heff EL, Lorenz AJ, Jannink J, Sorrells ME (2010) Plant breeding with genomic selection: Gain per unit time and cost. *Crop Sci* 50:1681–1690. <https://doi.org/10.2135/cropsci2009.11.0662>
- Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: A comparison of models. *Crop Sci* 52:146–160. <https://doi.org/10.2135/cropsci2011.06.0297>
- Hothorn T, Leisch F, Zeileis A, Hornik K (2005) The design and analysis of benchmark experiments. *J Comput Graph Stat* 14:675–699. <https://doi.org/10.1198/106186005X59630>
- Kohavi R (1995) Proceedings of the 14th international joint conference on artificial intelligence - Volume 2. pp 1137–1143
- Luan T, Woolliams JA, Lien S et al (2009) The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. *Genetics* 1126:1119–1126. <https://doi.org/10.1534/genetics.109.107391>
- Mendiburu F (2019) *Agricolae*: statistical procedures for agricultural research. R package version 1.3-3. <https://CRAN.R-project.org/package=agricolae>
- Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157(4):1819–1829
- Misztal I (2016) Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* 202:401–409. <https://doi.org/10.1534/genetics.115.182089>
- Misztal I, Legarra A, Aguilar I (2014) Using recursion to compute the inverse of the genomic relationship matrix. *J Dairy Sci* 97:3943–3952. <https://doi.org/10.3168/jds.2013-7752>
- Patterson HD, Williams ER (1976) A new class of resolvable incomplete block designs. *Biometrika* 63:83–92. <https://doi.org/10.1093/biomet/63.1.83>
- Pérez P, de los Campos G, (2014) Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 2:483–495
- Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161:209–228. <https://doi.org/10.1007/s10681-007-9449-8>
- Runcie D, Cheng H (2019) Pitfalls and Remedies for Cross Validation with Multi-trait Genomic Prediction Methods. *G3 Genes, Genomes, Genet* g3.400598.2019 . doi: <https://doi.org/10.1534/g3.119.400598>
- Shao J (1993) Linear model selection by cross-validation. *J Am Stat Assoc* 88:486–494. <https://doi.org/10.1016/j.jspi.2003.10.004>
- Signorell A (2021) DescTools: tools for descriptive statistics. R package version 0.99.41. <https://cran.r-project.org/package=DescTools>
- Singh P, Bhatia D (2017) Incomplete block designs for plant breeding experiments. *Agric Res J* 54:607–611. <https://doi.org/10.5958/2395-146x.2017.00119.3>
- Ta KN, Khong NG, Ha TL et al (2018) A genome-wide association study using a Vietnamese landrace panel of rice (*Oryza sativa*) reveals new QTLs controlling panicle morphological traits. *BMC Plant Biol* 18:1–15. <https://doi.org/10.1186/s12870-018-1504-1>
- Unterseer S, Bauer E, Haberer G et al (2014) A powerful tool for genome analysis in maize: Development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* 15:1–15. <https://doi.org/10.1186/1471-2164-15-823>
- VanRaden PM (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>
- Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012) Synbreed: A framework for the analysis of genomic prediction data using R. *Bioinformatics* 28:2086–2087. <https://doi.org/10.1093/bioinformatics/bts335>
- Wong TT (2015) Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit* 48:2839–2846. <https://doi.org/10.1016/j.patcog.2015.03.009>
- Wu X, Lund MS, Sun D et al (2015) Impact of relationships between test and training animals and among training animals on reliability of genomic prediction. *J Anim Breed Genet* 132:366–375. <https://doi.org/10.1111/jbg.12165>
- Würschum T, Abel S, Zhao Y (2014) Potential of genomic selection in rapeseed (*Brassica napus* L.) breeding. *Plant Breed* 133:45–51. <https://doi.org/10.1111/pbr.12137>

- Yates F (1936) Incomplete randomized blocks. *Ann Eugen* 7:121–140. <https://doi.org/10.1111/j.1469-1809.1936.tb02134.x>
- Yu X, Li X, Guo T et al (2016) Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nat Plants*. <https://doi.org/10.1038/nplants.2016.150>
- Zhang X, Pérez-Rodríguez P, Semagn K et al (2015) Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity (edinb)* 114:291–299. <https://doi.org/10.1038/hdy.2014.99>
- Zhang X, Sallam A, Gao L et al (2016) Establishment and optimization of genomic selection to accelerate the domestication and improvement of intermediate wheat-grass. *Plant Genome* 9:1–18. <https://doi.org/10.3835/plantgenome2015.07.0059>
- Zhao Y, Zeng J, Fernando R, Reif JC (2013) Genomic prediction of hybrid wheat performance. *Crop Sci* 53:802–810. <https://doi.org/10.2135/cropsci2012.08.0463>

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